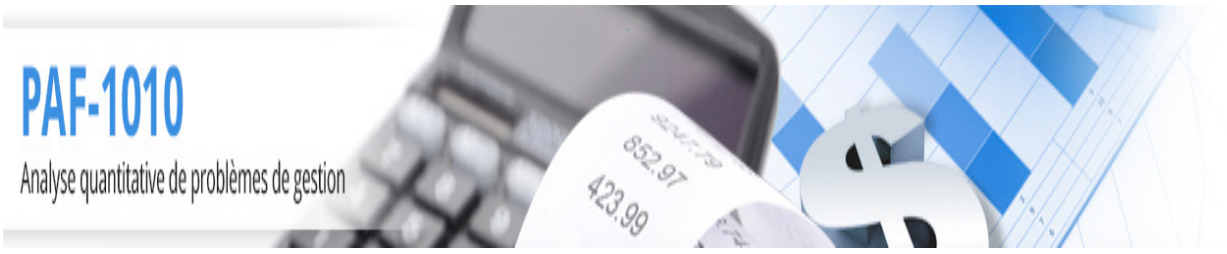
MODULE 9

ESTIMATION



Louis Houde
Département de Mathématiques et d'informatique
Université du Québec à Trois-Rivières

MODULE 9 Estimation

Objectifs et compétences

L'objectif de cette partie est de donner à l'étudiant les outils nécessaires pour appréhender les variations échantillonnales et les appliquer au problème d'estimation. L'étudiant devra identifier l'objectif de l'étude statistique et choisir la méthode appropriée pour apporter une solution raisonnable.

L'étudiant sera en mesure de

- identifier et calculer les estimateurs des principaux paramètres statistiques
- évaluer les probabilités approximatives sur la moyenne échantillonnale et sur le total
- identifier et utiliser les formules pour le calcul des principaux intervalles de confiance
- calculer la précision d'un intervalle de confiance et déterminer la taille de l'échantillon nécessaire pour obtenir une marge d'erreur donnée

9.1 Estimation ponctuelle

Estimateur

Le caractère qui est mesuré dans la population est associé à une variable aléatoire et de ce fait il y a des paramètres qui lui sont rattachés comme la moyenne μ, l'écart type σ, la probabilité de succès π, etc. L'estimation ponctuelle permet d'obtenir une approximation d'un paramètre de la population. Il est évidemment illusoire de vouloir connaître exactement un caractère aléatoire puisque sa principale propriété est justement de ne pas être connu avant la réalisation de l'expérience aléatoire et que les seules fenêtres que nous ayons sur ces phénomènes sont certaines de ses réalisations.

Si on prend par exemple la température du 8 janvier de l'année prochaine, elle ne peut être connue exactement puisque c'est une notion aléatoire et même la probabilité d'avoir une température de -4°C n'est pas connue exactement puisque le phénomène climatique menant à la journée du 8 janvier de l'année prochaine n'est pas entièrement maîtrisé. On est en présence d'une variable aléatoire qu'on veut connaître. Or on peut obtenir cette connaissance en regardant les caractéristiques des variables aléatoires pour appréhender une partie de l'information disponible sur ces phénomènes.

Exemple 1.1 ★Une entreprise désire connaître le salaire des cadres de même niveau dans les autres entreprises. Le paramètre d'intérêt est le salaire moyen μ tandis que l'estimation, ou approximation, sera le salaire moyen des cadres de même niveau observé dans les entreprises échantillonnées c'est-à-dire $\overline{X}$. On cherche à determiner si le salaire moyen est le même que dans les autres entreprises. Le salaire moyen des cadres est connu dans l'entreprise alors le problème est de vérifier si celui des autres entreprises est différent.

D'une façon plus générale, posons θ un paramètre quelconque de la population qui peut être un indice de centre, de dispersion, de probabilité ou autre. Un estimateur de ce paramètre, que nous notons $\widehat{\theta}$ est une fonction de l'échantillon qui donne une approximation de θ selon les observations disponibles, on note $\widehat{\theta}(X_1, X_2, \ldots X_n)$. C'est une v.a. puisqu'elle dépend de l'échantillon particulier qui a été choisi mais c'est aussi une approximation du paramètre d'intérêt θ.On s'intéresse dans un premier temps aux paramètres principaux de la variable aléatoire : moyenne, variance ou proportion de succès (μ, σ^2, π). D'autres paramètres peuvent aussi être d'intérêt mais c'est plus rare.

Il y a deux propriétés fondamentales des estimateurs qui sont importantes : il faut obtenir une approximation de la bonne chose en moyenne, c'est l'absence de biais et il faut que l'augmentation de l'information disponible se traduise par une meilleure approximation, c'est la convergence.

Mathématiquement cela se traduit par

- On dit qu'un estimteur est **non biaisé** si $E\left(\widehat{\theta}\right) = \theta$ c'est-à-dire l'approximation mesure la bonne chose en moyenne.
- On dit qu'un estimateur est **convergent** si $Var\left(\widehat{\theta}\right) \longrightarrow 0$ lorsque n converge vers l'infini.

Un estimateur convergent veut simplement dire que si on ajoute de l'information (taille de l'échantillon plus grande) alors la variation échantillonnale est moins grande d'où une plus grande précision de l'approximation.

Voici quelques estimateurs :

- Un estimateur ponctuel de la moyenne $E(X)$ est donné par $\overline{X} = 1/n \sum_{i=1}^{n} X_i$
- Un estimateur ponctuel de la variance σ^2 est donné par $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$
- Un estimateur ponctuel de l'écart type, σ, est donné par $S = \sqrt{S^2}$
- Un estimateur ponctuel de la probabilité de succès, π d'une expérience Binomiale est $p = 1/n \sum_{i=1}^{n} X_i$, où X_i est la v.a. qui donne 1 si un succès pour l'élément i de l'échantillon et 0 sinon, c'est-à-dire la proportion observée de succès dans l'échantillon.

Remarque 1.1 Il est important de faire la distinction entre l'estimateur et l'estimation. Dans le premier cas, c'est la v.a. qui permet d'obtenir une approximation du paramètre et puisque c'est une v.a. sa valeur n'est pas connue mais sa loi de probabilité peut être connue. L'estimation est la réalisation de cette variable aléatoire, c'est donc la valeur observée dans l'échantillon.

Remarque 1.2 On note généralement, mais pas toujours, par une lettre majuscule l'estimateur (la formule) et par une lettre minuscule l'estimation. Ainsi, on note $\overline{X}$ l'estimateur de μ et $\overline{x}$ l'estimation dans un cas précis.

Exemple 1.2 ★Un échantillon de 40 cigarettes d'une certaine marque a donné les teneurs en goudron (en mg) suivantes :

12,9	11,9	12,4	14,5	13,1	12,9	14,5	14,7	12,3	13,4	14,7
14,5	16,5	12,7	14,8	11,8	14,3	14,4	13,5	11,9	12,8	13,5
14,4	15,0	15,2	11,8	12,9	13,6	14,6	12,9	11,8	14,2	12,8
13,9	12,9	12,8	11,8	13,4	15,6	14,7				

La norme en vigueur recommande une teneur en goudron d'au plus 13 mg par cigarette. Donner une valeur estimée :

a) de la proportion des cigarettes de cette marque qui respectent la norme de la teneur en goudron

b) de la teneur en goudron moyenne des cigarettes de cette marque

c) de l'écart type de la teneur en goudron des cigarettes de cette marque.

Solution :

a) Une estimation de la proportion qui ne respecte pas la norme est donnée par le nombre de cigarettes qui ne respectent pas la norme dans l'échantillon divisé par le nombre de valeurs ($n = 40$)

$$p = 23/40 = .575$$

b) L'estimation de la teneur moyenne est

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{40} x_i = 13.558$$

c) Une estimation de l'écart type est

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{40}(x_i - \overline{x})^2} = \sqrt{5.207} = 1.187\,3$$

Exemple 1.3 ★Les diamètres de 20 vis produites par une machine sont les suivants :

1,05 1,04 1,06 1,02 1,03 1,04 1,07 1,09 1,02 1,03
1,05 1,03 1,09 1,07 1,03 1,05 1,07 1,04 1,02 1,01

Donner une valeur estimée pour le diamètre moyen des vis et pour la variance du diamètre.

Solution : Un estimateur de la moyenne est $\overline{X}$ et un estimateur de la variance est S^2. Or on observe

$$\overline{x} = 1.046 \text{ et } s^2 = 5.421 \times 10^{-4}$$

Exemple 1.4 ★Une étude sur la participation des femmes dans la vie active donne le taux en % de femmes qui travaillent et cela pour 19 villes américaines en 1968 et en 1972. Les résultats sont les suivants.

1968	42	50	52	45	43	55	45	34	45	54	42
1972	45	50	52	45	46	55	60	49	35	55	52

1968	51	49	54	50	58	49	56	63
1972	53	57	53	59	64	50	57	64

Donner une estimation du taux moyen de femmes dans la vie active en 1968 et en 1972.

Solution : On cherche une estimation du paramètre μ, le taux moyen de femmes dans la vie active. Pour 1968 on a $\overline{x}_{68} = 49.316$ et pour 1972 on a $\overline{x}_{72} = 52.684$.

Remarque 1.3 Il est parfois difficile de faire la différence entre les paramètres μ et π lorsqu'il est question de taux. C'est que le taux peut être une unité de mesure ou une estimation du paramètre π. Pour réussir à départager ces deux alternatives il faut regarder les mesures qui proviennent de l'échantillon. Dans l'exemple précédant on a un échantillon de 19 villes et pour chaque ville on a un taux en %. Le taux est donc une unité de mesure pour le caractère observé dans chaque ville et on s'intéresse au paramètre μ.

On s'intéresse au paramètre π si la mesure de chaque unité échantillonnale est un succès ou un échec : le taux est la proportion de succès dans l'échantillon. Ainsi si on prend un échantillon de 1000 femmes américaines et que l'on retrouve 514 de celles-ci dans la vie active alors on s'intéresse au paramètre π et le taux observé, 51.4% est une estimation de ce dernier.

Exemple 1.5 ★La SAAQ veut obtenir les taux de succès des examens théoriques pour obtenir le permis de conduire. Une enquête donnes les résultats suivants :

Bureau	1	2	3	4	5	6	7	8	9	10	11	12
Examens	120	75	84	200	103	140	167	69	87	107	124	100
Succès	84	58	62	147	92	117	120	41	47	84	95	65

On veut une estimation du taux moyen de succès et du taux général de succès.

Solution : Puisqu'on veut le taux moyen de succès on doit considérer qu'on a un échantillon de 12 bureaux et que pour chacun on a un taux de succès :

Bureau	1	2	3	4	5	6	7	8	9	10	11
Taux (%)	70.0	77.333	73.810	73.5	89.32	83.5 71	71.8 56	59.42	54.023	78.505	76.6

Il y a donc 12 valeurs et on demande la moyenne : $\overline{x} = 72.746$

Si on veut le taux global de succès on doit prendre tous les succès sur tous les examens : $p = 1012/1376 = 0.735\,47$.

Il faut faire attention à cette dernière valeur puisque c'est issu d'un échantillon aléatoire dans la seul mesure ou les personnes se présentent de façon aléatoire aux différents bureaux. Dans ce cas-ci les examens sont théoriques donc ils sont les mêmes pour tous.

9.2 Distribution d'échantillonnage

Les estimateurs sont des variables aléatoires donc ils ont une loi de probabilité, une moyenne, une variance etc. Le fait de connaître la distribution est intéressante pour obtenir une idée de la variation "raisonnable" des estimations. On utilisera cette distribution pour obtenir des marges d'erreur pour les estimations. Il y a une autre application à cette distribution d'échantillonnage : on peut calculer des probabilités approximatives sur des moyennes ou sur des totaux. On verra que dans certains problèmes très concret comme de déterminer le nombre de personnes maximal dans un ascenseur ou le poids de fret qu'un avion peut embarquer selon le nombre de passager, cette propriété est très intéressante.

Distribution de p

Considérons une v.a. $Bin\,(1, \pi)$, et un échantillon de taille n de cette distribution. La proportion de succès dans l'échantillon s'exprime comme étant

$$p = \frac{1}{n}\sum_{i=1}^{n} X_i$$

où chaque X_i prend la valeur 1 si on obtient un succès et 0 sinon. Cela revient à dire que p est simplement la proportion de succès observé dans l'échantillon.

Le théorème central limite (TCL) en statistique d'écrire les relations suivantes :

$$p \simeq N\left(\pi, \pi\left(1-\pi\right)/n\right)$$

et

$$\frac{p-\pi}{\sqrt{p\left(1-p\right)}}\sqrt{n} \simeq N\left(0,1\right)$$

si n est assez grand.

Cela veut dire qu'il est possible d'évaluer des probabilités sur la variable aléatoire p si les tailles d'échantillons sont assez grandes[1].

Exemple 2.1 ★★★Un joueur pense avoir une martingale[2] lui permettant de gagner à une machine automatique de Poker. Pour s'assurer que c'est vraiment une martingale, le joueur doit évaluer les probabilités réelles de gagner, π, étant donné son système de jeu. Il obtiendra p, la probabilité observée de gagner suite à ses expériences et la question est de déterminer si cette probabilité est intéressante par rapport à la probabilité de gagner sans la martingale. La société de loterie affirme que la probabilité de gagner est de 1/4 et le joueur obtient 1/3 pour 60 jeux. Peut-on dire que la valeur de 1/3 peut être uniquement dûe au hasard ?

Solution : Pour répondre à cette question, il faut calculer la probabilité de gagner au moins 20 jeux sur 60 en utilisant la martingale dans le contexte où la probabilité de gagner est réellement de 1/4 :

$$\Pr\left(p \geq \frac{1}{3} | \pi = \frac{1}{4}\right)$$

D'après le résultat ci-haut, on obtient

$$\begin{aligned}\Pr\left(p \geq \frac{1}{3} | \pi = \frac{1}{4}\right) &= \Pr\left(\frac{p - 1/4}{\sqrt{1/4\,(3/4)}}\sqrt{60} \geq \frac{1/3 - 1/4}{\sqrt{1/4\,(3/4)}}\sqrt{60} | \pi = \frac{1}{4}\right) \\ &= \Pr\left(Z \geq 0.149011\right) \text{ où } Z \sim N\left(0,1\right) \\ &\simeq 1 - 0.933\,2 = 0.066\,8\end{aligned}$$

Il est donc assez peu probable d'avoir observé une valeur de $1/3$ ou plus considérant la taille de l'échantillon et la valeur de référence, $1/4$. C'est une indication que la martingale est plus efficace que le hasard.

Exemple 2.2 ★★★Dans un état du sud des États-Unis la proportion de personnes en faveur de la vente libre d'armes à feu est historiquement de 60%. Lors d'un sondage auprès de 276 personnes on observe $p = 140/276 = 0.507\,25$. Ce résultat indique-t-il que le taux de 60% n'est plus valable ?

Solution : Pour répondre à la question on doit évaluer

$$\Pr\left(p \leq 0.50725 \mid \pi = 0.6\right)$$

c'est-à-direla probabilité d'observer une valeur aussi petite que $\frac{140}{276}$ si en réalité la vraie valeur est de $\pi = 0.6$ et que la taille d'échantillon est de $n = 276$.

[1] La notion de "grand", "très grand" et "très très grand" est assez flou en statistique. Elle découle des approximations qui en résultent. En général pour le paramètre π il est suffisant de prendre ≥ 20 si π est proche de 0.5, ≥ 30 si π est proche de 0.15 ou 0.85 et ≥ 50 si π est de l'ordre de 0.05 ou 0.95.

[2] "Une martingale est une technique permettant d'augmenter les chances de gain aux jeux de hasard tout en respectant les règles de jeu. Le principe dépend complètement du type de jeu qui en est la cible, mais le terme est accompagné d'une aura de mystère qui voudrait que certains joueurs connaissent des techniques secrètes mais efficaces pour tricher avec le hasard." (Wakipedia, http://fr.wikipedia.org/w/index.php?title=Martingale&oldid=25854896)

Or on sait que si $\pi = 0.6$ alors

$$p \simeq N\left(0.6, \frac{0.6\,(0.4)}{276}\right)$$

et ainsi

$$\begin{aligned}\Pr\left(p \leq 0.50725 | \pi = 0.6\right) &= \Pr\left(\frac{p-0.6}{\sqrt{0.6\times 0.4}}\sqrt{276} \leq \frac{0.50725-0.6}{\sqrt{0.6\times 0.4}}\sqrt{276} | \pi = 0.6\right) \\ &\simeq \Pr\left(Z \leq -3.145\,3\right) = .0008\end{aligned}$$

ce qui veut dire qu'il et fortement improbable que le taux réel soit inchangé.

Distribution de $\overline{X}$

La moyenne échantillonnale est une somme de variables aléatoires donc c'est une v.a. qui a une loi de probabilité. Le théorème central limite est un résultat très puissant qui permet entres autres de calculer des probabilités sur des moyennes ou des sommes de variables aléatoires avec une connaissance minimale, soit la moyenne et la variance de chacune. Ce théorème est lié au résultat suivant :

Proposition 2.1 Soit X une v.a. de loi normale de moyenne μ et de variance σ^2 représentant la mesure d'un caractère dans une population. Considérons un échantillon provenant de cette population alors

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Exemple 2.3 ★★Dans une entreprise les ressources humaines font passer un test d'aptitudes à chaque candidat. On sait que le score à ce test est une v.a. de loi $N\left(60, 31\right)$. Sur 10 candidats, quelle est la probabilité que le score moyen soit plus grand que 62 ?

Solution : On a $\overline{X} \sim N\left(60, 31/10\right)$ et on cherche $\Pr\left(\overline{X} > 62\right)$:

$$\begin{aligned}\Pr\left(\overline{X} > 62\right) &= \Pr\left(\frac{\overline{X}-60}{\sqrt{3.1}} > \frac{62-60}{\sqrt{3.1}}\right) \\ &= \Pr\left(Z > 1.135\,9\right) = .128\end{aligned}$$

où Z est une v.a. $N\left(0,1\right)$.

Exemple 2.4 ★★(suite) Pour un poste vacant, il y a 5 candidats provenant d'une université américaine qui se présentent et le score moyen de ces 5 candidats est de 55. Peut-on croire que les candidats sont moins "apte" que la moyenne de la population selon le score au test ?

On cherche $\Pr\left(\overline{X} < 55\right)$, or sur 5 individus, $\overline{X} \sim N\left(60, 31/5\right)$ et ainsi

$$\begin{aligned}\Pr\left(\overline{X} < 55\right) &= \Pr\left(\frac{\overline{X} - 60}{\sqrt{31/5}} < \frac{55 - 60}{\sqrt{31/5}}\right) \\ &= \Pr\left(Z < -2.008\right) = 0.022322\end{aligned}$$

On peut donc dire qu'il est peu probable que ces candidats proviennent d'une population ayant un score moyen au test de 60.

Proposition 2.2 Soit X une variable provenant d'une population qui est régie par une loi $N\left(\mu, \sigma^2\right)$, et X_i, $i = 1, 2, ..., n$ un échantillon alors

$$\frac{\overline{X} - \mu}{S}\sqrt{n} \sim t_n$$

C'est le résultat principal de la publication[3] de W.S. Gosset, alias Student, en 1908, celle qui a immortalisé le nom de Student plutôt que celui de Gosset...

Si la taille de l'échantillon est assez grande on sait que $t_n \simeq N\left(0, 1\right)$ et ainsi on peut obtenir une approximation de la probabilité sur $\overline{X}$ en utilisant la loi $N\left(0, 1\right)$. On considère généralement[4] qu'une taille de 30 est assez bonne pour calculer une probabilité. Le TCL (théorème central limite) permet de calculer des probabilités sur $\overline{X}$ quelque soit la distribution des observations :

Théoreme 2.3 Soit X une variable aléatoire de moyenne μ et de variance σ^2. Considérons un échantillon provenant de cette population alors si n est assez grand,

$$\frac{\overline{X} - \mu}{\sigma}\sqrt{n} \simeq N\left(0, 1\right)$$

et

$$\frac{\overline{X} - \mu}{S}\sqrt{n} \simeq N\left(0, 1\right)$$

Cela veut dire qu'il est possible d'obtenir une approximation d'une probabilité calculée sur $\overline{X}$ en ne connaissant que la moyenne et éventuellement la variance de la mesure. L'approximation dépend de la taille de l'échantillon et si la taille est infinie alors on a exactement une loi normale. Dans la pratique on considère qu'une taille de 30 est raisonnable pour obtenir une approximation valable[5].

[3] "The probable error of a mean". Biometrika 6 (1): 1–25. March 1908

[4] Généralement veut dire "pour des probabilités relativement grande" et une précision raisonnable. Cela veut dire que cette approximation n'est pas particulièrement adaptée pour des probabilités très petites ou très grandes et que la précision ne doit pas être de plus de trois chiffres significatifs.

[5] Cette "taille raisonnable" de 30 n'est pas une borne fixe qui convient à tous les problèmes : il faut considérer

Exemple 2.5 ★★★Dans une population le rythme cardiaque est de 80 battements par minute pour les femmes avec un écart type de 7.

Quelle est la probabilité d'observer une moyenne de moins de 78

- avec un échantillon de 10 personnes ?
- avec un échantillon de 50 personnes ?
- avec un échantillon de 1000 personnes ?

Solution : Posons X la v.a. qui donne le nombre de battements par minute, on a $\mu = 80$ et $\sigma = 7$ et on veut $\Pr\left(\overline{X} < 78\right)$.

On sait que si n est assez grand, quelque soit la distribution de la v.a. X,

$$\overline{X} \simeq N\left(80, 49/n\right)$$

On peut alors écrire

$$\begin{aligned}\Pr\left(\overline{X} < 78\right) &= \Pr\left(\frac{\overline{X} - 80}{7/\sqrt{n}} < \frac{78 - 80}{7/\sqrt{n}}\right) \\ &\simeq \Pr\left(Z < \frac{78 - 80}{7/\sqrt{n}}\right) = \Pr\left(Z < -\frac{2}{7}\sqrt{n}\right)\end{aligned}$$

- Si on observe un échantillon de taille $n = 10$,

$$\Pr\left(\overline{X} < 78\right) \simeq \Pr\left(Z < -\frac{2}{7}\sqrt{10}\right) = \Pr\left(Z < -.\,903\,51\right) = .\,174\,87$$

- Si on observe un échantillon de taille $n = 50$,

$$\Pr\left(\overline{X} < 78\right) \simeq \Pr\left(Z < -\frac{2}{7}\sqrt{50}\right) = \Pr\left(Z < -2.\,020\,3\right) = 2.\,169\,2 \times 10^{-2}$$

- Si on observe un échantillon de taille $n = 50$,

$$\Pr\left(\overline{X} < 78\right) \simeq \Pr\left(Z < -\frac{2}{7}\sqrt{1000}\right) = \Pr\left(Z < -9.\,035\,1\right) = 0$$

Il faut cependant tenir compte du fait que pour $n = 10$, l'approximation de la probabilité par une loi normale ne sera pas très bonne surtout pour des probabilités faibles.

Exemple 2.6 ★★★Dans une certaine population, le poids des individus est une variable ayant une moyenne égale à 60 kg et un écart type égal à 15 kg. Un ascenseur a une capacité égale à 2200 kg. Calculer

a) la probabilité que 36 individus pèsent ensemble plus de 2200 kg.

Solution : Posons T la v.a. qui donne le total des poids,

$$T = 36 * \overline{X}$$

Puisque la taille de l'échantillon est assez grande, on a

$$\overline{X} \simeq N\left(60, 15^2/36\right)$$

que lorsque la distribution des valeurs est fortement asymétrique il est recommandé de prendre une taille beaucoup plus grande.

et on veut $\Pr(T > 2200)$:

$$\begin{aligned}\Pr(T > 2200) &= \Pr\left(36\overline{X} > 2200\right) = \Pr\left(\overline{X} > \frac{2200}{36}\right) \\ &= \Pr\left(\frac{\overline{X} - 60}{15}\sqrt{36} > \frac{61.111 - 60}{15}\sqrt{36}\right) \\ &= \Pr(Z > .4444) = .32836\end{aligned}$$

b) le nombre maximum d'individus tel que la probabilité que le poids total soit de plus de 2200 kg soit au plus de 1%.

Solution : S'il y a n individus, alors si on pose T_n, la v.a. qui donne le total des n observations,

$$T_n = n\overline{X}$$

et

$$\overline{X} \simeq N\left(60, 15^2/n\right)$$

On veut trouver la taille maximale (n maximum) qui solutionne l'équation

$$\Pr(T_n > 2200) < 0.01$$

Cette dernière inégalité revient à

$$\begin{aligned}\Pr\left(n\overline{X} > 2200\right) &< 0.01 \\ \Pr\left(\overline{X} > 2200/n\right) &< 0.01 \\ \Pr\left(Z > \frac{2200/n - 60}{15}\sqrt{n}\right) &< 0.01\end{aligned}$$

Or d'après la table de la loi normale cela revient à considérer l'inégalité suivante :

$$\frac{2200/n - 60}{15}\sqrt{n} > 2.33$$

Or c'est la même chose que de considérer

$$\begin{aligned}2200 - 60n &> 2.33\sqrt{n}15 \\ 2200 - 60n - 34.95\sqrt{n} &> 0\end{aligned}$$

Il y a deux façons de résoudre cette inéquation en fonction de n. La première est de calculer la valeur de $2200 - 60n - 34.95\sqrt{n}$ pour quelques valeurs de n et de vérifier si la condition est respecté (> 0) :

n	$2200 - 60n - 34.95\sqrt{n}$	
20	843.7	> 0
25	525.25	> 0
30	208.57	> 0
35	-106.77	< 0
32	82.293	> 0
33	19.228	> 0
34	-43.792	< 0

on remarque que pour toutes les valeurs de $n < 33$ l'équation est vérifiée et que pour 34 et plus l'équation n'est pas vérifier. Ainsi le nombre maximum de personnes pour que le poids soit inférieur à 2200 avec une probabilité de 99% est de 33.

★★★★L'autre façon de trouver la valeur de n est de constater que l'équation

$$2200 - 60n - 34.95\sqrt{n} = 0$$

est une équation du second dégré de la forme

$$a\sqrt{n}^2 + b\sqrt{n} + c = 0$$

où $a = -60$, $b = -34.95$ et $c = 2200$.

En regardant vos notes de mathématique de secondaire 4 vous découvrirez que cette équation admet comme solutions[6]

$$\left\{\sqrt{n} = \frac{1}{2a}\left(-b + \sqrt{(b^2 - 4ac)}\right)\right\}$$
$$\left\{\sqrt{n} = \frac{1}{2a}\left(-b - \sqrt{(b^2 - 4ac)}\right)\right\}$$

soit

$$\sqrt{n} = \frac{1}{2(-60)}\left(-(-34.95) - \sqrt{\left((-34.95)^2 - 4(-60)(2200)\right)}\right) = 5.7711$$
$$\sqrt{n} = \frac{1}{2(-60)}\left(-(-34.95) + \sqrt{\left((-34.95)^2 - 4(-60)(2200)\right)}\right) = -6.3536$$

Evidemment $\sqrt{n}$ ne peut être négatif d'où la solution $\sqrt{n} = 5.7711$ ou $n = 33.306$. Puisqu'il faut un nombre entier de personnes il faut mettre une limite de 33 personnes.

9.3 Estimation par intervalle

Lorsqu'on est en présence d'une estimation ponctuelle le résultat c'est-à-dire l'estimation est une valeur qui devrait approcher le paramètre d'intérêt. Un échantillon de taille 10 qui donne une estimation de 11.7 pour la moyenne est intuitivement moins intéressant qu'un échantillon de 1000 qui donne une estimation de 11.2 pour la moyenne. Il faut intégrer cette notion pour obtenir une estimation qui donne une valeur et des bornes pour cette valeur. C'est la notion d'**intervalle de confiance**.

Définition 3.1 Soit θ un paramètre, un intervalle de confiance de niveau $1-\alpha$ pour θ est un intervalle aléatoire (B_i, B_s) tel que

$$\Pr(\theta \in (B_i, B_s)) = 1 - \alpha$$

La valeur α est généralement petite (0.1, 0.05 ou 1%) pour que la confiance $(1-\alpha)$ soit grande. L'intervalle de confiance s'interprète de la façon suivante : c'est certain à $(1-\alpha) * 100\%$ que la valeur réelle du paramètre se retrouve dans l'intervalle.

[6] Habituellement on parle d'équation du second degrés de la forme

$$ax^2 + bx + c = 0$$

et la solution donnée est pour x. Cela est équivalent à considérer $x = \sqrt{n}$ dans le problème.

Généralement, mais pas toujours, les intervalles de confiances sont symétriques par rapport à l'estimation ponctuelle et ainsi on obtient la formulation suivante :

$$\theta \in \left(\widehat{\theta} - E, \widehat{\theta} + E\right)$$

où E est la **marge d'erreur** (ou précision) de l'intervalle. Pour simplifier la notation on utilise parfois l'expression

$$\theta \in \left(\widehat{\theta} \pm E\right)$$

Lors de la présentation des résultats il est assez fréquent de rencontrer les deux formes, soient $\left(\widehat{\theta} - E, \widehat{\theta} + E\right)$ et $\left(\widehat{\theta} - E, \widehat{\theta} + E\right)$. Ce sont deux informations identiques quant à leur contenu mais complémentaires quant à l'interprétation.

Exemple 3.1 ★Un journal publie les résultats d'un sondage sur la satisfaction des citoyens envers leur administration municipale. L'article dit qu'il y a 33.4% des électeurs qui sont à tout le moins satisfait avec une erreur d'au plus 2.8 points de % et cela dans 19 cas sur 20.

On est en présence d'un intervalle de confiance de niveau 95% (19/20). La marge d'erreur est de 0.028 et l'intervalle est donné par

$$\begin{aligned} \pi &\in (0.334 - 0.028\ ,\ 0.334 + 0.028) \\ \pi &\in (0.306\ ,\ 0.362\ ) \end{aligned}$$

Construction d'un IC

La création de la formule qui donne un intervalle de confiance est assez simple dans son principe : il suffit de partir d'une relation connue qui implique le paramètre avec une probabilité fixée et de modifier les relations pour obtenir une relation sous la forme

$$\Pr\left(\widehat{\theta} \in (I_i; I_s)\right) = 1 - \alpha$$

Un technique efficace pour construire un I.C. pour un paramètre, est de prendre un estimateur du paramètre dont on connaît la distribution ou une approximation de celle-ci pour ensuite isoler le paramètre.

Exemple 3.2 ★★★Considérions une population et un paramètre tel que $E(X) = \mu, V(X) = \sigma^2$ et supposons un échantillon de taille $n \geq 30$.

La théorie de l'échantillonnage permet de dire que

$$\sqrt{n}\left(\overline{X} - \mu\right)/S \sim N(0,1).$$

On peut alors poser l'équation suivante qui découle directement de la définition des points critiques :

$$\Pr\left(-z_{\alpha/2} < \frac{\left(\overline{X} - \mu\right)}{S}\sqrt{n} < z_{\alpha/2}\right) \simeq 1 - \alpha$$

Si on isole le paramètre μ, on obtient

$$\Pr\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) \simeq 1 - \alpha$$

On obtient ainsi un intervalle aléatoire tel que la moyenne est dans cet intervalle avec une probabilité approximative de $1 - \alpha$. On a ainsi

$$\mu \in \left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

avec une probabilité de $1 - \alpha$.

Exemple 3.3 ★Un psychologue d'entreprise vient de construire un test pour évaluer le sentiment d'appartenance des employés à l'entreprise. Chez 34 employés il observe $\overline{x} = 90.3$ et $s = 15.7$. Quel est l'intervalle de confiance de niveau 90% pour le score moyen au test dans l'entreprise ?

Solution : On cherche un IC de niveau 90% pour μ. On a déterminé ci-haut cet intervalle approximatif :

$$\begin{aligned}
\mu &\in \left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) \\
\mu &\in \left(90.3 - 1.645\frac{15.7}{\sqrt{34}}; 90.3 + 1.645\frac{15.7}{\sqrt{34}}\right) \\
\mu &\in (85.871; 94.729)
\end{aligned}$$

avec confiance 90%. Cela veut dire qu'on est presque certain (avec une confiance de 90%) que la vraie valeur du paramètre est dans l'intervalle $(85.871; 94.729)$.

Principaux intervalles de confiances

On s'intéresse généralement aux paramètres π, μ et σ^2. Ce sont des caractéristiques qui résument bien la distribution des valeurs.

Le paramète π ne requiert que peu de conditions pour obtenir une bonne approximation de l'intervalle de confiance : le fait d'avoir un échantillon aléatoire et une taille d'échantillon assez grande par rapport aux probabilités à estimer ou le cas selon lequel la population est normale ou du moins très proche.

Pour le paramètre μ le premier donne un intervalle de confiance dont les bornes sont calculées en considérant une distribution normale des valeurs observées tandis que le deuxième est calculé selon une distribution quelconque des observations en considérant une taille d'échantillon relativement grande.

Finalement lintervalle de confiance pour la variance demande nécessairement une distribution normale des valeurs (observations).

Voici le tableau qui donne les formules pour ces intervalles de confiance :

Paramètre	**Hypothèses**	**IC** de niveau 1-α
π	$Bin(1,\pi)$ n grand	$\left(p - z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$
μ	$N(\mu,\sigma^2)$	$\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$
μ	n grand	$\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$
σ^2	$N(\mu,\sigma^2)$	$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right)$

La première et la troisième ligne du tableau donne des intervalles de confiance approximatifs tandis que les deux autres sont des intervalles exacts. Il faut toujours tenir compte des approximations dans l'interprétation des résultats. Si un intervalle de confiance est approximatif alors les bornes sont plus ou moins exactes selon le degré d'approximation et la taille de l'échantillon.

Remarque 3.1 Un intervalle de confiance est directement liée à la taille de l'échantillon et à la confiance. Si la confiance augmente, la longueur de l'intervalle augmente et si la taille de l'échantillon augmente alors la longueur de l'intervalle diminue. L'idée est alors de trouver une confiance raisonnable pour une présicion raisonnable.

Remarque 3.2 Il est possible de donner un intervalle de confiance de niveau 100% pour le paramèter π:

$$\pi \in [0,1]$$

Or cette réalité n'est d'aucune utilité puisque c'est trivial.

Exemple 3.4 ★Suite à un sondage auprès de 80 étudiants des CEGEP, il y en a 41.2% qui s'inscriront à l'UQTR. Pour faire une prédiction budgétaire, un IC de niveau 95% est construit pour le paramètre π, la probabilité qu'un étudiant s'inscrive. En se basant sur cet échantillon,

l'IC approximatif est donné par

$$\begin{aligned}
\pi &\in \left(p - z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}; p + z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right)\\
\pi &\in \left(0.412 - z_{0.025}\sqrt{\frac{0.412\,(1-0.412)}{80}}; 0.412 + z_{0.025}\sqrt{\frac{0.412\,(1-0.412)}{80}}\right)\\
\pi &\in \left(0.412 - 1.96\sqrt{\frac{0.412\,(1-0.412)}{80}}; 0.412 + 1.96\sqrt{\frac{0.412\,(1-0.412)}{80}}\right)\\
\pi &\in (.304\,14; .519\,86)
\end{aligned}$$

avec confiance 95%.

Présentation des intervalles de confiances

Lorsqu'un intervalle de confiance doit être calculé, certaines informations doivent être données pour que le lecteur soit en mesure de comprendre le cheminement et les limites de l'interprétation. Dans un premier temps le paramètre pour lequel un intervalle de confiance est défini et le niveau de confiance doit être donné et justifié . Le paramètre ne peut être donné que si on définit la v.a. et certaines de ses caractéristiques. On parle alors de la population de référence. Par la suite les hypothèses supplémentaires sur la population qui sont nécessaires pour évaluer l'IC doivent être données et éventuellement justifiées par l'analyse descriptive ou d'autres méthodes.

Précision d'un intervalle de confiance

Un intervalle de confiance est souvent de la forme $\widehat{\theta} \pm l/2$. La longueur de l'intervalle de confiance est l et sa précision[7] est $l/2$. Un intervalle de confiance pour le paramètre μ avec une taille d'échantillon de 17 et une distribution normale pour les observations a comme présicion

$$E = t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$$

La notion de précison est ce qui manquait à l'estimation ponctuelle pour interpréter correctement les résultats. Avec la notion de marge d'erreur (E) ou précision, il est possible de comparer deux études qui ont mesurer la même caractéristique mais avec des tailles d'échantillon différentes.

[7] Il y a des paramètres pour lesquels l'intervalle de confiance n'est pas symétrique par rapport à un estimateur. On garde cependant la même définition de l'intervalle de confiance : la longueur est l et la précision est $l/2$

9.4 Taille d'échantillon

La notion d'intervalle de confiance peut aussi être utilisée pour calculer la taille de l'échantillon nécessaire dans un problème en particulier pour obtenir au moins une précision donnée.

Proportion

Dans le cas de l'intervalle de confiance pour une proportion, la formule est

$$\pi \in \left(p - z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right)$$

ce qui veut dire que sa longueur est

$$2z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

et la précision est

$$E = z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

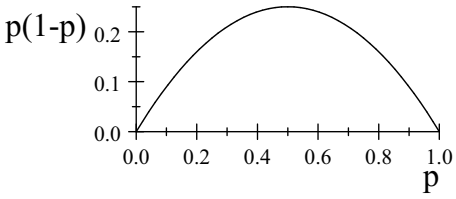
Si on veut déterminer la taille nécessaire pour obtenir au moins une certaine précision alors

$$n \geq \frac{\left(z_{\alpha/2}\right)^2 p\,(1-p)}{E^2}$$

Le signe d'inégalité provient du fait que si pour $n = 50$ on obtient la précision alors il en sera de même pour $n = 51$ ou pour toutes les valeurs plus grande que 50.

Dans cette formule il y a deux inconnues, n et p. Or p est la valeur qu'on observera si on prend un échantillon : elle ne peut être connue exactement puisqu'on veut justement déterminer le nombre d'unités à échantillonner pour évaluer p...

Une façon de contourner le problème est d'observer le graphique de $p\,(1-p)$ pour toutes les valeurs possibles de p:



Le pire des cas c'est-à-dire celui pour lequel la valeur est maximale est pour $p = 0.5$ cela veut dire que si on veut être certain dans tous les cas d'avoir une précision de E alors on remplace p dans la formule par 0.5. On a la solution suivante :

$$n \geq \frac{\left(z_{\alpha/2}\right)^2 0.5\,(1-0.5)}{E^2}$$

Exemple 4.1 ★Considérons le cas d'un sondage auprès de la population pour estimer la proportion de personnes en faveur de la légalisation du cannabis. Si on veut une précision de 3 points 19 fois sur 20, on a la formule suivante

$$n \geq \frac{\left(z_{\alpha/2}\right)^2 p\left(1-p\right)}{E^2} \tag{9.1}$$

avec $z_{0.025} = 1.96$. Puisqu'on ne sait pas la valeur de p a priori on utilisera 0.5 dans la formule :

$$n \geq 426\,8.4 \times 0.5\left(1-0.5\right) = 1067.1$$

On prendra $n = 1068$ puisque 1067 qui est plus petit que 1067.1 ne nous assure pas de la précision demandée dans tous les cas.

Cette méthode du "pire des cas" n'est pas toujours acceptable. Prenons l'exemple d'une étude sur le taux de restaurants qui ont fait une fraude sur la TVQ en 2007. On demande une précision de 0.01 sur le paramètre π en considérant une confiance de 95%. Selon la méthode du pire des cas la taille de l'échantillon requis est de

$$\begin{aligned} n &\geq \frac{\left(z_{\alpha/2}\right)^2 p\left(1-p\right)}{E^2} \\ &= \frac{1.95^2 0.25}{0.01^2} = 9506.3 \end{aligned}$$

soit au moins 9507 restaurants dont la comptabilité doit être inspectée.

Supposons qu'après enquête on observe un taux de 7%. Nous aurons une précision de

$$\begin{aligned} E &= z_{\alpha/2}\frac{\sqrt{p\left(1-p\right)}}{\sqrt{n}} \\ &= 1.96\frac{\sqrt{0.07 * 0.93}}{\sqrt{9507}} = 0.00513 \end{aligned}$$

C'est une précision beaucoup plus petite que celle demandée. Beaucoup de temps et d'argent auront été investis dans cette étude inutilement.

Pour palier ce problème on peut remettre en question la solution du "pire des cas" lorsqu'on a une information a priori sur le taux qu'on devrait observer lors de l'étude. Dans l'exemple du taux de fraude on peut avoir des raisons de penser qu'il est de l'ordre de 5%. En utilisant cette valeur dans la formune cela donne une taille de

$$n \geq \frac{\left(1.95\right)^2 0.05\left(0.95\right)}{0.01^2} = 1806.2$$

soit une taille de 1807, plus de 5 fois plus petite.

Une deuxième technique pour déterminer la taille de l'échantillon pour le paramètre π lorsque la valeur "prévue" est éloignée de 0.5 est de trouver des renseignements permettant d'avoir un ordre de grandeur pour la valeur puis d'utiliser cette information dans l'équation 9.1.

Cette technique a le désavantage de ne pas procurer une précision suffisante si l'a priori sur la

valeur de π est mauvaise.

Il existe une troisième technique qui permet de déterminer la taille par des essais/erreurs. Un premier échantillon de 100 unités est étudié. Ce dernier permet de calculer une approximation de π, soit p_{100}. On utilise cette valeur dans la formule 9.1 pour obtenir une taille d'échantillon. Cette dernière taille donne une seconde estimation de π qui sera utilisée pour déterminer si la précision est suffisante. Si ce n'est pas le cas il faut utiliser la dernière valeur de p dans la formule 9.1. Il est possible de continuer de la sorte jusqu'à ce que la précision désirée soit atteinte.

Chacune de ces méthodes a des avantages et des inconvénients. La première est la seule qui assure au moins la précision demandée mais elle peut aussi conduire à une taille trop grande inutilement. La deuxième peut facilement conduire à une précision insuffisante pour le problème. La troisième solution suppose qu'il est possible de faire l'enquête en plusieurs parties et ce n'est pas toujours réaliste.

Moyenne

La détermination de la taille de l'échantillon pour une précision donnée sur le paramètre μ est semblable au cas d'une proportion sauf que la technique du "pire des cas" n'est pas disponible.

L'intervalle de confiance pour le paramètre moyenne et un échantillon de grande taille est donné par

$$\mu \in \left(\overline{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}} \right)$$

ce qui donne comme précision

$$E = z_{\alpha/2} \frac{S}{\sqrt{n}}$$

Si on veut la taille d'échantillon nécessaire pour obtenir une précision donnée alors

$$n \geq \frac{z_{\alpha/2}^2 S^2}{E^2} \tag{9.2}$$

Or il y a deux inconnues : n et S. La première est nécessaire tandis que le deuxième pose un problème ; comment peut-on avoit la valeur de S avant de l'observer...

Une première solution est de chercher dans la littérature ou dans des enquêtes similaires pour obtenir une valeur réaliste de S. Par la suite il suffit de l'insérer dans l'équation 9.2 pour obtenir une taille d'échantillon.

S'il n'y a pas de données disponibles sur le comportement de S dans le contexte de l'enquête alors il est possible de prendre un échantillon relativement petit, disons $n = 20$ puis de calculer une valeur pour S. En utilisant la formule 9.2 on obtient une première approximation de la taille d'échantillon. Il est alors possible, après avoir observer d'autres unités statistiques d'obtenir

une taille plus précise.

Remarque 4.1 On utilise la formule avec le point critique $z_{\alpha/2}$ pour trouver la taille même si on a des raisons de penser que les observations sont normales. En fait cela provient du fait que si on utiliser la formule de l'intervalle de confiance pour population normale alors la précision est donnée par

$$E = \frac{t_{n-1;\alpha/2} S}{\sqrt{n}} \tag{9.3}$$

et il est impossible d'isoler n puisqu'il se retrouve au dénominateur et dans le point critique $t_{n-1;\alpha/2}$ et l'expression de ce dernier est une fonction très complexe de n.

S'il est certain que la population est normale et qu'on s'attend à une taille d'échantillon petite il est possible d'utiliser la formule 9.3 en itérant sur la valeur de n : on prend une valeur n arbitraire puis on calcule la précision. Si cette dernière n'est pas assez petite alors on prend une autre valeur plus grande. Si la précision donne une valeur plus petite que celle désirée alors on prend un n plus petit. Le processus est répété jusqu'à ce qu'on obtienne la précision désirée avec une taille n et que la valeur de E est trop grande pour une taille $n+1$.

Cette méthode est longue et demande l'utilisation d'un logiciel permettant de donner automatiquement la valeur de l'équation 9.3 comme Excel par exemple.

Remarque 4.2 Dans tous les cas la taille de l'échantillon n'est qu'une approximation puisque la valeur S qui est utilisée est approxmative. C'est pourquoi d'utiliser la formule 9.2 dans tous les cas est une stratégie raisonnable.

Exemple 4.2 ★Une entreprise de consultation informatique crée des pages web pour ses clients. Dans le but d'évaluer le nombre moyen de modifications que les clients demandent en cours de création, l'entreprise doit faire un échantillonnage sur les contrats déjà fait. Si la précision doit être de ± 1 pour un intervalle de confiance de niveau 95% sur le nombre moyen de modifications quelle est la taille d'échantillon nécessaire ?

Solution : Considérons X la v.a. qui donne le nombre de modifications lors de la création d'une page web et μ l'espérance de cette v.a..Pour avoir une précision de 1 il faut selon la formule 9.2 $n \geq 3.841\,6 S^2$. Supposons qu'une autre expérience permette de dire que S^2 est approximativement de 8 alors pour obtenir la précision demandée il faut $n \geq 30.733$. Il faut donc un échantillon de taille $n = 31$.

9.5 Résumé

Dans ce module on introduit la notion d'inférence statistique par le biais de l'estimation ponctuelle et l'estimation par intervalle. On a un échantillon c'est-à-dire une variable aléatoire qui se

répète n fois et on veut connaître les caractéristiques la variable aléatoire : $\mu,\ \pi$ ou σ^2. Les indices de moyenne et de variance sont les approximations de μ et σ^2 respectivement tandis que la proportion de succès dans l'échantillon est l'approximation du paramètre π.

Un élément fondamental de la statistique est le théorème central limite (TCL) qui permet de calculer la probabilité d'une moyenne ou d'un total en ayant uniquement μ et σ^2 comme information :

$$\frac{\overline{X}-\mu}{\sigma}\sqrt{n} \simeq N(0,1)$$

Cette relation (exacte si les mesures sont normales) permet de faire le calcul d'une probabilité sur $\overline{X}$, $T = n\overline{X}$ et $p = \overline{X}$.

L'estimation d'un paramètre est précisée avec la notion d'intervalle de confiance qui donne en plus de la valeur du paramètre, la marge d'erreur de l'approximation pour une confiance déterminé a priori. Pour évaluer un intervalle de confiance il suffit de choisir la bonne formule selon le paramètre d'intérêt :

Paramètre	**Hypothèses**	**IC** de niveau 1-α
π	$Bin(1,\pi)$ n grand	$\left(p - z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$
μ	$N(\mu,\sigma^2)$	$\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right)$
μ	n grand	$\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$
σ^2	$N(\mu,\sigma^2)$	$\left(\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right)$

Cette notion d'intervalle de confiance permet de donner la taille d'échantillon nécessaire pour obtenir une certaine précision dans l'estimation d'un paramètre par un intervalle de confiance. La formule pour le paramètre μ est donnée par $n \geq \frac{z^2_{\alpha/2}S^2}{E^2}$ tandis que celle pour le paramètre π est donnée par $n \geq \frac{z^2_{\alpha/2}p(1-p)}{E^2}$ où p est 0.5 s'il n'y a pas d'information et la valeur supposée si elle est donnée.