

Chapitre 2

Estimation ponctuelle et par intervalle

Christelle Dumas

Contents

1 Concepts d'estimation et application à la moyenne	4
1.1 Définitions	4
1.2 Critères d'évaluation d'un estimateur 1: le biais	5
1.3 Critères d'évaluation d'un estimateur 2: l'efficacité	7
1.3.1 Efficacité d'un estimateur sans biais	7
1.3.2 Propriétés des estimateurs	9
1.4 Application des concepts au cas de la moyenne d'un échantillon	9
1.5 Estimation par intervalle	10
1.6 Calcul d'un intervalle de confiance pour une moyenne	12
1.6.1 Principe	12
1.6.2 Application	13
2 Estimateur ponctuel et par intervalle de la variance, σ^2, de la population	14
2.1 Rappels du chapitre précédent	14
2.1.1 Cas de la moyenne connue	14
2.1.2 Cas de la moyenne inconnue	14
2.2 Propriétés de l'estimateur de la variance	15
2.2.1 Variance de l'estimateur de la variance	15
2.2.2 Propriétés de convergence	16
2.2.3 Quid de la variance empirique non modifiée?	16
2.3 Distribution de l'estimateur de la variance	16
2.3.1 Distribution de l'estimateur de la variance pour une population normale de moyenne inconnue	17

2.3.2	Cas particulier où la moyenne est connue (et la population normale)	18
2.3.3	Distribution de l'estimateur de la variance pour une population de loi inconnue	18
2.3.4	Résumé des résultats	19
2.4	Construction d'un intervalle de confiance pour la variance	19
2.5	Application	20
3	Estimateur de la moyenne dans le cas où la variance est inconnue	21
3.1	Distribution de l'estimateur de la moyenne quand σ est inconnue	21
3.1.1	Cas d'une population normale	21
3.1.2	Cas d'une population quelconque avec $n \geq 30$	23
3.1.3	Résumé	23
3.2	Application et construction d'un intervalle de confiance	23
4	Estimateur de la proportion	24
4.1	Propriétés de l'estimateur de la proportion	24
4.2	Estimation par intervalle de la proportion	24
4.3	Application	25

Nous avons, lors du chapitre précédent, trouvé des outils pour décrire un échantillon. Mais incidemment, nous avons aussi déjà mis en évidence des résultats d'estimation sur lesquels nous allons revenir. L'objet de ce chapitre est de formaliser un petit peu ce que nous venons de voir et de le généraliser. Nous allons donc commencer par revenir sur la notion d'estimateur et d'estimation et les définir proprement.

L'enjeu de ceci est le suivant: comme nous l'avons déjà mentionné, la statistique inférentielle est le procédé qui consiste à obtenir des informations sur une population sur la base d'un échantillon. Nous avons déjà donné un exemple de ce procédé: nous avons vu que si nous calculions la moyenne d'un échantillon, alors nous avons de bonnes chances d'être proche de la moyenne pour la population entière. Ceci s'appelle "estimer" la moyenne de la population. C'est par exemple ce que l'on fait lorsqu'on fait une étude de marché: on essaie de savoir si le produit va plaire et donc sera acheté en posant la question à un plus petit nombre d'individus. Par ailleurs, rappelez-vous que ce que nous avons déjà vu pour la moyenne était un cas particulier; l'objectif ici est aussi d'étendre les résultats lorsque les hypothèses que nous avons faites ne sont pas valables.

Mais pour être complet, il faut aussi rappeler que le calcul étant fait sur un échantillon, il y a des incertitudes, qu'on appelle erreurs d'échantillonnage. Donc par exemple, si l'on estime qu'environ 55% de la population des mangeurs de burgers serait prêt à acheter un nouveau McTruc, on n'est pas sûr que l'objectif de 50% soit atteint. La théorie de l'estimation a pour objet de traiter de ces erreurs d'échantillonnage. Le fabricant de burgers peut vouloir savoir quel est le risque que son objectif de 50% ne soit pas atteint. Au lieu de fournir le chiffre de 55% qui est une estimation que l'on appellera ponctuelle de la proportion d'acheteurs potentiels du nouveau sandwich, on peut fournir une fourchette, par exemple 49%-61% dans laquelle on est sûrs à 95% que la proportion réelle d'acheteurs se trouve. On appellera ça une estimation par intervalle. On verra aussi que souvent plusieurs estimateurs d'une même quantité sont disponibles; la théorie de l'estimation permet de choisir entre les estimateurs.

1 Concepts d'estimation et application à la moyenne

1.1 Définitions

Commençons par définir un estimateur:

Définition 1 *Un estimateur (ponctuel) $\hat{\Theta}_n$ d'un paramètre inconnu θ est une statistique $f(X_1, \dots, X_n)$ qui, quel que soit l'échantillon, doit prendre des valeurs proches de θ .*

Définition 2 *Une estimation (ponctuelle) $\hat{\theta}_n$ d'un paramètre inconnu θ est une valeur $\hat{\theta}_n = f(x_1, \dots, x_n)$ prise par un estimateur $\hat{\theta}_n$ de θ sur un échantillon particulier (x_1, \dots, x_n)*

Qu'est-ce que θ ? θ est un paramètre qui représente, d'une façon ou d'une autre, la population. On veut connaître sa valeur pour avoir une information sur la population. θ peut être beaucoup de choses:

- la moyenne d'une population: ex, le salaire moyen des travailleurs français;
- la proportion d'une population: ex, la proportion d'acheteurs de burgers, la proportion d'actifs au chômage;
- la variance de population: ex, la variance du salaire (ça donne une mesure des inégalités salariales);
- l'effet d'une variable sur une autre (ex: effet du diplôme sur le salaire) dans le chapitre économétrie;
- mais aussi des choses plus compliquées que l'on n'apprendra pas à estimer dans ce cours comme les paramètres de certaines lois (ex: taux d'arrivée d'offres de travail)...

Une définition floue Comme vous le remarquez peut-être, la définition d'un estimateur est relativement floue: que veut dire "prendre des valeurs proches"? Si par exemple, je m'amuse à estimer la moyenne d'une population en prenant la moyenne échantillonnale -1. Est-ce que ça serait suffisamment proche? peut-être que oui si on mesure le nombre moyen de kilomètres effectués en voiture d'un francilien par an car ce nombre est très grand. Mais si on mesure le nombre de mètres effectués par un escargot en une journée, l'estimation va être assez loin de la valeur moyenne.

Important Par ailleurs, vous noterez que ce qui compte, c'est la valeur que l'on obtient quel que soit l'échantillon. Il ne s'agit donc pas de savoir si la valeur qu'on obtient sur un échantillon est proche mais de savoir si le procédé qui consiste à calculer cette statistique donnera, dans l'absolu, des valeurs proches du paramètre que l'on cherche à calculer. Rappel: un estimateur, en tant que statistique, est une variable aléatoire.

Enfin, notez la différence entre un estimateur (statistique et donc mode de calcul pour effectuer une estimation) et une estimation (valeur obtenue). Même si, *in fine*, on veut être en mesure d'obtenir des estimations de certains paramètres, ce qui va nous intéresser dans ce chapitre sont plutôt les estimateurs et leurs propriétés. En bref, il faut commencer par choisir la méthode avant de faire le calcul (ou pour le dire autrement: réfléchir avant de se ruer sur sa calculatrice).

Exemple: nous avons déjà rappelé que la moyenne d'échantillon est un estimateur de la moyenne de la population. $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur de $E(X)$ tandis que $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ en est une estimation. Mais par abus d'écriture, on utilise souvent $\hat{\theta}$ à la fois pour un estimateur et une estimation de θ .

1.2 Critères d'évaluation d'un estimateur 1: le biais

La définition d'un estimateur étant relativement floue, il convient de proposer des critères d'évaluation des estimateurs afin de choisir entre différentes statistiques.

Prenons un exemple: considérons trois estimateurs concurrents de la moyenne de la population m construits à partir d'un échantillon aléatoire de trois observations (taille $n = 3$).

$$\begin{aligned} T_n^1 &= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3 \\ T_n^2 &= \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3 \\ T_n^3 &= \frac{1}{6}X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3 \end{aligned}$$

Soit une population de 5 individus dont le caractère (taille, par exemple) vaut: 165, 166, 168, 170, 172. On peut calculer les trois différentes estimations de cette moyenne à l'aide des trois estimateurs proposés sur différents échantillons de taille 3 comme on l'a fait au chapitre précédent. Comment choisir entre ces différents estimateurs? un premier critère qui semble évident est que l'on veut qu'en espérance, l'estimateur soit égal à la valeur θ inconnue. Ceci revient à demander que l'estimateur soit sans biais:

Définition 3 $\hat{\Theta}_n$ est un *estimateur non biaisé* (ou sans biais) du paramètre $\theta \Leftrightarrow E(\hat{\Theta}_n) = \theta$.

En fait, si l'on reprend l'exemple précédent, on peut très facilement montrer que les 3 estimateurs proposés sont des estimateurs sans biais de la moyenne. En effet:

$$\begin{aligned} E(T_n^1) &= E\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) = \mu \\ E(T_n^2) &= E\left(\frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3\right) = \frac{1}{2}E(X_1) + \frac{1}{4}E(X_2) + \frac{1}{4}E(X_3) = \mu \\ E(T_n^3) &= E\left(\frac{1}{6}X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3\right) = \frac{1}{6}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{3}E(X_3) = \mu \end{aligned}$$

De ce point de vue, il faudra proposer un autre critère d'évaluation, que nous verrons plus loin.

Mais avant cela, demandons-nous si nous connaissons des estimateurs qui avaient l'air a priori acceptables mais qui se sont révélés être biaisés. Au chapitre précédent, nous avons considéré la variance empirique comme estimateur de la variance de la population, or nous avons montré que:

$$E(\tilde{S}_n) = \frac{n-1}{n}\sigma^2$$

la variance empirique non modifiée est donc un estimateur biaisé de σ^2 .

Le biais On appelle **biais** la différence entre la valeur espérée de l'estimateur et le paramètre que l'on souhaite définir:

$$B(\theta) = E[\hat{\Theta}_n] - \theta$$

(bien sûr, le biais vaut zéro si l'estimateur est non biaisé). Contrairement à l'erreur d'échantillonnage qui est aléatoire, **le biais est une erreur systématique**. Notez que le biais $B(\theta)$ est une valeur certaine (bien que généralement inconnue) et non pas aléatoire.

Graphique pour représenter un estimateur non biaisé et un estimateur biaisé.

Si l'on revient à l'exemple de la variance empirique non modifiée, le biais vaut:

$$B(\sigma^2) = E(\tilde{S}_n^2) - \sigma^2 = \left(\frac{n-1}{n} - 1\right)\sigma^2 = -\frac{1}{n} \cdot \sigma^2$$

ce qui signifie que la variance empirique sous-estime systématiquement la variance de la population. Mais on peut aussi remarquer que lorsque n devient grand, ce qui est généralement le cas en statistique, le biais tend vers 0. Avec un échantillon de grande taille, on peut donc utiliser un estimateur biaisé à condition que le biais tende vers 0 lorsque n tend vers l'infini. C'est la définition d'un estimateur asymptotiquement sans biais.

Définition 4 $\hat{\Theta}_n$ est un *estimateur asymptotiquement non biaisé* (ou sans biais) du paramètre $\theta \Leftrightarrow \lim_{n \rightarrow +\infty} E(\hat{\Theta}_n) = \theta$.

La variance empirique non modifiée est donc un estimateur asymptotiquement sans biais de la variance σ^2 . La proposition suivante est immédiate:

Proposition 5 Si un estimateur est sans biais alors il est asymptotiquement sans biais.

Par contraste avec un estimateur asymptotiquement sans biais, on appelle aussi un estimateur sans biais un **estimateur sans biais à distance finie**. À noter: si l'échantillon est petit, il n'est pas intéressant d'utiliser un estimateur biaisé à distance finie mais asymptotiquement sans biais.

On utilise assez rarement des estimateurs biaisés asymptotiquement. Ce serait néanmoins le cas si j'estimais le nombre de kms effectués en voiture par un francilien en 1 an en utilisant la moyenne sur un échantillon aléatoire à laquelle j'ôte 1. Dans ce cas là, le biais est de -1 et ne diminue pas en augmentant la taille de l'échantillon. L'estimateur est donc asymptotiquement biaisé.

1.3 Critères d'évaluation d'un estimateur 2: l'efficacité

Nous avons vu, avec l'exemple des 3 moyennes pondérées de façon différente que l'on pouvait avoir plusieurs estimateurs sans biais d'un même paramètre. Peut-être ces estimateurs ne sont-ils pas tous aussi bons et il vaut mieux en retenir un plutôt qu'un autre. Il nous faut donc des critères supplémentaires d'évaluation d'un estimateur. Dans ce cours, on ne traitera que le cas des estimateurs asymptotiquement sans biais.

1.3.1 Efficacité d'un estimateur sans biais

Dans le cas des trois moyennes, on pourrait résumer les différences entre les estimations de la façon suivante: dans un cas (le premier), on prend en compte les observations de la même façon (pondération égale), tandis que dans les deux autres, certaines observations sont plus prises en compte que d'autres. Quelle approche vous semble la plus informative dans ce cas-là?

En fait, l'utilisation de la totalité de l'information permet d'obtenir des résultats plus précis ainsi que le montrent les variances des estimateurs ¹ :

$$V(T_n^1) = V\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \frac{1}{9}V(X_1) + \frac{1}{9}V(X_2) + \frac{1}{9}V(X_3) = \frac{1}{3}\sigma^2$$

¹Rappel : $V(T_n^i) = \frac{1}{n^2} \sum_{i=1}^n V(X_i)$, car nous supposons les variables *i.i.d.* (indépendantes, identiquement distribuées) et donc la variance d'une somme est égale à la somme des variances.

$$V(T_n^2) = V\left(\frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_4\right) = \frac{1}{4}V(X_1) + \frac{1}{16}V(X_2) + \frac{1}{16}V(X_3) = \frac{3}{8}\sigma^2$$

$$V(T_n^3) = V\left(\frac{1}{6}X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3\right) = \frac{1}{36}V(X_1) + \frac{1}{4}V(X_2) + \frac{1}{9}V(X_3) = \frac{7}{18}\sigma^2$$

Ceci suggère que l'on peut ajouter ce critère.

Définition 6 *Entre deux estimateurs sans biais T_n^1 et T_n^2 du même paramètre θ , on dit que l'estimateur le plus efficace est celui dont l'erreur aléatoire est la plus faible, i.e. celui dont la variance est la plus faible.*

Si $V(T_n^2) < V(T_n^1)$ alors $T_n^2 \succ T_n^1$ (T_n^2 est préféré à T_n^1).

Que représente l'efficacité? Imposer que l'estimateur soit sans biais signifie que si l'on prenait un grand nombre d'échantillons, on devrait retrouver le paramètre d'intérêt. Ceci ne dit pas si lorsqu'on prend un seul échantillon, le paramètre estimé est très proche du vrai paramètre. Si par contre on sait que la variance de l'estimateur est petite, alors on a peu de chances de s'écarter de la vraie valeur du paramètre. Un estimateur avec une petite variance autour de la vraie valeur θ est donc préférable à un estimateur avec une grande variance autour de θ . Graphe.

Dans l'exemple précédent, $V(T_n^1) < V(T_n^2) < V(T_n^3)$ et donc le premier estimateur est le meilleur estimateur des trois. Il convient de remarquer que la moyenne arithmétique de l'échantillon est le meilleur estimateur.

En pratique, ce que l'on cherche à faire généralement est construire directement un estimateur sans biais et efficace.

Définition 7 *On dit d'un estimateur sans biais qu'il est efficace s'il n'existe pas d'autre estimateur sans biais dont la variance est plus petite:*

T_n^ est efficace si et seulement si:*

$$E(T_n^*) = \theta \quad \text{et} \quad [\forall T_n \text{ tel que } E(T_n) = \theta, \quad V(T_n^*) \leq V(T_n)]$$

Adoptons cette démarche et considérons maintenant, l'estimateur T_n de la forme $T_n = aX_1 + bX_2$ et construisons cet estimateur de façon à ce qu'il soit sans biais et que sa variance soit minimale. Calculons l'espérance et la variance de cet estimateur, il vient :

$$\begin{aligned} E(T_n) &= E(aX_1 + bX_2) = aE(X_1) + bE(X_2) = (a + b)\mu \\ V(T_n) &= V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) = (a^2 + b^2)\sigma^2 \end{aligned}$$

Il s'agit donc de résoudre, *in fine*, le programme de minimisation suivant :

$$\begin{cases} \min_{a,b} a^2 + b^2 \\ \text{s.c. } a + b = 1 \end{cases}$$

Il vient $a = b = 1/2$. Le meilleur estimateur est donc :

$$T_n = \frac{1}{2}X_1 + \frac{1}{2}X_2$$

Le meilleur estimateur de μ est donc la variable aléatoire moyenne d'échantillon \bar{X}_n . Dans l'exemple précédent, la variance minimale est alors $\frac{1}{4}\sigma^2$.

Borne inférieure de la variance Vous remarquerez que même en choisissant l'estimateur qui a la variance minimale, celle-ci n'est pas égale à 0. C'est logique, quel que soit l'estimateur choisi, il y a toujours une erreur d'échantillonnage qui fait que l'on ne peut estimer le paramètre avec une totale certitude (sur un échantillon). En fait, il y a une borne inférieure pour la variance (appelée borne de Ramer-Crao), c'est-à-dire qu'on ne peut indéfiniment diminuer la variance des estimateurs.

1.3.2 Propriétés des estimateurs

Définition 8 *Un estimateur sans biais (asymptotiquement ou à distance finie) dont la variance tend vers 0 est un **estimateur convergent**:*

$$\lim_{n \rightarrow \infty} V(\hat{\Theta}_n) = 0$$

Plus n augmente et plus la loi de l'estimateur est concentrée autour du vrai paramètre inconnu. Ainsi, plus n augmente, moins, il y a de chance que l'on s'éloigne de la vraie valeur. La variabilité de l'estimateur T_n est mesurée par sa variance et représente l'erreur aléatoire (par opposition au biais qui représente l'erreur systématique).

Cette propriété peut être illustrée graphiquement - Inclure graphique.

1.4 Application des concepts au cas de la moyenne d'un échantillon

Dans le chapitre précédent, nous avons déjà dérivé un ensemble important de résultats sur la moyenne échantillonnale. Nous avons notamment caractérisé sa distribution dans le cas où:

- σ connue;
- la population est normale ou la taille de l'échantillon est grande.

Prenons les propriétés des estimateurs et regardons si elles sont vérifiées par la moyenne d'échantillon, en tant qu'estimateur de la moyenne (de la population) (i.e. \bar{X}_n estimateur de μ).

L'estimateur de moyenne est-il sans biais? On a montré que $E(\bar{X}_n) = E(X) = \mu$. Ceci prouve donc bien que l'estimateur de la moyenne est sans biais à distance finie. Par conséquent, il est aussi asymptotiquement sans biais.

L'estimateur de la moyenne est-il à variance minimale? On a montré sur un exemple simple ($T_n = \frac{1}{2}X_1 + \frac{1}{2}X_2$) que la moyenne d'un échantillon de taille 2 est l'estimateur linéaire de plus petite variance. Cette démonstration est généralisable pour n quelconque.

Ceci dit, il n'est pas inutile de comparer le cas du tirage avec remise et celui du tirage sans remise. En effet, nous nous souvenons que dans le cas du tirage avec remise, la variance de la moyenne échantillonnale (ie variance de l'estimateur) valait σ^2/n alors que dans le cas du tirage sans remise, elle vaut $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$. Pour comparer ces deux valeurs, il suffit de comparer $\frac{N-n}{N-1}$ à 1. Or:

$$n \geq 1 \Leftrightarrow N - n \leq N - 1 \Leftrightarrow \frac{N - n}{N - 1} \leq 1 \Leftrightarrow \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1} \leq \frac{\sigma^2}{n}$$

La précision de l'estimateur de la moyenne est donc meilleur dans une tirage sans remise que dans un tirage avec remise. Ceci est logique: pour un tirage avec remise, il faut prendre en compte le fait qu'on peut avoir tiré deux fois la même observation qui, la seconde fois n'apporte en théorie pas d'information supplémentaire.

Donc à choisir entre deux modes de tirage, il vaut mieux prendre un tirage sans remise pour estimer une moyenne puisque celui-ci permettra d'avoir une estimation plus précise.

L'estimateur de la moyenne est-il convergent? On a vu que la variance de l'estimateur de la moyenne est $\frac{\sigma^2}{n}$ dans le cas avec remise et $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$ dans le cas sans remise. Cependant, lorsque n est grand, $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \sim \frac{\sigma^2}{n}$, ce qui nous ramène au cas avec remise.

Or nous avons vu que si l'estimateur est centré, alors il est convergent si et seulement si $\lim_{n \rightarrow \infty} V(T_n) = 0$. Dans notre cas, T_n est la moyenne échantillonnale et, pour n grand, $V(\bar{X}_n) = \sigma^2/n$ dont la limite vaut 0 quand n tend vers l'infini. L'estimateur de la moyenne est donc bien convergent.

1.5 Estimation par intervalle

À quoi sert l'estimation par intervalle? L'estimation ponctuelle n'est pas pleinement satisfaisante car même si on sait qu'un estimateur est sans biais et convergent, on ne sait pas si pour l'échantillon particulier dont on dispose, l'estimation est proche de la vraie valeur du paramètre. Par exemple, si on calcule une moyenne sur un échantillon, on est quasiment sûr de se tromper

en disant que c'est exactement la moyenne de la population: l'estimateur de la moyenne ne vaut a priori pas la vraie valeur de la moyenne (ie la moyenne de la population). De par ses propriétés d'absence de biais et de convergence, on sait que, en moyenne l'estimation obtenue va être proche de la vraie valeur du paramètre et que la précision va être de plus en plus grande lorsque la taille de l'échantillon augmente, mais:

- on ne dispose que d'un échantillon
- et il est de taille finie

donc ces propriétés bien que nécessaires, sont loin d'être suffisantes pour garantir que l'estimation ponctuelle que nous sommes en train de calculer est suffisamment précise pour qu'on puisse s'en servir. Ainsi, si l'on sait qu'un estimateur est sans biais, on veut quand même avoir une notion de l'ampleur de l'erreur aléatoire ou erreur d'échantillonnage pour savoir à quel point on peut avoir confiance dans l'estimation obtenue. Pour cela, on peut faire appel au concept d'estimation par intervalle.

Définition 9 *L'estimation par intervalle de confiance d'un paramètre θ de valeur inconnue consiste à calculer à partir d'un estimateur $\hat{\theta}$ un intervalle dans lequel il est plausible de trouver la vraie valeur du paramètre (θ). Cet intervalle est défini de la façon suivante par deux limite aléatoires LI et LS en tenant compte d'une probabilité $1 - \alpha$ donnée a priori et aussi élevée que l'on désire:*

$$P(LI \leq \theta \leq LS) = 1 - \alpha$$

LI est la variable aléatoire "limite inférieure de l'intervalle" et LS est la variable aléatoire "limite supérieure de l'intervalle".

Pour un échantillon particulier, on obtient la réalisation de l'intervalle en calculant les valeurs numériques prises par LI et LS . Cet intervalle est accompagné d'un niveau de confiance égal à $1 - \alpha$. On dit que l'intervalle aléatoire $[LI, LS]$ (ou même sa réalisation $[li, ls]$, qui n'est pas aléatoire) est un intervalle de confiance pour θ de probabilité (ou de niveau de confiance) $1 - \alpha$.

Remarque 1 L'intervalle ainsi défini est aléatoire puisque des limites sont des variables aléatoires qui prendront des valeurs numériques fonction des observations de l'échantillon; ces valeurs seront différentes d'un échantillon à l'autre. Toutefois, la probabilité qu'un intervalle donné englobe la vraie valeur du paramètre est $1 - \alpha$. Autrement dit, si on affirme que l'intervalle $[li, ls]$ contient la valeur du paramètre, on ne se trompe en moyenne qu' α fois sur 100.

Remarque 2 Il y a un arbitrage entre la taille de l'intervalle et le niveau de confiance demandé. Si on veut éviter de se tromper dans ses conclusions (on veut un niveau de confiance élevé, 99% par exemple), on sera obligé d'avoir un intervalle de confiance très large. À l'opposé, un intervalle de confiance pourra être petit et donc plus précis si on ne requiert pas un fort niveau de confiance.

Remarque 3 Le niveau de confiance est noté $1-\alpha$. Il accompagne nécessairement toute réalisation d'un intervalle de confiance dont les limites ne sont pas des variables aléatoires mais des valeurs numériques. Par conséquent, le paramètre est ou n'est pas entre ces deux limites; mais il y a une probabilité d'erreur lorsque nous affirmons qu'il y est. Par ailleurs, puisque θ n'est pas une variable aléatoire mais une valeur certaine (bien qu'inconnue), il est faux de dire que la vraie valeur du paramètre a, disons, 95% de chances de se trouver dans l'intervalle. Il faut plutôt dire que l'intervalle a 95% de chances de recouvrir la valeur exacte du paramètre.

Remarque 4 α prend généralement de petites valeurs (habituellement 10%, 5% ou 1%), il désigne le risque que la prévision effectuée par l'intervalle de confiance ne se réalise pas.

1.6 Calcul d'un intervalle de confiance pour une moyenne

Présentons la procédure habituelle pour calculer un intervalle de confiance en construisant l'intervalle de confiance pour μ dans le cas où la population est normale et de variance connue.

1.6.1 Principe

On veut trouver deux statistiques LI et LS telles que

$$P(LI \leq \mu \leq LS) = 1 - \alpha$$

Choix du meilleur estimateur ponctuel Pour obtenir un intervalle de confiance le plus précis possible à α donné, il convient de prendre le meilleur estimateur ponctuel. Nous avons précédemment vu que le meilleur estimateur de μ est la moyenne d'échantillon \bar{X} .

Loi de probabilité de l'estimateur On sait que si la population obéit à une loi normale dont on connaît la variance, les fluctuations de l'écart-réduit $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ suivent une loi normale centrée réduite.

Calcul de l'intervalle En partant de cela, on construit l'intervalle aléatoire

$$\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

qui a une probabilité $1 - \alpha$ de contenir la vraie valeur de μ et ce avant même que l'échantillon soit prélevé. En effet:

$$\begin{aligned} P\left(\bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) &= P\left(-z_{\alpha/2} \leq \frac{\mu - \bar{X}_n}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \\ &= 1 - \alpha \end{aligned}$$

1.6.2 Application

Après 9 semaines de cours, en allant à Cergy 3 fois par semaine environ, vous disposez de 27 observations pour le temps d'attente du RER le matin. Vous avez donc calculé la moyenne du temps d'attente et trouvé qu'elle était égale à 7mn. La SNCF par ailleurs vous informe que l'écart-type est de 4mn. L'objectif étant d'arriver à l'heure en cours, vous voulez calculer un intervalle de confiance à 90%, 95% et 99% pour le temps moyen d'attente du RER.

Nous avons déjà rappelé comment estimer la moyenne d'une population en utilisant la moyenne échantillonnale, appelée plus couramment estimateur de la moyenne. On a vu que nous disposions de la totalité des résultats dans le cas où la variance de la population était connue. Nous reviendrons sur le cas où elle n'est pas connue après avoir montré comment on peut estimer la variance d'une population. L'objet de cette deuxième partie du chapitre 2 est de fournir des estimations ponctuelles et par intervalle des moments d'une population ainsi que de discuter les propriétés des estimateurs.

2 Estimateur ponctuel et par intervalle de la variance, σ^2 , de la population

Dans un souci de simplicité, on ne traitera que le cas du tirage avec remise.

2.1 Rappels du chapitre précédent

2.1.1 Cas de la moyenne connue

Dans le chapitre précédent, nous avons montré que si la moyenne μ était connue, alors la variance échantillonnale semi-empirique, définie comme suit:

$$\mathcal{V}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

avait la propriété suivante:

$$E(\mathcal{V}_n) = \sigma^2$$

On peut donc dire que, dans le cas où la moyenne est connue (i.e. certaine, i.e. non estimée), alors la variance semi-empirique est un estimateur sans biais de la variance.

2.1.2 Cas de la moyenne inconnue

Ce cas est nettement plus intéressant et nous avons déjà montré que la variance empirique non modifiée, définie comme suit:

$$\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X}_n)^2$$

était un estimateur biaisé de la variance puisque:

$$E(\tilde{S}_n^2) = \frac{n-1}{n} \cdot \sigma^2$$

Cependant, nous avons noté que la variance empirique non modifiée, bien que biaisée à distance finie est sans biais asymptotiquement, puisque:

$$B(\sigma^2) = \frac{n-1}{n} \cdot \sigma^2 - \sigma^2 = -\frac{1}{n} \cdot \sigma^2 \xrightarrow{n \rightarrow \infty} 0$$

Dans la mesure où nous n'avons pour l'instant pas calculé la variance de l'estimateur, nous ne pouvons pas dire si cet estimateur est convergent ou non.

Passons maintenant au cas de la variance empirique modifiée, dont la définition est rappelée ici:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Puisque $E(S_n^2) = \sigma^2$, la variance empirique modifiée est un estimateur sans biais à distance finie de la variance. De ce point de vue, il est donc préférable à la variance empirique non modifiée.

Nous allons dorénavant nous concentrer sur ce dernier estimateur et allons dériver ses propriétés de façon à:

- être en mesure de dépasser le stade de la simple estimation, notamment pouvoir donner une mesure de la précision de l'estimateur, c'est-à-dire fournir une estimation par intervalle de confiance;
- dériver le cas où la variance est inconnue pour l'estimateur de la moyenne.

2.2 Propriétés de l'estimateur de la variance

2.2.1 Variance de l'estimateur de la variance

Nous avons déjà calculé le premier moment (espérance) de l'estimateur de la variance qu'est la variance empirique modifiée et montré qu'il était égal à la variance. Passons maintenant au moment centré d'ordre 2 qu'est la variance. La variance de la variance d'échantillonnage S_n^2 s'écrit:

$$V(S_n^2) = E[(S_n^2 - E(S_n^2))^2] = E[(S_n^2)^2] - [E(S_n^2)]^2$$

Par souci de simplicité, nous omettons les calculs et obtenons la formule suivante pour n grand:

$$V(S_n^2) \simeq \frac{1}{n} \cdot (\mu_4 - \sigma^4)$$

où μ_4 est le moment centré d'ordre 4, à savoir:

$$\begin{aligned} \mu_4 &= E[(X - E(X))^4] \\ &= m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4 \end{aligned}$$

avec m_k moment d'ordre k ($m_k = E(X^k)$).

On a donc caractérisé la variance (ou erreur aléatoire) de l'estimateur de la variance à partir de caractéristiques propres à la population.

Cas d'une population normale Dans le cas où la population est normale ($X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$), la formule pour $V(S_n^2)$ se simplifie en:

$$V(S_n^2) = \frac{2\sigma^4}{n-1}$$

2.2.2 Propriétés de convergence

Nous sommes maintenant en mesure de conclure sur les propriétés de l'estimateur de la variance, et notamment de savoir s'il est convergent. On a déjà dit qu'il était sans biais, il suffit donc de regarder si sa variance tend vers 0 lorsque n devient très grand.

Comme les valeurs m_k sont fixées et ne dépendent pas de l'échantillon ni de sa taille n . Comme $\frac{1}{n} \rightarrow_{n \rightarrow \infty} 0$, on trouve que:

$$\lim_{n \rightarrow \infty} V(S_n^2) = 0$$

et l'estimateur de la variance est donc convergent. Pour rappel, ceci signifie que la précision de l'estimateur va être de plus en plus grande (ou l'erreur d'estimation de plus en plus petite) à mesure que la taille de l'échantillon croît. Pour n infini, l'estimation est exacte.

2.2.3 Quid de la variance empirique non modifiée?

On obtient très facilement les mêmes propriétés pour la variance empirique non modifiée puisque $\tilde{S}_n^2 = \frac{n-1}{n} \cdot S_n^2$, en effet:

$$V(\tilde{S}_n^2) = V\left(\frac{n-1}{n} \cdot S_n^2\right) = \left(\frac{n-1}{n}\right)^2 V(S_n^2)$$

or, quand n tend vers l'infini, la fraction $\left(\frac{n-1}{n}\right)^2$ tend vers 1 et donc les deux variances sont semblables. La variance empirique non modifiée est donc un estimateur asymptotiquement sans biais et convergent. Pour un grand nombre d'observations, il est totalement équivalent d'utiliser l'un ou l'autre des estimateurs.

2.3 Distribution de l'estimateur de la variance

Souvenez-vous, au chapitre 1, nous avons caractérisé la loi de probabilité de l'estimateur de la moyenne et ceci avait été très utile pour trouver des informations sur le degré de précision de l'estimation de la moyenne. Notamment, cela permettait de dire entre quelle et quelle valeur la moyenne devait être comprise avec une probabilité de 95%. Jusqu'ici, nous n'avons pas d'information comparable pour l'estimateur de la variance et même si l'on connaît l'erreur aléatoire

de cet estimateur, cela ne suffit pas pour donner des encadrements de la vraie valeur avec une probabilité associée à l'encadrement. De la même façon, on va chercher à caractériser la loi de l'estimateur de la variance, après avoir décrit ses moments. Nous ne traiterons que le cas où la population est distribuée normalement.

2.3.1 Distribution de l'estimateur de la variance pour une population normale de moyenne inconnue

Commençons par considérer la statistique suivante $\frac{n-1}{\sigma^2} \cdot S_n^2$. Si l'on substitue S_n par sa valeur:

$$\frac{n-1}{\sigma^2} \cdot S_n^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n \left[\frac{X_i - \bar{X}_n}{\sigma} \right]^2$$

or

$$\frac{X_i - \bar{X}_n}{\sigma} \rightsquigarrow \mathcal{N}(0, 1).$$

Rappel 10 Si Y_1, \dots, Y_k sont k variables aléatoires normales centrées réduites alors

$$Y_1^2 + Y_2^2 + \dots + Y_k^2 \rightsquigarrow \chi_\nu^2$$

où ν est le nombre de degrés de liberté des variables (Y_1, \dots, Y_k) . Si les k variables sont indépendantes alors $\nu = k$.

Dans notre cas, on peut définir

$$Y_i = \frac{X_i - \bar{X}_n}{\sigma}.$$

Y_i suit une normale centrée réduite donc la somme des Y_i^2 va suivre une χ^2 . Demandons-nous quel est le degré de liberté des Y_i . En fait, il y a une relation linéaire entre les Y_i puisque

$$\sum_i Y_i = \frac{1}{\sigma} \sum_i (X_i - \bar{X}_n) = \frac{1}{\sigma} (n\bar{X}_n - n\bar{X}_n) = 0$$

donc il n'y a que $n - 1$ degrés de liberté dans la famille (Y_1, \dots, Y_n) . Par conséquent,

$$\frac{n-1}{\sigma^2} \cdot S_n^2 = \sum_i \left[\frac{X_i - \bar{X}_n}{\sigma} \right]^2 = \sum_i Y_i^2 \rightsquigarrow \chi_{n-1}^2.$$

Approximation par la loi normale lorsque n est grand Dans le cas d'un grand échantillon ($n \geq 30$), on peut approcher la loi du χ^2 par une loi normale. En effet:

Rappel 11 Une loi du χ^2 à n degrés de liberté s'approche par une loi normale d'espérance n et de variance $2n$ lorsque n est grand.

Ainsi:

$$\frac{n-1}{\sigma^2} \cdot S_n^2 \rightsquigarrow \mathcal{N}(n-1, 2(n-1)).$$

Par conséquent,

$$S_n^2 \rightsquigarrow \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n-1}\right)$$

Nous retrouvons bien les résultats précédents sur les moments de S_n^2 dans le cas de la loi normale.

2.3.2 Cas particulier où la moyenne est connue (et la population normale)

Dans le cas où la moyenne est connue et non estimée, nous avons vu que l'estimateur de la variance

$$\mathcal{V}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

était non biaisé. Il convient donc de l'utiliser si l'on connaît μ dans la mesure où il est nécessairement plus précis que la variance empirique modifiée. La démonstration faite précédemment pour S_n^2 s'adapte très facilement pour \mathcal{V}_n mais il faut faire attention à ce que les n variables aléatoires $Y_i = \frac{X_i - \mu}{\sigma}$ sont cette fois indépendantes. Par conséquent, le nombre de degrés de libertés est n et non pas $n-1$. En bref, si μ est connue, alors:

$$\frac{n}{\sigma^2} \cdot \mathcal{V}_n \rightsquigarrow \chi_n^2.$$

2.3.3 Distribution de l'estimateur de la variance pour une population de loi inconnue

Lorsque la loi du caractère est inconnue et donc a priori pas normale, on va utiliser le théorème central-limite pour dériver la distribution de l'estimateur de la variance. On ne pourra appliquer ce théorème que dans le cas où n est suffisamment grand.

Pour appliquer le TCL, il faut voir S_n^2 comme une moyenne. Posons donc:

$$Y_i = (X_i - \bar{X}_n)^2$$

Table 1: Distribution de la variance empirique d'une population normale, pour un tirage avec remise ou sans remise dans une population infinie

Moyenne μ	Variance empirique	Distribution	Approximation pour n grand
Inconnue	$S_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2$	$\frac{n-1}{\sigma^2} \cdot S_n^2 \rightsquigarrow \chi_{n-1}^2$	$S_n^2 \rightsquigarrow \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n-1}\right)$
Connue	$\mathcal{V}_n = \frac{1}{n} \sum_i (X_i - \mu)^2$	$\frac{n}{\sigma^2} \cdot \mathcal{V}_n \rightsquigarrow \chi_n^2$	$\mathcal{V}_n \rightsquigarrow \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$

Dans ce cas,

$$S_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_i Y_i = \frac{n}{n-1} \cdot \bar{Y}_n$$

On va donc pouvoir appliquer le TCL sur \bar{Y}_n . Lorsque n est grand, on peut approcher la loi de \bar{Y}_n par une normale:

$$\bar{Y}_n \rightsquigarrow \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

Il nous faut donc calculer les moments de Y :

$$E(Y) = E([X - \bar{X}_n]^2) = V(X) = \sigma^2$$

$$V(Y) = \sigma_Y^2 = V([X - \bar{X}_n]^2) = E([X - \bar{X}_n]^4) - [E(X - \bar{X}_n)^2]^2 = \mu_4 - \sigma^4$$

Comme on est dans le cas où n est grand, $\frac{n}{n-1}$ est proche de 1 et par conséquent:

$$S_n^2 \rightsquigarrow \mathcal{N}\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right)$$

Vous noterez qu'on retrouve bien les résultats obtenus précédemment sur les moments de S_n^2 .

2.3.4 Résumé des résultats

Le tableau 1 résume les principaux résultats obtenus dans la dernière section.

2.4 Construction d'un intervalle de confiance pour la variance

Maintenant que l'on a spécifié la loi des estimateurs de la variance, on est en mesure de construire un intervalle de confiance. On donne la formule pour le

cas où μ est inconnue, sachant que le cas où elle est connue en découle très simplement (voir application).

Puisque $\frac{n-1}{\sigma^2} \cdot S_n^2 \rightsquigarrow \chi_{n-1}^2$, on pose $\chi_{\alpha;n-1}^2$ la valeur telle que pour C variable aléatoire de χ^2 à $n-1$ degrés de liberté:

$$P(C \geq \chi_{\alpha;n-1}^2) = 1 - \alpha.$$

Par conséquent,

$$P\left(\chi_{1-\alpha/2;n-1}^2 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{\alpha/2;n-1}^2\right) = 1 - \alpha.$$

En réarrangeant les termes, nous allons obtenir un intervalle de confiance pour σ^2 et donc σ :

$$\begin{aligned} 1 - \alpha &= P\left(\chi_{1-\alpha/2;n-1}^2 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{\alpha/2;n-1}^2\right) \\ &= P\left(\frac{1}{\chi_{\alpha/2;n-1}^2} \leq \frac{\sigma^2}{(n-1)S_n^2} \leq \frac{1}{\chi_{1-\alpha/2;n-1}^2}\right) \\ &= P\left(\frac{S_n^2(n-1)}{\chi_{\alpha/2;n-1}^2} \leq \sigma^2 \leq \frac{S_n^2(n-1)}{\chi_{1-\alpha/2;n-1}^2}\right) \\ &= P\left(S_n \sqrt{\frac{n-1}{\chi_{\alpha/2;n-1}^2}} \leq \sigma \leq S_n \sqrt{\frac{n-1}{\chi_{1-\alpha/2;n-1}^2}}\right) \end{aligned}$$

Lorsque n est grand et que vous voulez utiliser l'approximation par la loi normale, il suffit de faire la même chose que ce qu'on a vu pour l'estimation par intervalle d'une moyenne (voir exemple ci-dessous).

2.5 Application

Un sondage est effectué auprès de jeunes diplômés de l'université afin d'estimer leurs salaires ainsi que les inégalités salariales. On recueille l'information auprès de 25 individus 1 an après qu'ils aient obtenus leurs diplômes. On suppose que la loi des salaires est normale. Un rapide calcul donne que la moyenne sur l'échantillon est de 22000 euros, tandis que l'écart-type de l'échantillon est de 2500 euros.

- Quelle variance faut-il utiliser: $\mathcal{V}_n, \tilde{S}_n^2, S_n^2$? Comment calcule-t-on ξ_n^2 à partir de l'information donnée?
- Distribution de la variance empirique utilisée?
- Intervalles de confiance de σ à 90%? 95%? 99%?

- La précision étant relativement faible, on décide de doubler la taille de l'échantillon. Faisons l'hypothèse que la moyenne et l'écart-type de l'échantillon ne changent pas. Nouveaux intervalles de confiance?
- Maintenant, on suppose que l'on connaît la vraie moyenne et qu'elle vaut 22200 euros. On suppose d'ailleurs qu'on a de nouveau uniquement 25 observations. Quel estimateur de la variance utilise-t-on? Donner son estimation ponctuelle puis les intervalles de confiance de σ ?

3 Estimateur de la moyenne dans le cas où la variance est inconnue

3.1 Distribution de l'estimateur de la moyenne quand σ est inconnue

Prendre en compte l'incertitude sur σ Maintenant que nous savons comment estimer une variance, nous pouvons revenir sur le cas où nous cherchons à estimer une moyenne sans connaître la variance. Pour rappel, nous savons que si X suit une loi normale avec n quelconque ou si X suit une loi quelconque avec n grand et dans le cas d'un tirage avec remise, alors:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0, 1)$$

où μ et σ sont respectivement l'espérance et la variance de X .

Le problème de cette formule est que, si l'on ne connaît pas σ , il faut le remplacer par son estimateur. Or nous avons déjà vu que remplacer la vraie valeur d'un paramètre par son estimateur, qui est incertain, change les résultats. Rappelez-vous par exemple que la variance semi-empirique est un estimateur sans biais de la variance mais que lorsqu'on remplace μ par son estimateur \bar{X}_n pour obtenir la variance empirique non modifiée alors on trouve que cette dernière est un estimateur biaisé de la variance. De la même façon, lorsqu'on va remplacer σ par son estimateur, on va introduire de l'incertitude qui n'existait pas auparavant et qu'il faut prendre en compte dans la distribution, notamment si la taille de l'échantillon est petit (et que par conséquent, l'estimation que l'on peut avoir de σ est peu précise).

3.1.1 Cas d'une population normale

Si $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$ alors on sait que:

$$\frac{n-1}{\sigma^2} S_n^2 \rightsquigarrow \chi^2(n-1).$$

Rappel 12 Soit Z une variable aléatoire de loi normale centrée réduite et Y une variable aléatoire de distribution χ^2 à ν degrés de liberté, lorsque ces deux variables sont indépendantes, la variable aléatoire T définie formellement comme:

$$T = \frac{Z}{\sqrt{\frac{Y}{\nu}}}$$

suit une loi de Student à ν degrés de liberté. Il est alors possible de montrer que l'espérance mathématique et la variance de la variable T vérifient respectivement:

$$E(T) = 0 \quad \text{pour } \nu > 1 \quad \text{et} \quad V(T) = \frac{\nu}{\nu - 2} \quad \text{pour } \nu > 2$$

Appliquons ce résultat à notre cas: on pose

$$Y = \frac{n-1}{\sigma^2} S_n^2$$

où Y est une v.a. du χ^2 à $\nu = n - 1$ degrés de liberté et

$$Z = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

puis on construit la variable aléatoire de Student:

$$T = \frac{Z}{\sqrt{\frac{Y}{\nu}}} = \frac{\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{\frac{n-1}{\sigma^2} S_n^2}{n-1}}} = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{\sqrt{n-1}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}}} = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}}$$

La statistique T correspond bien à la statistique Z dans laquelle on a substitué σ à son estimateur; or on vient de montrer que T suivait une loi de Student à $\nu = n - 1$ degrés de liberté:

$$T = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \rightsquigarrow T_{n-1}$$

Lorsque la taille de l'échantillon est grande, la loi de Student tend à se comporter comme une loi normale d'espérance nulle et de variance $\frac{\nu}{\nu-2}$. Ainsi, de façon plus formelle, nous avons, pour n grand:

$$T \rightsquigarrow \mathcal{N}\left(0, \frac{n-1}{n-3}\right)$$

qui est proche d'une normale centrée réduite pour n suffisamment grand.

3.1.2 Cas d'une population quelconque avec $n \geq 30$

Sans entrer dans le détail, retenons simplement que dans le cas où la population est quelconque mais n grand, alors la statistique T définie précédemment suit une normale centrée réduite.

$$T \underset{n \rightarrow \infty}{\rightsquigarrow} \mathcal{N}(0, 1)$$

3.1.3 Résumé

On s'intéresse aux fluctuations de l'estimateur de la moyenne autour de son objectif qu'est la moyenne $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$, les résultats obtenus sont résumés dans le tableau 2.

Table 2: Distribution de la moyenne dans le cas d'un tirage avec remise ou sans remise dans une grande population ($s = \frac{1}{n-1} \sum_i x_i - \bar{x}_n$)

Variance	Population	Taille de l'échantillon	Ecart-réduit	Loi
Connue	Inconnue	$n < 30$	-	-
Connue	Inconnue	$n \geq 30$	$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$	$\rightsquigarrow \mathcal{N}(0, 1)$
Connue	Normale	quelconque	$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$	$\rightsquigarrow \mathcal{N}(0, 1)$
Inconnue	Inconnue	$n < 30$	-	-
Inconnue	Inconnue	$n \geq 30$	$Z = \frac{\bar{X}_n - \mu}{s/\sqrt{n}}$	$\rightsquigarrow \mathcal{N}(0, 1)$
Inconnue	Normale	$n < 30$	$Z = \frac{\bar{X}_n - \mu}{s/\sqrt{n}}$	$\rightsquigarrow T_{n-1}$
Inconnue	Normale	$n \geq 30$	$Z = \frac{\bar{X}_n - \mu}{s/\sqrt{n}}$	$\rightsquigarrow \mathcal{N}(0, 1)$

3.2 Application et construction d'un intervalle de confiance

On reprend l'exemple précédent des 25 observations tirées dans une population normale de salaires dont les moments empiriques étaient: $\bar{X}_n = 22000$ et $\tilde{S}_n = 2500$. On est dans le cas où σ est inconnu mais estimé.

- Estimation ponctuelle de μ ?
- Distribution de l'estimateur de la moyenne?
- Intervalles de confiance pour μ à 90%, 95%, 99%?

4 Estimateur de la proportion

Il nous reste à décrire comment fournir une estimation ponctuelle et par intervalle de la proportion d'un échantillon qui possède un caractère qualitatif X .

4.1 Propriétés de l'estimateur de la proportion

Nous avons déjà obtenu des résultats sur la distribution de la proportion empirique. En effet, nous avons vu que pour $\hat{P} = \frac{1}{n} \sum_{i=1}^n X_i$ où la variable aléatoire X prend la valeur 1 si l'individu possède le caractère et 0 sinon,

$$\begin{aligned} E(\hat{P}) &= p \\ V(\hat{P}) &= \frac{p(1-p)}{n} \end{aligned}$$

Par conséquent, la proportion d'échantillon est un estimateur intéressant de la proportion p puisque c'en est un estimateur sans biais (à distance finie). De plus,

$$\lim_{n \rightarrow \infty} \frac{p(1-p)}{n} = 0$$

donc la variance de l'estimateur \hat{P} tend vers 0 quand n devient très grand. L'estimateur de la proportion est donc un estimateur convergent.

Par ailleurs, nous avons aussi montré que la loi de l'estimateur \hat{P} est une Bernouilli, qui peut s'approcher par:

- une loi normale si n est grand et p proche de 0.5,
- une loi de Poisson si n est grand et p proche de 0.

Nous sommes donc en mesure de fournir une estimation ponctuelle de p à l'aide de l'estimateur \hat{P} . Passons donc maintenant à l'estimation par intervalle.

4.2 Estimation par intervalle de la proportion

On se placera uniquement dans le cas où n est grand et p pas trop proche des bornes (0 ou 1), de façon à pouvoir utiliser l'approximation par loi normale. Dans ce cas,

$$\hat{P}_n \rightsquigarrow \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

en se ramenant à une variable aléatoire centrée réduite, on obtient:

$$Z = \frac{\hat{P}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$$

Le terme au dénominateur correspond à l'écart-type de l'estimateur mais on ne connaît justement pas p puisque c'est ce que l'on cherche. Un estimateur de la variance de l'estimateur est donné par:

$$\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}$$

et dans ce cas la variable centrée réduite Z est toujours distribuée selon une loi normale centrée réduite, donc:

$$Z = \frac{\widehat{P}_n - p}{\sqrt{\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}}} \rightsquigarrow \mathcal{N}(0, 1)$$

On peut donc maintenant construire un intervalle de confiance symétrique pour p de la façon suivante:

$$P\left(-z_{\alpha/2} \leq \frac{\widehat{P}_n - p}{\sqrt{\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

En réarrangeant les termes, on obtient:

$$P\left(\widehat{P}_n - z_{\alpha/2} \sqrt{\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}} \leq p \leq \widehat{P}_n + z_{\alpha/2} \sqrt{\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}}\right) = 1 - \alpha.$$

d'où l'intervalle de confiance à $1 - \alpha$:

$$\left[\widehat{P}_n - z_{\alpha/2} \sqrt{\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}}; \widehat{P}_n + z_{\alpha/2} \sqrt{\frac{\widehat{P}_n(1 - \widehat{P}_n)}{n}} \right]$$

4.3 Application

M. Sarkal et Mme Royosy s'affrontent aux élections présidentielles. Un sondage sur 1000 personnes donne gagnant Mme Royosy à 53%.

- Estimation ponctuelle de la proportion p de votants pour Mme Royosy?
- Intervalle de confiance à 90%? 95%? 99%?
- Nombre d'observations nécessaires pour garantir que $p > 0.5$?