Chapitre 2
Estimation d'autres paramètres

Nous avons traité en détail du problème de l'estimation de la moyenne $\bar{y}_U$ d'une population. Mais la moyenne n'est pas le seul paramètre intéressant. Voici quatre autres paramètres qu'on a souvent intérêt à connaître ou à estimer :

- Le *total* t_y d'une population: par exemple, la production totale de blé dans les fermes d'une certaine région.
- Une *proportion p*: par exemple, la proportion des employés d'une compagnie qui serait favorable à un plan de soins dentaires ;
- Un *effectif* N_c : par exemple, le *nombre* d'employés favorables à un plan de soins dentaires.
- Un *quotient R* : *par* exemple, le nombre de postes de radio par personne dans les ménages d'une population.

Les concepts fondamentaux de l'estimation sont les mêmes, quel que soit le paramètre. Ce que nous ferons donc ici, c'est appliquer ces mêmes concepts à de nouvelles situations. Il n'y a que les formules qui changent, et encore : les trois premiers paramètres sont à un tel point liés à la moyenne $\bar{y}_U$ qu'on pourrait à la limite utiliser les mêmes formules. Le quotient R est un rapport de deux moyennes, et son étude exigera qu'on développe des outils particuliers à partir de zéro.

Rappelons ce que nous avons établi au sujet de la moyenne:

- L'estimateur $\bar{y}$ est sans biais, c'est-à-dire, $\mu_{\bar{y}} = \bar{y}_U$.
- L'écart-type des $\bar{y}$ est $\sigma_{\bar{y}} = \sqrt{1-f}\,\frac{S}{\sqrt{n}}$, où S est l'écart-type de la population, défini par $S = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1}}$.
- L'écart-type de $\bar{y}$ peut être estimé par $\hat{\sigma}_{\bar{y}} = \sqrt{1-f}\,\frac{s}{\sqrt{n}}$, où s est l'écart-type de l'échantillon, défini par $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$.
- Un intervalle de confiance approximatif à 95 % pour $\bar{y}_U$ est donné par [$\bar{y}$ - 2 $\hat{\sigma}_{\bar{y}}$; $\bar{y}$ + 2 $\hat{\sigma}_{\bar{y}}$].

2.1 Estimation d'un total

Considérons une population de comptes à recevoir de laquelle on a tiré un échantillon. Il est probablement nécessaire d'estimer la valeur moyenne $\bar{y}_U$ des comptes. Mais ce qui importe aussi, c'est de savoir quelle est la valeur *totale* de tous les comptes de la population. Comme le montre le prochain exemple (qui reprend les données de l'exemple 1.3.1), le passage de la moyenne $\bar{y}_U$ au total τ se fait naturellement et sans difficulté:

***Exemple* 2.1.1** *Estimation d'un total*

D'une population de N = 8 427 comptes à recevoir, on prélève un échantillon de taille n = 30 afin d'estimer la valeur totale des comptes des comptes de la population. Voici les résultats, en dollars:

240,82	232,50	740,81	860,32	224,10	7,15	324,12	240,12	190,08	182,75
160,21	148,22	132,19	119,25	113,85	108,30	107,10	101,19	99,21	93,12
88,13	80,15	78,13	72,15	67,13	65,14	41,10	32,17	10,02	9,15

a) Estimer le total τ de la population

b) Déterminer un intervalle de confiance à 95 % pour le total τ de la population.

Solution

La façon la plus simple d'aborder le problème, c'est de traiter d'abord de la moyenne. Nous avons les données suivantes :

$$\bar{y} = \frac{4968,68}{30}, \quad s = 189,553936$$

a) L'estimateur de la moyenne est

$$\bar{y} = \frac{4968,68}{30} = 165,6226667$$

Pour estimer la valeur *totale* des comptes, il suffira de multiplier cette valeur par le nombre de comptes dans la population. L'estimation de τ, que nous dénoterons par T, est donc

$$T = 8427 \times 165,6226667 = 1\ 395\ 702\ \$$$

(Remarquez que nous avons conservé beaucoup de décimales dans $\bar{y}$ *et s*. C'est nécessaire, car lorsqu'on multiplie par N on risque de grossir l'erreur d'arrondi).

b) Ici aussi, nous commençons par déterminer un intervalle de confiance pour la moyenne. Nous avons

$$\hat{\sigma}_{\bar{y}} = \sqrt{1-\frac{n}{N}}\frac{s}{\sqrt{n}} = \sqrt{1-\frac{30}{8427}}\frac{189,553936}{\sqrt{30}} = 34,54599923$$

L'intervalle de confiance pour μ est donc

$$\bar{y} - 2\hat{\sigma}_{\bar{y}} \leq \bar{y}_U \leq \bar{y} + 2\hat{\sigma}_{\bar{y}}$$
$$165,6226667-2(34,54599923) \leq \bar{y}_U \leq 165,6226667+2(34,54599923)$$
$$96,530668 \leq \bar{y}_U \leq 234,714665$$

On obtient l'intervalle de confiance pour le total en multipliant les bornes par N

$$8427 \times 96,530668 \leq \tau \leq 8427 \times 234,714665$$
$$813464 \leq \tau \leq 1977940$$

■

2.2 Estimation d'une proportion

Un autre paramètre qu'on veut souvent estimer est p, la *proportion* des membres d'une population qui possèdent une certaine caractéristique, ou qui appartiennent à une certaine classe $\mathcal{C}$. p peut être la proportion de comptes impayés dans une population de comptes; ou la proportion de pièces défectueuses dans un lot de pièces fabriquées; ou la proportion de personnes exprimant telle ou telle opinion dans un sondage.

Une proportion est en fait une moyenne, de sorte qu'il n'est pas théoriquement nécessaire de développer de nouvelles techniques. Mais c'est une moyenne particulière qui mérite, compte tenu de son importance, d'être traitée séparément. Nous allons donc reprendre la démarche que nous avons suivie pour estimer $\overline{y}_U$: proposer un estimateur, vérifier s'il est sans biais, déterminer son écart-type, et finalement estimer l'écart-type.

L'estimateur Supposons que pour estimer la proportion de fumeurs dans une population, on prélève un échantillon de n personnes, et qu'on observe X fumeurs. L'estimateur naturel de p est naturellement la proportion échantillonnale de fumeurs, dénotée par $\hat{p}$:

$$\hat{p} = \frac{X}{n}$$

Cet estimateur est sans biais, c'est-à-dire,

$$\mu_{\hat{p}} = p$$

Écart-type de l'estimateur L'écart-type de l'estimateur est

$$\boxed{\sigma_{\hat{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \approx \sqrt{1-f}\sqrt{\frac{p(1-p)}{n}}}$$

Estimateur de $\sigma_{\hat{p}}$ On pourrait penser que pour estimer $\sigma_{\hat{p}}$, il suffirait de remplacer p dans l'expression de $\sigma_{\hat{p}}$ par $\hat{p}$. Ce n'est pas tout à fait faux, et on peut le faire sans graves conséquences. Mais l'estimateur suivant est légèrement meilleur :

$$\boxed{\hat{\sigma}_{\hat{p}} = \sqrt{1-f}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}}$$

Comme d'habitude, l'intervalle de confiance est donné par

$$\boxed{\hat{p}-2\hat{\sigma}_{\hat{p}} \le p \le \hat{p}+2\hat{\sigma}_{\hat{p}}}$$

***Exemple* 2.2.1** *Estimation d'une proportion*

Le tableau A.11 présente les résultats d'un sondage fait auprès de 138 étudiants de l'UQAM choisis parmi les étudiants d'un programme dont l'effectif est $N = 988$. On leur a demandé, entre autre, de s'exprimer sur l'énoncé « Le singe et l'homme ont un ancêtre commun ». Il y en 53 parmi les 138 qui se sont déclarés d'accord. Soit p la proportion d'étudiants du programme qui sont d'accord. L'estimation $\hat{p}$ de p est

$$\hat{p} = \frac{53}{138} = 0{,}3841$$

L'estimation de l'écart-type de $\hat{p}$ est

$$\hat{\sigma}_{\hat{p}} = \sqrt{1-\frac{138}{988}}\sqrt{\frac{\left(\frac{53}{138}\right)\left(1-\frac{53}{138}\right)}{138-1}} = 0{,}03854$$

Voici, finalement, un intervalle de confiance à 95% pour p :

$$0{,}3841 \pm 2(0{,}03854) = 0{,}3841 \pm 0{,}0771$$

On estime donc que la proportion de personnes dans le programme qui croient que l'homme et le singe ont un ancêtre commun est de 38,41%, avec une marge d'erreur de 7,771%. Ou encore, on peut affirmer avec à peu près 95% de confiance que 30,7% $\leq p \leq$ 46,16% ■

Remarque *Il est utile de constater que l'écart-type de $\hat{p}$ est borné par une limite supérieure: le produit $p(1-p)$ atteint son maximum lorsque p = ½, et donc*

$$\sigma_{\hat{p}} = \sqrt{1-\frac{n}{N}}\frac{S}{\sqrt{n}} = \sqrt{1-\frac{n}{N}}\frac{\sqrt{Np(1-p)/(N-1)}}{\sqrt{n}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}$$

Nous verrons plus tard que cette propriété est utile pour évaluer à l'avance la taille de l'échantillon nécessaire pour obtenir un niveau de précision donné. ■

2.3 Estimation d'un effectif

Tout comme avec le total $t_y = N\,\bar{y}_U$ d'une variable quantitative, nous pouvons définir *l'effectif* $N_c = Np$ d'une certaine classe $\mathcal{C}$ de la population : le *nombre* d'unités qui appartiennent à la classe plutôt que la *proportion*. La solution est parfaitement naturelle : On estime p *(par* $\hat{p}$), puis on multiplie $\hat{p}$ par N. L'estimateur de N_c est donc

$$\hat{N}_c = N\hat{p}$$

De même pour l'intervalle de confiance : on détermine un intervalle de confiance pour p, puis on multiplie les bornes *par* N :

***Exemple* 2.3.1** Dans l'exemple 2.2.1, supposons qu'on veuille estimer le nombre de personnes dans le programme qui seraient intéressés à assister à un bal de fin de baccalauréat. Et que, parmi les 138 répondants, 46 se sont déclarés intéressés.

a) Estimer le nombre de personnes N_c qui sont intéressés au bal
b) Déterminer un intervalle de confiance pour N_c

Solution

a) On estime la *proportion p* d'abord. $\hat{p} = \frac{46}{138} = 1/3$

Puisque $\hat{p}$ = 0,333 (un tiers des gens sont intéressés) et *que* N = 988, nous estimons le nombre de personnes intéressés par

$$\hat{N}_c = 988 \times (1/3) = 329{,}3333.$$

Puisqu'il s'agit d'un effectif, nous arrondissons, et estimons donc que le nombre de personnes intéressées à 329.

b) Déterminons d'abord un intervalle de confiance pour p. L'écart-type de l'estimateur est

$$\hat{\sigma}_{\hat{p}} = \sqrt{1-\frac{n}{N}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{1-\frac{138}{988}}\sqrt{\frac{(1/3)(2/3)}{138-1}} = 0{,}03735635$$

L'intervalle de confiance pour p est donc

$$\hat{p} \pm 2\,\hat{\sigma}_{\hat{p}} = 1/3 \pm 2(0{,}03735635) = 1/3 \pm 0{,}0747127.$$

L'intervalle de confiance pour l'effectif est donc

$$988 \times (1/3 \pm 0{,}0747127) = [256\ ;\ 403]$$

Le nombre de personnes intéressées se trouve quelque part entre 256 et 403 . ■

2.4 Estimation d'un quotient

Considérons le tableau 2.4.1 qui présente les données d'un échantillon de 50 logements tirés d'une population de 1 500 logements. Les variables sont:

x: Nombre de personnes dans le logement
y: Superficie du logement, en mètres carrés

Supposons que l'objectif est d'estimer le nombre de *mètres carrés par personne* dans les logements de la population. Il s'agit donc d'un paramètre nouveau, un *quotient*, que nous dénoterons par R , soit

$$R = \frac{\textit{Nombre total de mètres carrés dans la population}}{\textit{Nombre total de personnes dans la population}} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

Si on divise numérateur et dénominateur par N, on constate que R est aussi un quotient de moyennes :

$$R = \frac{\bar{y}_U}{\bar{x}_U},$$

où $\bar{y}_U$ et $\bar{x}_U$ sont les moyennes de y et de x , respectivement.

Encore une fois, on devine ce que devrait être l'estimateur de R : puisqu'il s'agit d'estimer le quotient de deux totaux (ou moyennes), nous utiliserons comme estimateur le même quotient, calculé à partir de l'échantillon, soit

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

L'estimateur $\hat{R}$ est légèrement biaisé, mais le biais est négligeable lorsque n est grand. L'écart-type de $\hat{R}$ est donné *approximativement* par

$$\sigma_{\hat{R}} \approx \frac{\sqrt{1-f}}{\sqrt{n}\,\bar{x}_U} \frac{\sqrt{S_y^2 + R^2 S_x^2 - 2RS_{xy}}}{\sqrt{n}} = \frac{\sqrt{1-f}}{\sqrt{n}\,\bar{x}_U} \sqrt{\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}}$$

***Tableau* 2.4.1** *Données sur un échantillon de* 50 *logements*
x: Nombre de personnes; y: superficie du logement (*en* m^2)

x	y	x	y	x	y	x	y	x	y
2	87	7	116	4	95	4	98	4	95
5	112	3	90	7	122	6	105	4	92
4	95	2	85	3	92	2	83	4	98
6	95	2	86	1	80	2	85	2	83
4	95	5	112	6	95	3	96	2	83
5	108	6	115	6	115	3	96	6	120
4	98	5	108	2	83	2	87	3	90
4	92	6	100	2	86	6	95	3	90
2	86	4	92	1	80	5	96	5	104
3	94	4	92	4	95	4	98	4	95

Pour estimer cet écart-type, on remplace dans l'expression ci-dessus les paramètre inconnus par leurs estimateurs : S_y^2, S_x^2 et S_{xy} par leurs estimateurs respectifs s_y^2, s_x^2, et s_{xy} ; R par $\hat{R}$; et $\bar{x}_U$ par $\bar{x}$.

Nous avons alors

$$\boxed{\hat{\sigma}_{\hat{R}} = \frac{\sqrt{1-f}}{\bar{x}} \frac{\sqrt{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}}}{\sqrt{n}} = \frac{\sqrt{1-f}}{\sqrt{n}\,\bar{x}} \sqrt{\frac{\sum_{i\in\omega}(y_i - \hat{R}x_i)^2}{n-1}}}$$

***Exemple* 2.4.1** *Estimation d'un quotient*

a) Estimer R, le nombre de mètres carrés par personne à partir des données présentées au tableau 2.4.1
b) Déterminer un intervalle de confiance pour R

Solution

Nous avons:

$$n = 50;\ \Sigma x_i = 193;\ \Sigma y_i = 4790;\ \bar{x} = 3{,}86;$$
$$s_x^2 = 2{,}5310204;\ s_y^2 = 111{,}06122;\ s_{xy} = 14{,}461224$$

a) L'estimation du quotient est donc

$$\hat{R} = \frac{\sum_{i=1}^{50} y_i}{\sum_{i=1}^{50} x_i} = \frac{4790}{193} = 24{,}818653$$

Nous estimons donc que, dans la population, l'espace dans les logements est de 24,8 m^2 par personne.

b) On estime l'écart-type de $\hat{R}$ par

$$\hat{\sigma}_{\hat{R}} = \frac{\sqrt{1-f}}{\bar{x}} \frac{\sqrt{s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}}}{\sqrt{n}}$$
$$= \frac{\sqrt{1-\frac{50}{1500}}}{3{,}86} \frac{\sqrt{111{,}06122 + 24{,}818653^2(2{,}5310204) - 2(24{,}818653)(14{,}461224)}}{\sqrt{50}}$$
$$= 1{,}11159$$

On construit un intervalle de confiance selon la règle générale suivie jusqu'ici, c'est-à-dire, en prenant pour marge d'erreur deux fois l'écart-type estimé de l'estimateur :

$$\hat{R} - 2\hat{\sigma}_{\hat{R}} \le R \le \hat{R} + 2\hat{\sigma}_{\hat{R}}$$

$$24{,}818653 - 2(1{,}11159) \leq R \leq 24{,}818653 + 2(1{,}11159)$$

$$22{,}60 \leq R \leq 27{,}94$$

La précision est excellente. Nous verrons plus loin pourquoi. ■

2.5 Estimation de la moyenne et du total d'un domaine

Il arrive qu'on s'intéresse non seulement à la population échantillonnée, mais aussi à certaines sous-populations, ou sous-ensembles de la population. Une sous-population est appelée *domaine* lorsque certains de ses paramètres sont des objets d'étude. Considérons un échantillon tiré d'une population $\mathcal{P}$, et supposons qu'on veuille estimer, à part certains paramètres de $\mathcal{P}$, les paramètres d'un domaine $\mathcal{D}$.

Par exemple, les données du tableau 2.5.1. Ce sont des données sur un échantillon de 50 professeurs choisis dans une population (nommons-la $\mathcal{P}$) de 200 professeurs. La variable d'intérêt y est le salaire, et l'échantillon a été tiré dans $\mathcal{P}$ pour estimer la moyenne des salaires $\bar{y}_U$. Supposons maintenant qu'après avoir estimé $\bar{y}_U$, on s'intéresse en particulier aux salaires des *femmes* de la population. Les femmes constituent alors un *domaine* $\mathcal{D}$.

***Tableau* 2.5.1** *Données tirées du tableau* A.02
Salaire et sexe de 50 *professeurs*

Salaire	*Sexe*	*Salaire*	*Sexe*	*Salaire*	*Sexe*	*Salaire*	*Sexe*	*Salaire*	*Sexe*
28390	F	41098	F	52927	F	41951	M	56463	M
32660	F	41198	F	54146	F	44044	M	56951	M
33659	F	42195	F	54146	F	45467	M	57683	M
35000	F	44044	F	54268	F	51220	M	58171	M
35976	F	45854	F	54390	F	51707	M	58780	M
36098	F	50000	F	62561	F	51951	M	58902	M
36929	F	50000	F	65732	F	53049	M	60000	M
36951	F	51585	F	29878	M	54390	M	62195	M
38659	F	51951	F	33537	M	54756	M	62317	M
40366	F	51951	F	35854	M	56220	M	64146	M

Comment estimer la moyenne $\bar{y}_{U_d}$ des salaires des femmes? L'échantillon comprend n_d = 27 femmes. Peut-on considérer cet échantillon comme un échantillon de 27 femmes tirées du domaine $\mathcal{D}$?[1] La réponse est oui. On procède donc comme d'habitude, en se limitant aux femmes de l'échantillon. Nous distinguons deux situations, selon que N_d est connu ou pas.

[1] La question surprend et entraîne souvent la réaction : pourquoi donc pas? On peut effectivement le faire, mais ce n'est pas aussi évident qu'on le pense. La raison est que la théorie développée jusqu'ici s'applique à un échantillon aléatoire simple. Mais nous n'avons pas tiré un échantillon aléatoire simple de 27 femmes dans une population de femmes. Nous avons tiré un échantillon aléatoire simple de 50 *personnes* dans une population de *personnes*, et nous nous sommes retrouvés avec 27 femmes. Ce n'est pas la même chose.

N_d connu

Lorsque N_d est connu, le problème ne présente aucune difficulté nouvelle. On estime la moyenne et le total en se limitant aux seules femmes de l'échantillon.

Voici les salaires des 27 femmes de l'échantillon :

28390	35976	38659	42195	50000	52927	54390	36951	54268
32660	36098	40366	44044	51585	54146	62561	41198	35000
33659	36929	41098	45854	51951	54146	65732	50000	51951

L'estimateur de la moyenne $\bar{y}_{U_d}$ des femmes est la moyenne $\bar{y}_d$ des 27 femmes de l'échantillon, qui est

$$\bar{y}_d = 45\ 286{,}44.$$

C'est donc notre estimation de $\bar{y}_{U_d}$.

L'écart-type de $\bar{y}_d$ sera estimé, comme d'habitude, par

$$\boxed{\hat{\sigma}_{\bar{y}_d} = \sqrt{1-\frac{n_d}{N_d}}\frac{s_d}{\sqrt{n_d}}}$$

Ici $n_d = 27$; s_d est l'écart-type corrigé des femmes de l'échantillon, $s_d = 9\ 564{,}466$. Mais là surgit un problème : que *vaut* N_d, le nombre de femmes dans le domaine? Dans certaines situations, N_d est connu; dans d'autres pas. Supposons qu'il est connu, $N_d = 81$ (c'est le nombre de femmes dans la population, qui est présentée au complet au tableau A.01). Nous avons alors

$$\hat{\sigma}_{\bar{y}_d} = \sqrt{1-\frac{n_d}{N_d}}\frac{s_d}{\sqrt{n_d}} = \sqrt{1-\frac{27}{81}}\frac{9564{,}466}{\sqrt{27}} = 1\ 502{,}911$$

L'intervalle de confiance est donc

$$\bar{y}_d \pm 2\hat{\sigma}_{\bar{y}_d} = 45\,286{,}44 \pm 2(1502{,}911),$$

$$42\,280{,}618 \le \bar{y}_{U_d} \le 48\,292{,}26$$

Estimation du total t_{yd} du domaine

Pour estimer le total t_{yd} du domaine, on multiplie l'estimation de la moyenne, $\bar{y}_d$, par N_d, puisque N_d est connu; et on détermine un intervalle de confiance en multipliant les bornes par N_d. Ainsi donc, si N_d est connu, $N_d = 81$, l'estimation du total est

$$T_d = N_d \times \bar{y}_d = 81\times 45\ 286{,}44 = 3\ 668\ 202$$

et l'intervalle de confiance pour τ_d est

$$81 \times 42\,280{,}618 \le \tau_d \le 81 \times 48\,292{,}26$$

$$3\ 424\,730 \le \tau_d \le 3\,911\,673$$

N_d inconnu

Estimation de la moyenne

Mais supposons que N_d ne soit pas connu. Signalons d'abord qu'en ce qui concerne l'estimation d'une moyenne on pourrait s'en passer : il suffirait de laisser tomber le facteur de correction $\sqrt{1-n_d/N_d}$ dans la formule $\hat{\sigma}_{\bar{y}_d} = \sqrt{1-\frac{n_d}{N_d}}\frac{s_d}{\sqrt{n_d}}$, dont l'influence est, dans la plupart des cas, négligeable. Mais s'il n'est pas négligeable, on peut toujours l'estimer. Dans un échantillon de taille 50, on a observé 27 femmes, soit 54 %. Si on suppose que c'est aussi le pourcentage de femmes dans la population, alors on estime leur nombre par $\hat{N}_d$ = 0,54×200 = 108. Remarquez qu'à ce moment, la fraction

$$\frac{n_d}{\hat{N}_d} = \frac{27}{(27/50)\times 200} = \frac{50}{200} = \frac{n}{N}$$

Donc ici, estimer $\hat{N}_d$ équivaut tout simplement à remplacer n_d/N_d par n/N. Dans le cas présent, cette substitution ne semble pas tout à fait anodine, car alors on a

$$\hat{\sigma}_{\bar{y}_d} = \sqrt{1-\frac{n}{N}}\frac{s_d}{\sqrt{n_d}} = \sqrt{1-\frac{50}{200}}\frac{9564,466}{\sqrt{27}} = 1594,078$$

une estimation assez éloignée de la première. L'intervalle de confiance est

$$42098 \le \bar{y}_{U_d} \le 48475$$

Estimation du total

L'estimateur est prévisible : on multiplie $\bar{y}_d$ par $\hat{N}_d$ plutôt que par N_d. L'estimateur devient

$$\hat{T}_d = \hat{N}_d \times \bar{y}_d = 108 \times 45\,286,44 = 4\,890\,936$$

La différence entre les deux estimations de t_d (T_d et $\hat{T}_d$) est majeure : une différence de 33 % par rapport à T_d. Le fait de ne pas connaître N_d fait perdre beaucoup de précision.

Comment estimer l'écart-type de $\hat{T}_d$? Hélas, ici il n'est pas question de multiplier $\hat{\sigma}_{\bar{y}_d}$ par $\hat{N}_d$ (car $\hat{N}_d$ est une variable aléatoire et non un nombre fixe). La solution, cependant, n'est pas difficile, bien qu'à première vue elle puisse sembler artificielle. Elle consiste à supposer que tous les hommes de la population ont un salaire nul. Pourquoi? Parce qu'alors, le total du domaine est égal au total de la population entière—et le problème est réduit à l'estimation d'un total, tout bonnement. On commence donc par poser égal à 0 tout salaire correspondant à un homme. L'échantillon présenté au tableau 2.5.1 se voit donc transformé en celui-ci :

28390	36929	42195	51951	54390	0	0	0	0
32660	36951	44044	51951	62561	0	0	0	0
33659	38659	45854	52927	65732	0	0	0	
35000	40366	50000	54146	0	0	0	0	
35976	41098	50000	54146	0	0	0	0	
36098	41198	51585	54268	0	0	0	0	

À partir d'ici, on estime un total, tout simplement. On a n = 50, N = 200. La moyenne échantillonnale, qu'on dénote par $\bar{y}'$, est la moyenne des 50 données ci-dessus, incluant les zéros; de même, l'écart-type corrigé s' est l'écart-type corrigé des 50 données, incluant les zéros.

On obtient les résultats suivants :

$$\bar{y}' = 24454,68 \; ; \; s' = 23840,52487.$$

On estime donc le total par

$$N\bar{y}' = 200(24454,68) = 4890936$$

On a

$$\hat{\sigma}_{\bar{y}'} = \sqrt{1-\frac{n}{N}}\,\frac{s'}{\sqrt{n}} = \sqrt{1-\frac{50}{200}}\,\frac{23840,52487}{\sqrt{50}} = 2919,856057$$

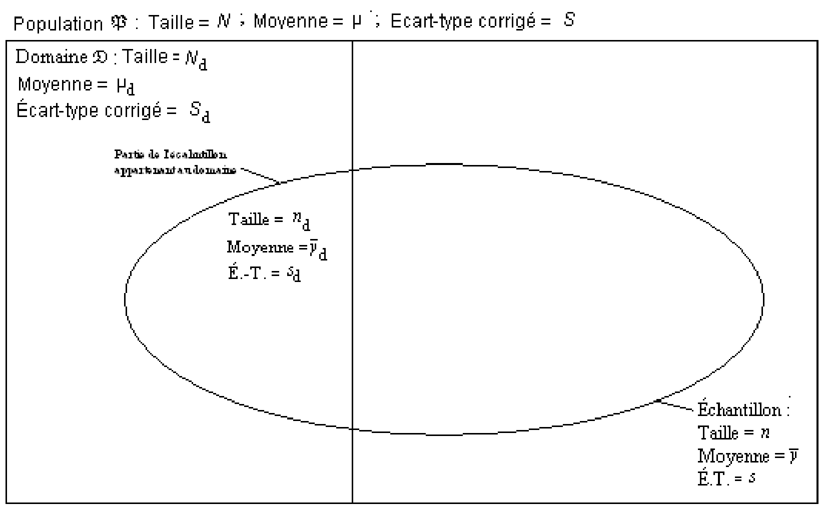
Pour obtenir un intervalle de confiance pour le total on multiplie l'intervalle $\bar{y}' \pm 2\hat{\sigma}_{\bar{y}'}$ par N = 200.

$$\bar{y}' \pm 2\hat{\sigma}_{\bar{y}'} = 24454,68 \pm 2(2919,856057).$$

Multiplié par N = 200, cet intervalle donne

$$3\,722\,994 \le \tau_d \le 6\,058\,878$$

Le schéma suivant résume la notation utilisée dans cette section :



2.6 Intervalle de confiance exact pour p

Une proportion p dans une population finie est de la forme $p = N_1/N$, où $N = N_1+N_2$, N_1 et N_2 sont des entiers positifs, $N_1 \le N$. L'estimateur de p est de la forme $\hat{p} = X/n$, où X est de loi hypergéométrique de paramètres n, N_1, et N_2. On peut déterminer un intervalle de confiance pour N_1 (et donc pour p) sans recourir à l'approximation normale. Pour un intervalle de niveau α, la procédure est la suivante : Soit x la valeur observée de X.

Limite supérieure :

La limite supérieure est la plus grande valeur de N_1 pour laquelle $P(X \le x \mid N_1) \ge \alpha/2$

Limite inférieure :

La limite supérieure est la plus petite valeur de N_1 pour laquelle $P(X \geq x \mid N_1) \geq \alpha/2$

Exemple : D'une population de taille $N = 70$ on tire un échantillon de taille $n = 25$. On observe $X = 10$. Déterminer un intervalle de confiance à 95 % pour N_1.

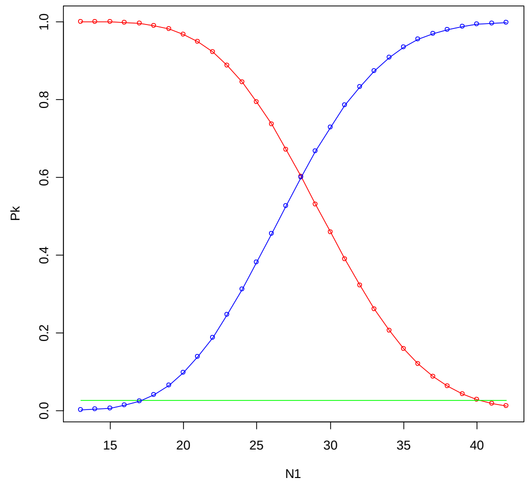
Solution Pour $N_1 = 13$ à 39 (l'intervalle de valeurs dans lequel on trouvera presque sûrement les limites), on calcule $P(X \geq 10 \mid N_1)$ et $P(X \leq 10 \mid N_1)$. Voici les résultats :

N_1	$P(X \geq 10 \mid N_1)$	$P(X \leq 10 \mid N_1)$
13	0.001	1.000
14	0.003	1.000
15	0.007	0.999
16	0.013	0.997
17	0.024	0.994
18	0.041	0.989
19	0.065	0.980
20	0.097	0.967
21	0.138	0.948
22	0.188	0.921
23	0.246	0.887
24	0.311	0.845
25	0.381	0.794
26	0.454	0.736
27	0.527	0.671
28	0.598	0.602
29	0.666	0.530
30	0.729	0.458
31	0.784	0.388
32	0.833	0.322
33	0.873	0.261
34	0.907	0.206
35	0.933	0.159
36	0.953	0.120
37	0.969	0.087
38	0.979	0.062
39	0.987	0.043
40	0.992	0.028
41	0.995	0.018
42	0.997	0.011

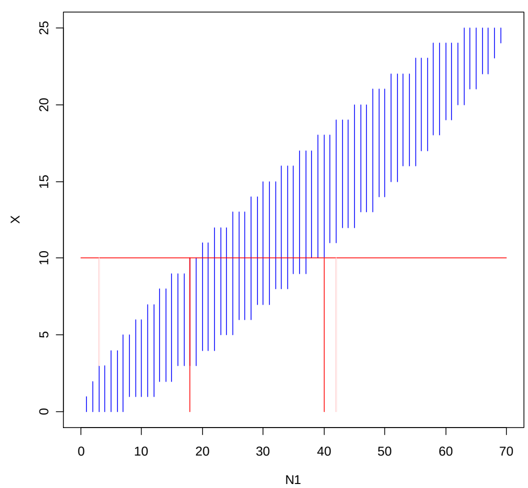
On conclut donc que $18 \leq N_1 \leq 40$.

L'approximation normale aurait donné $17 \leq N_1 \leq 39$

Voici un graphique représentant ces probabilités :



Autre présentation graphique : Pour chaque valeur de N_1 on déterminer un intervalle $[a(N_1) ; b(N_1)]$ de valeurs de X telles que $P[a(N_1) \leq X \leq b(N_1) \mid N_1] \geq 1\text{-}\alpha$. Chaque ligne bleue verticale dans le graphique ci-dessous représente l'intervalle correspondant à une valeur N_1 donnée. L'intervalle de confiance découlant d'une valeur observée $X = x$ est l'ensemble des valeurs de N_1 pour lesquelles l'intervalle $[a(N_1) ; b(N_1)]$ contient x.



2.7 Résumé

Voici un résumé des paramètres, leur estimateur, l'écart-type, et l'estimateur de l'écart-type.

Paramètre	*Estimateur*	*Écart-type de l'estimateur*	*Estimateur de l'écart-type de l'estimateur*
Moyenne $\overline{y}_U$	$\overline{y}$	$\sigma_{\overline{y}} = \sqrt{1-f}\,\dfrac{S}{\sqrt{n}}$	$\hat{\sigma}_{\overline{y}} = \sqrt{1-f}\,\dfrac{s}{\sqrt{n}}$
Total $\tau = N\mu$	$T = N\,\overline{y}$	$\sigma_T = N\sigma_{\overline{y}}$	$\hat{\sigma}_T = N\hat{\sigma}_{\overline{y}}$
Proportion p	$\hat{p} = \dfrac{X}{n}$	$\sigma_{\hat{p}} = \sqrt{\dfrac{N-n}{N-1}}\sqrt{\dfrac{p(1-p)}{n}}$	$\hat{\sigma}_{\hat{p}} = \sqrt{1-f}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}}$
Effectif $N_c = Np$ *d'une classe* $\mathcal{C}$	$\hat{N}_c = N\,\hat{p}$	$\sigma_{\hat{N}_c} = N\sigma_{\hat{p}}$	$\hat{\sigma}_{\hat{N}_c} = N\hat{\sigma}_{\hat{p}}$
Un quotient R = $\dfrac{\overline{y}_U}{\overline{x}_U}$	$\hat{R} = \dfrac{\overline{y}}{\overline{x}}$	$\sigma_{\hat{R}} \approx \dfrac{\sqrt{1-f}}{\mu_x}\dfrac{\sqrt{S_y^2+R^2S_x^2-2RS_{xy}}}{\sqrt{n}}$	$\hat{\sigma}_{\hat{R}} = \dfrac{\sqrt{1-f}}{\overline{x}}\dfrac{\sqrt{s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{xy}}}{\sqrt{n}}$
Moyenne $\overline{y}_{U_d}$ *d'un domaine* $\mathcal{D}$	$\overline{y}_d$: *Moyenne du domaine dans l'échantillon*		$\sqrt{1-\dfrac{n_d}{N_d}}\,\dfrac{s_d}{\sqrt{n_d}}$ *ou* $\sqrt{1-\dfrac{n}{N}}\,\dfrac{s_d}{\sqrt{n_d}}$ *selon que* N_d *est connu ou pas*
Total $\tau_d = N_d\overline{y}_{U_d}$ *d'un domaine* (N_d *connu*)	$T_d = N_d\,\overline{y}_d$		$N_d\sqrt{1-\dfrac{n_d}{N_d}}\,\dfrac{s_d}{\sqrt{n_d}}$
Total $\tau_d = N_d\mu_d$ *d'un domaine* (N_d *inconnu*)	$\hat{T}_d = \hat{N}_d\,\overline{y}_d = N\,\overline{y}'$ *où* $\hat{N}_d = \dfrac{n_d}{n}N$		$N\sqrt{1-f}\,\dfrac{s'}{\sqrt{n}}$

2.7 Exercices

2.1 D'une population de 850 transactions, on prélève un échantillon de taille 20. Les montants des transactions obtenus dans l'échantillon sont:

23,9	110,52	79,95	146,65	19,51	26,62	27,67	65,79	12,38	12,44
72,14	57,37	62,93	135,88	15,22	36,35	31,98	7,39	33,05	46,96

a) Estimer le total de la population et l'écart-type de l'estimateur.
b) Déterminer un intervalle de confiance à 95 % pour le total de la population.

2.2 À la fin d'une journée, le gérant d'un marché prélève un échantillon de 40 coupons de caisse parmi les 350 de la journée dans le but d'estimer le montant total des ventes de produits pharmaceutiques. Il trouve que seulement 10 coupons parmi les 40 comprenaient des achats de pharmaceutiques. Les montants pour ces 10 coupons sont:

21,28	2,32	18,76	16,36	4,59	4,99	6,67	7,28	8,1	6,6

a) Estimer le montant total des ventes en pharmaceutiques et estimer l'écart-type de votre estimateur.
b) Supposons qu'on fasse un travail préliminaire pour séparer, dans la population, les coupons qui comprennent des produits pharmaceutiques de ceux qui n'en comprennent pas, et qu'on trouve que dans la population 80 des coupons ont des produits pharmaceutiques. Estimer le total en utilisant cette information et estimer l'écart-type de l'estimateur.

2.3 D'une population de 880 comptes à payer, on prélève un échantillon de 30 comptes. On s'intéresse particulièrement aux comptes qui sont dus depuis 6 mois ou plus. Dans l'échantillon on trouve 17 comptes dus depuis 6 mois ou plus. Les montants de ces 17 comptes sont les suivants:

141,22	34,97	12,50	99,95	49,44	109,7	7,04	229,95	50,55
55,41	68,86	110,75	27,88	15,04	4,01	25,25	18,97	

a) Estimer le montant total de tous les comptes dus depuis 6 mois ou plus, et estimer l'écart-type de l'estimateur.
b) Supposons qu'on ait déterminé auparavant que 460 des comptes de la population sont dus depuis 6 mois ou plus. Estimer le total en utilisant cette information et estimer l'écart-type de l'estimateur.

2.4 La valeur aux livres des 1 230 comptes à recevoir d'une compagnie est de 136 220 $. Pour vérifier l'exactitude de ce montant, vous prélevez un échantillon de 25 comptes. Vous observez les montants suivants:

112,34	98,25	17,32	21,05	23,45	7,38	240,23	230,18	332,12	32,06
18,25	12,35	234,54	19,87	12,34	12,50	23,45	15,34	14,32	12,36
10,23	112,50	124,99	167,35	18,45					

a) Déterminer un intervalle de confiance à 95 % pour le montant total réel τ.
b) Déterminez un intervalle de confiance à 95 % pour la différence entre le total réel et la valeur aux livres. Peut-on conclure que la valeur aux livres est erronée?

2.5 Le tableau A.02 présente des données sur un échantillon de 50 professeurs d'université. Le nombre de professeurs dans la population est 200.

a) Estimez la proportion des professeurs engagés en 1980 ou avant. Estimez l'écart-type de votre estimateur (le nombre de professeurs engagés en 1980 ou avant est de 26 parmi 50).
b) Estimez le nombre de professeurs engagés en 1980 ou avant. Estimez l'écart-type de votre estimateur.
c) Estimez la moyenne du salaire en 2001 des professeurs engagés en 1980 ou avant. Estimez l'écart-type de votre estimateur.
d) Estimez le total des salaires de 2001 des professeurs engagés en 1980 ou avant. Estimez l'écart-type de votre estimateur. Faites le calcul de deux façons: (*i*) en supposant connu le fait que dans la population 120 personnes on été engagés en 1980 ou avant; et (*ii*) sans faire cette supposition.

2.6 Considérez l'échantillon de 50 professeurs présenté au tableau A.02 tiré d'une population de taille 200. On doit estimer le total des salaires de 2001 pour les deux domaines: a) le domaine 1, constitué de ceux qui ont 20 ans d'expérience ou moins; et b) le domaine 2, constitué de ceux qui ont 30 ans d'expérience ou moins. Voici quelques données :

Domaine	$\bar{y}_d$	s_d	$\bar{y}$	s'	n_d	$\bar{y}'$	N_d
exp≤20	44 628	9 878	48 447	16 944	8	7 141	24
exp≤30	49 357	10 580	48 447	25 368	32	31589	114

Estimez le total de chaque domaine, ainsi que l'écart-type de l'estimateur sous les deux hypothèses: l'une étant que vous connaissez les tailles des domaines dans la population; l'autre étant que vous ne la connaissez pas.

a) Est-ce qu'on gagne beaucoup à connaître la taille des domaines?
b) Est-ce que l'un des deux domaines permet une meilleure estimation du total? Utilisez l'écart-type de l'estimateur, ainsi que son coefficient de variation comme critère de qualité.

2.7 On prélève un échantillon de 25 librairies dans une ville ayant 250 librairies afin d'estimer le nombre total de livres espagnols vendus dans la ville au courant du mois dernier. Voici les résultats:

Nombre de livres espagnols	0	1	2	3	4	5	6	7	8
Nombre de librairies	14	3	2	4	0	0	0	1	1

a) Estimez le nombre total de livres espagnols vendus dans la ville au courant du mois dernier, et estimez l'écart-type de votre estimateur.
b) Supposons maintenant que vous apprenez qu'une étude faite auprès de la population entière a révélé que 175 librairies ne vendent jamais de livres espagnols. Utilisez cette information pour arriver à une deuxième estimation du nombre total de livres espagnols vendus au courant du mois dernier. Laquelle des deux estimations vous semble-t-elle plus précise?

2.8 Utilisez les données de l'échantillon présenté au tableau A.02 pour estimer le pourcentage d'augmentation de salaire entre l'entrée en fonction et 2001. Estimez le coefficient de variation de votre estimateur, et expliquez pourquoi il est si grand.

2.9 Une commission de transport vend des cartes d'abonnement par l'entremise des dépanneurs de la ville. Il y a en tout $N = 3\,221$ dépanneurs dans la ville. Vous voulez estimer le nombre moyen $\bar{y}_U$ de cartes vendues par les dépanneurs *qui en ont vendu au moins une*. En d'autres termes vous voulez estimer le paramètre

$$\bar{y}_U = \frac{Nombre\ \ total\ \ de\ \ cartes\ \ vendues}{Nombre\ \ de\ \ depanneurs\ \ qui\ \ en\ \ ont\ \ vendu\ \ au\ \ moins\ \ une}$$

Vous prélevez un échantillon de 100 dépanneurs. Voici la distribution du nombre de cartes vendues:

Nombre de cartes vendues	0	1	2	3	4	5	6	7
Nombre de dépanneurs	30	3	8	15	21	16	5	2

a) Estimez le nombre moyen de cartes vendues (parmi les dépanneurs qui en ont vendu au moins une), ainsi que l'écart-type de votre estimateur.
b) En fait, le numérateur du paramètre que vous voulez estimer est connu: le nombre total de cartes vendues est de 9773. Proposez un deuxième estimateur de $\bar{y}_U$ (il s'agirait de remplacer le dénominateur par une estimation).
c) Supposez maintenant que le nombre total de cartes vendues n'est pas connu et estimez-le. Estimez l'écart-type de votre estimateur.

2.10 Dans une étude sur la santé, on prélève un échantillon de 40 ménages afin d'obtenir une estimation de la proportion des gens qui font de l'exercice régulièrement. Les résultats sont présentés dans le tableau 2.7.1. La population est très grande.

Tableau 2.7.1 *Enquête sur la santé*
x: Nombre de personnes dans le ménage
y: Nombre de personnes qui font de l'exercice régulièrement

x	y	x	y	x	y	x	y	Calculs
4	0	6	6	5	1	3	0	
4	4	2	0	6	0	3	3	$\bar{x} = 4$; $\bar{y} = 1{,}7$
3	3	3	3	4	3	5	1	
4	0	2	2	2	2	6	0	$S_x = \sqrt{2}$; $S_y = \sqrt{2{,}93333}$
3	0	4	4	3	1	8	0	
5	4	4	3	3	0	1	0	
5	0	4	1	4	4	3	2	$S_{xy} = -0{,}025641$
3	3	4	0	5	5	3	3	
3	0	4	0	6	1	4	1	
4	2	4	4	7	0	4	2	

a) Estimez la proportion des gens qui font de l'exercice régulièrement, en considérant que vous avez un échantillon de n = 40 ménages et que vous estimez un *quotient*. Estimez l'écart-type de votre estimateur.

b) Estimez la proportion des gens qui font de l'exercice régulièrement, en considérant que vous avez un échantillon de n = 160 *personnes* et que vous estimez une proportion. Estimez l'écart-type de votre estimateur.

Vous allez constater que votre estimation est la même dans les deux cas, mais que l'écart-type est beaucoup plus petit dans le deuxième. Cette deuxième approche est erronée parce que l'échantillon est un échantillon de ménages et non de personnes. Elle donne une impression trop optimiste de la qualité de l'estimateur.

2.11 Considérez l'échantillon de 50 paroisses présenté au tableau A.04. C'est un échantillon des 210 paroisses présentées au tableau A.03. Dans chacun des cas suivants, estimez le paramètre et l'écart-type du paramètre, et déterminez un intervalle de confiance à 95 % (approximatif) pour le paramètre. La population est de taille N = 210.

a) La proportion de paroisses avec 10 naissances ou plus.
b) Le nombre de paroisses ayant 10 naissances ou plus
c) Le nombre total de mariages dans la population
d) Le nombre de mariage par habitant
e) Le nombre total de décès dans la strate 1.

2.12 Considérons la population de N = 8 unités pour lesquels sont définies deux variables, x et y, dont les valeurs sont données dans le tableau suivant:

y	8	10	67	44	66	56	89	99
x	3	6	24	27	30	36	51	57

Voici les valeurs de l'estimateur $\hat{R}$ pour les 56 échantillons possibles de taille n = 3:

2,57576	2,07937	1,94048	2,04938	1,88172	2,1	1,83333	2,47638
1,72222	2,10256	1,92222	2,02299	1,66667	2,11429	1,65789	1,88406
2,15385	2,07143	1,7	1,90476	1,66667	2,09009	1,65833	2,01515
1,64444	1,96667	1,69792	1,5942	1,73684	1,90991	1,71852	1,89655
1,78333	1,63636	1,76577	1,70238	2,18519	1,89744	1,80342	1,94444
1,77273	1,74074	2,12281	1,7	1,91954	1,93182	1,79675	1,84259
2,2037	1,73563	2,38333	1,83333	1,96078	1,78495	1,84058	1,69444

[*Voici quelques données sur ce tableau*: Somme: 106,4367; *Somme des carrés* : 204,7971]

a) Montrez que l'estimateur $\hat{R}$ est biaisé, comme on le sait, et exprimez une opinion sur l'importance du biais dans ce cas.

b) Calculez la variance (la vraie variance, à partir du tableau ci-dessus) de $\hat{R}$, et comparez avec la variance telle que calculée par la formule approximative $\sigma^2_{\hat{R}}$. (Rappelez-vous que la formule de $\sigma^2_{\hat{R}}$ n'est qu'une approximation et ne donne pas la vraie variance).

c) Quelle est la probabilité de se tromper de plus de 20% dans l'estimation du quotient?

2.13 Voici plus de détail (pages suivantes) sur les 56 échantillons issus de la population de taille 8 décrite dans le numéro précédent.

a) Comparez $\sigma_{\hat{R}}$ avec la quantité obtenue par la formule usuelle $\frac{\sqrt{1-n/N}}{\sqrt{n}\quad \bar{x}_U}\sqrt{S_y^2 + R^2 S_x^2 - 2RS_{xy}}$ = $\frac{\sqrt{1-n/N}}{\sqrt{n}\quad \bar{x}_U}\sqrt{\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}}$. Est-ce cette dernière quantité qui serait estimée par $\hat{\sigma}_{\hat{R}}$ plutôt que $\sigma_{\hat{R}}$? Dites ce que vous en pensez à la lumière de vos calculs.

b) Est-ce que s_{xy} est un estimateur sans biais de S_{xy} d'après ces données?

$\bar{y}$	$\bar{x}$	s_x	s_y	s_{xy}	$\hat{R}$	$\hat{\sigma}_{\hat{R}}$
28,33330	11	11,3578	33,5012	378,5	2,5758	0,2208910
20,66670	12	13,0767	20,232	264	1,7222	0,1020230
28,00000	13	14,7986	32,9242	486	2,1538	0,0889220
24,66670	15	18,2483	27,1539	495	1,6444	0,0954090
35,66670	20	26,8887	46,1988	1241,5	1,7833	0,0542830
39,00000	22	30,348	51,9711	1576,5	1,7727	0,0503540
39,66670	18	13,0767	29,7377	339	2,2037	0,3766690
47,00000	19	14,1774	33,7787	466,5	2,4737	0,1906680
43,66670	21	16,7033	31,3741	448,5	2,0794	0,3921180
54,66670	26	24,0624	41,8848	953,5	2,1026	0,3062730
58,00000	28	27,2213	46,1628	1201,5	2,0714	0,2974040
39,33330	20	14,7986	29,2803	416	1,9667	0,1883450

Suite

$\bar{y}$	$\bar{x}$	s_x	s_y	s_{xy}	$\hat{R}$	$\hat{\sigma}_{\hat{R}}$
36,00000	22	17,0587	24,98	426	1,6364	0,0623270
47,00000	27	24	40,5832	972	1,7407	0,0489450
50,33330	29	27,0555	45,8294	1238	1,7356	0,0445140
43,33330	23	17,5784	31,0054	515	1,8841	0,2152090
54,33330	28	24,0624	41,7413	989,5	1,9405	0,1479070
57,66670	30	27	46,0688	1228,5	1,9222	0,1467580
51,00000	30	24,5561	40,7308	994,5	1,7	0,0686680
54,33330	32	27,2213	45,5229	1233,5	1,6979	0,0634970
65,33330	37	29,5973	49,9032	1477	1,7658	0,0291060
40,33330	19	11,3578	28,6764	278,5	2,1228	0,3573230
47,66670	20	12,49	32,6241	394	2,3833	0,1941730
44,33330	22	15,0997	30,2379	379	2,0152	0,3668880
55,33330	27	22,6495	40,7717	862,5	2,0494	0,2837490
58,66670	29	25,865	45,0814	1103,5	2,023	0,2750840
40,00000	21	13,0767	28,2135	354	1,9048	0,1790160
36,66670	23	15,3948	23,8607	367	1,5942	0,0244650
47,66670	28	22,5167	39,6274	891,5	1,7024	0,0339020
51,00000	30	25,632	44,911	1150,5	1,7	0,0305430
44,00000	24	15,8745	29,8664	444	1,8333	0,2003690
55,00000	29	22,5167	40,6325	897	1,8966	0,1337540
58,33330	31	25,5147	44,9926	1129	1,8817	0,1321170
51,66670	31	22,9129	39,6779	905	1,6667	0,0588950
55,00000	33	25,632	44,5084	1137	1,6667	0,0553260
66,00000	38	27,8747	48,7545	1359	1,7368	0,0050570
59,00000	27	3	13	-1,5	2,1852	0,2499030
55,66670	29	6,245	11,5036	-15,5	1,9195	0,2883060
66,66670	34	14,7986	22,5019	267,5	1,9608	0,2322480
70,00000	36	18,2483	27,6225	439,5	1,9444	0,2242600
63,00000	30	6	6,0828	-33	2,1	0,2782050
74,00000	35	14,1774	13	178,5	2,1143	0,2306130
77,33330	37	17,5784	18,7705	323,5	2,0901	0,2307560
70,66670	37	13,5277	16,8028	159,5	1,9099	0,2276710
74,00000	39	16,7033	22,3383	304,5	1,8974	0,2183050
85,00000	44	17,5784	16,3707	285	1,9318	0,1855750
55,33330	31	4,5826	11,0151	19	1,7849	0,1615670
66,33330	36	13,0767	22,5019	271,5	1,8426	0,1178350
69,66670	38	16,5227	27,6827	434,5	1,8333	0,1144250
63,00000	38	12,1244	23,3024	280,5	1,6579	0,0494770
66,33330	40	15,3948	28,9194	443,5	1,6583	0,0472670
77,33330	45	15,8745	29,2973	465	1,7185	0,0211410
70,33330	39	10,8167	16,9214	153	1,8034	0,1255100
73,66670	41	14,1774	22,5019	289	1,7967	0,1202640
84,66670	46	14,1774	16,9214	239	1,8406	0,0928010
81,33330	48	10,8167	22,5019	243	1,6944	0,0411830

2.14 Après avoir annoncé son produit lors d'une émission de télévision locale, le propriétaire d'une chaîne de magasins fait faire une petite enquête auprès de 500 personnes de la région couverte par la station de télévision (17 324 habitants). Voici un résumé des résultats :

		Ont vu l'annonce		
		Oui	Non	Total
Ont vu l'émission	Oui	250	100	350
	Non	0	150	150
	Total	250	250	500

a) Déterminez un intervalle de confiance à 95 % pour la proportion de la population qui ont vu l'annonce.
b) Déterminez un intervalle de confiance à 95 % pour la proportion de la population qui ont vu l'émission.
c) Déterminez un intervalle de confiance à 95 % pour la proportion de ceux qui ont vu l'annonce parmi ceux qui ont vu l'émission
d) Déterminez un intervalle de confiance à 95 % pour le nombre de personnes qui ont vu l'annonce.
e) Déterminez un intervalle de confiance à 95 % pour le nombre de personnes qui ont vu l'annonce, en utilisant plutôt le fait, connu indépendamment, que 13 000 personnes ont vu l'émission.

2.15 On tire un échantillon aléatoire simple de n succursales parmi les N succursales d'une banque. Dans chacun des cas énumérés dans le cadre de gauche (Paramètre à estimer, ci-dessous), identifier le paramètre qu'il s'agit d'estimer. Faites votre choix dans le cadre de droite

Paramètre à estimer

a) Le salaire moyen des employés
b) Le salaire moyen des femmes employées
c) Le nombre de succursales dont le gérant est un homme
d) Le nombre moyen d'employés dans les succursales situées en banlieue
e) La proportion de femmes parmi les employés
f) La proportion de succursales dont le gérant est un homme
g) La masse salariale moyenne des succursales situées en banlieue
h) La masse salariale moyenne des succursales

Liste des réponses possibles

A : La moyenne $\overline{y}_U$ d'une certaine variable Y
B : La moyenne $\overline{y}_{U_d}$ d'une certaine variable Y dans un domaine $\mathcal{D}$.
C : Le total τ_d d'une certaine variable Y dans un domaine
D : Le total τ_y d'une certaine variable Y
E : Le nombre N_c d'unités appartenant à une certaine classe $\mathcal{C}$
F : Le quotient $R = \overline{y}_U / \overline{x}_U$ de deux variables Y et X
G : La proportion p d'unités appartenant à une certaine classe
H : Aucun de ces paramètres

2.16 Compléter chacune des phrases de la colonne de gauche. Choisir la suite dans la colonne de droite et inscrire la lettre correspondante.

<u>*Phrase*</u>

a) Dire que $\overline{y}_d$, l'estimateur de $\overline{y}_U$ par la différence, est sans biais, c'est dire que …

b) Dire qu'un estimateur $\hat{\sigma}$ de l'écart-type de la population σ est sans biais, c'est dire que …

c) Dire que $\hat{R}$ n'est pas sans biais pour R, c'est dire que…

d) Dire que l'estimateur de la variance de $\overline{y}$ est sans biais, c'est dire que …

e) Dire que $\overline{y}_d$, l'estimateur de $\overline{y}_U$ par la différence, est meilleur que $\overline{y}$, c'est dire que …

<u>*Liste des suites possibles*</u>

A. $\hat{R} \neq R$
B. $\hat{\sigma} = \sigma$
C. $\mu_{\hat{\sigma}} = \sigma$
D. $\mu_{\hat{\sigma}^2_{\overline{y}}} = \sigma^2_{\overline{y}}$
E. $\mu_{\hat{R}} \neq R$
F. $\mu_{\hat{\mu}_{yd}} = \overline{y}_U$
I. $\mu_{\hat{\sigma}} = \overline{y}_U$
J. $\hat{\sigma}^2_{\overline{y}} = \sigma^2_{\overline{y}}$
K. $\sigma_{\hat{\mu}_{yd}} < \sigma_{\overline{y}}$
L. $\hat{\mu}_{yd} = \overline{y}_U$
M. Aucune de ces suites