Estimation

$https://www.imo.universite-paris-saclay.fr/\ pansu/web_ifips/proba_S5-IFIPS_estimation.tex$

4 avril 2020

Estimer ne coûte presque rien, Estimer incorrectement coûte cher. Vieux proverbe chinois.

1 Introduction

Dans de nombreux domaines (scientifiques, économiques, épidémiologiques...), on a besoin de connaître certaines caractéristiques d'une population. Mais, en règle générale, on ne peut pas les évaluer facilement du fait de l'effectif trop important des populations concernées. La solution consiste alors à estimer le paramètre cherché à partir de celui observé sur un échantillon plus petit.

L'idée de décrire une population à partir d'un échantillon réduit, à l'aide d'un "multiplicateur", n'a été imaginée que dans la seconde moitié du XVIIIème siècle, notamment par l'école arithmétique politique anglaise. Elle engendra une véritable révolution : l'observation d'échantillons permettait d'éviter des recensements d'une lourdeur et d'un prix exorbitants. Toutefois, on s'aperçut rapidement que les résultats manquaient d'exactitude. Nous savons maintenant pourquoi : on ne prenait en considération ni la *représentativité* de l'échantillon, ni les *fluctuations* d'échantillonnage. C'est là que le hasard intervient.

La première précaution à prendre est donc d'obtenir un échantillon représentatif. Nous pourrons en obtenir un par tirage au sort (voir le chapitre précédent sur l'échantillonnage aléatoire simple) : le hasard participe donc au travail du statisticien qui l'utilise pour pouvoir le maîtriser ! Mais, même tiré au sort, un échantillon n'est pas l'image exacte de la population, en raison des fluctuations d'échantillonnage. Lorsque, par exemple, on tire au sort des échantillons dans un urne contenant 20 % de boules blanches, on obtient des échantillons où la proportion de boules blanches fluctue autour de 20 %. Ces fluctuations sont imprévisibles : le hasard peut produire n'importe quel écart par rapport à la proportion de la population (20 %). Cependant, on

s'en doute, tous les écarts ne sont pas également vraisemblables : les très grands écarts sont très peu probables. Au moyen du calcul des probabilités, le statisticien définit un intervalle autour du taux observé, intervalle qui contient probablement le vrai taux : c'est "l'intervalle de confiance" ou, plus couramment, la "fourchette".

Si l'on ne peut connaître le vrai taux par échantillonnage, peut-on au moins le situer avec certitude dans la fourchette ? Non. Le hasard étant capable de tous les caprices, on ne peut raisonner qu'en termes de probabilités, et la fourchette n'a de signification qu'assortie d'un certain risque d'erreur. On adopte souvent un risque de 5 % : cinq fois sur cent, le taux mesuré sur l'échantillon n'est pas le bon, le vrai taux étant en dehors de la fourchette. On peut diminuer le risque d'erreur mais alors la fourchette grandit et perd de son intérêt. Bien entendu, il existe une infinité de fourchettes, une pour chaque risque d'erreur adopté. On doit trouver un compromis entre le risque acceptable et le souci de précision.

Dans ce cours, nous allons apprendre à estimer à l'aide d'un échantillon :

— Dans le cas d'un caractère quantitatif la moyenne m et l'écart-type d'une population.

— Dans le cas d'un caractère qualitatif, la proportion p de la population.

Ces estimations peuvent s'exprimer par une seule valeur (estimation ponctuelle), soit par un intervalle (estimation par intervalle de confiance). Bien sûr, comme l'échantillon ne donne qu'une information partielle, ces estimations seront accompagnées d'une certaine marge d'erreur.

2 L'estimation ponctuelle

2.1 Définition

Estimer un paramètre, c'est en chercher une valeur approchée en se basant sur les résultats obtenus dans un échantillon. Lorsqu'un paramètre est estimé par un seul nombre, déduit des résultats de l'échantillon, ce nombre est appelé *estimation ponctuelle* du paramètre.

L'estimation ponctuelle se fait à l'aide d'un *estimateur*, qui est une variable aléatoire d'échantillon. L'estimation est la valeur que prend la variable aléatoire dans l'échantillon observé.

2.2 Propriétés des estimateurs ponctuels

Lorsqu'on utilise fréquemment des estimateurs ponctuels on souhaite qu'ils possèdent certaines propriétés. Ces propriétés sont importantes pour choi-

sir le meilleur estimateur du paramètre correspondant, c'est-à-dire celui qui s'approche le plus possible du paramètre à estimer. Un paramètre inconnu peut avoir plusieurs estimateurs. Par exemple, pour estimer le paramètre m, moyenne d'une population, on pourrait se servir de la moyenne arithmétique, de la médiane ou du mode. Les qualités que doit posséder un estimateur pour fournir de bonnes estimations sont décrites ci-après.

Définition

Estimateur non biaisé. On note θ le paramètre de valeur inconnue, $\hat{\theta}$ l'estimateur de θ.

Un estimateur est *sans biais* si la moyenne de sa distribution d'échantillonnage est égale à la valeur θ du paramètre de la population à estimer, c'est-à-dire si $E(\hat{\theta}) = \theta$.

Si l'estimateur est biaisé, son biais est mesuré par l'écart suivant :

$$\text{BIAIS} = E(\hat{\theta}) - \theta.$$

Exemple

On a vu que $E(\bar{X}) = m$. Donc la moyenne d'échantillon $\bar{X}$ est un estimateur sans biais du paramètre m, moyenne de la population.

En revanche, la médiane d'échantillon M_e est un estimateur biaisé lorsque la population échantillonnée est asymétrique.

Exemple

Nous avons vu également que $E(\Sigma^2_{ech}) = \frac{n-1}{n}\sigma^2_{pop}$. Donc Σ^2_{ech} est un estimateur biaisé du paramètre σ^2_{pop}, variance de la population.

C'est pour cette raison que l'on a introduit la variance d'échantillon $S^2 = \frac{n}{n-1}\Sigma^2_{ech}$ qui est un estimateur sans biais de σ^2_{pop}, puisque $E(S^2) = \sigma^2_{pop}$.

L'absence de biais, à elle toute seule, ne garantit pas que nous avons un bon estimateur. En effet, certains paramètres peuvent avoir plusieurs estimateurs sans biais. Le choix parmi les estimateurs sans biais s'effectue en comparant les variances des estimateurs. En effet, un estimateur sans biais mais à variance élevée peut fournir des estimations très éloignées de la vraie valeur du paramètre.

Définition

Estimateur efficace. Un estimateur sans biais est efficace si sa variance est la plus faible parmi les variances des autres estimateurs sans biais.

Ainsi, si $\hat{\theta}_1$ et $\hat{\theta}_2$ sont deux estimateurs sans biais du paramètre θ, l'estimateur $\hat{\theta}_1$ est plus efficace que $\hat{\theta}_2$ si

$$E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta \quad \text{et } V(\hat{\theta}_1) < V(\hat{\theta}_2).$$

Définition

Estimateur convergent. Un estimateur $\hat{\theta}$ est convergent si sa distribution tend à se concentrer autour de la valeur inconnue à estimer, θ, à mesure que la taille d'échantillon augmente, c'est-à-dire si $n \to \infty$.

Par exemple, $\bar{X}$ est un estimateur convergent puisque $V(\bar{X}) = \frac{\sigma^2_{pop}}{n}$ tend vers 0.

Remarque

Un estimateur sans biais et convergent est dit absolument correct.

Ces trois propriétés sont les principales qualités que nous recherchons pour un estimateur. Nous n'insisterons pas sur les propriétés mathématiques que doivent posséder les estimateurs.

Conséquences : nous a appris que

$$\begin{aligned} E(\bar{X}) = m \quad \text{et} \quad V(\bar{X}) &= \frac{\sigma^2_{pop}}{n}, \\ E(S^2) = \sigma^2_{pop} \quad \text{et} \quad V(S^2) &= \frac{2\sigma^4_{pop}}{n-1}, \\ E(F) = p \quad \text{et} \quad V(F) &= \frac{pq}{n}. \end{aligned}$$

On peut donc affirmer que :

— $\bar{X}$ est un estimateur absolument correct de la moyenne m pour un caractère quantitatif.

— S^2 est un estimateur absolument correct de la variance pour un caractère quantitatif.

— F est un estimateur absolument correct de la proportion p pour un caractère qualitatif.

Nous pourrons donc estimer m par $\bar{X}$, σ^2_{pop} par S^2, p par F.

Mais les estimations ponctuelles bien qu'utiles, ne fournissent aucune information concernant la précision des estimations, c'est-à-dire qu'elles ne tiennent pas compte de l'erreur possible dans l'estimation, erreur attribuable aux fluctuations d'échantillonnage. Quelle confiance avons-nous dans une valeur unique ? On ne peut répondre à cette question en considérant uniquement l'estimation ponctuelle obtenue des résultats de l'échantillon. Il faut lui associer un intervalle qui permet d'englober avec une certaine fiabilité, la vraie valeur du paramètre correspondant.

3 Estimation par intervalle de confiance

3.1 Définition

L'estimation par intervalle d'un paramètre inconnu θ consiste à calculer, à partir d'un estimateur choisi $\hat{\theta}$, un intervalle dans lequel il est vraisemblable que la valeur correspondante du paramètre s'y trouve. L'intervalle de confiance est défini par deux limites LI et LS auxquelles est associée une certaine probabilité, fixée à l'avance et aussi élevée qu'on le désire, de contenir la valeur vraie du paramètre. La probabilité associée à l'intervalle de confiance et exprimée en pourcentage est égale à S où S est le seuil de confiance ou niveau de confiance de l'intervalle, exprimé également en pourcentage. Autrement dit,

$$P(LI \leq \theta \leq LS) = S,$$

où

- LI est la limite inférieure de l'intervalle de confiance.
- LS est la limite supérieure de l'intervalle de confiance.
- S est la probabilité associée à l'intervalle d'encadrer la vraie valeur du paramètre.

LI et LS sont appelées les *limites de confiance* de l'intervalle et sont des quantités qui tiennent compte des fluctuations d'échantillonnage, de l'estimateur $\hat{\theta}$ et du seuil de confiance S. La quantité 1 - S est égale à la probabilité, exprimée en pourcentage, que l'intervalle n'encadre pas la vraie valeur du paramètre. On note $\alpha = 1 - S$. α s'appelle le risque ou le *seuil de signification* de l'intervalle.

A quoi correspond l'intervalle de confiance ?

Si nous répétons l'expérience un grand nombre de fois (prélever un grand nombre de fois un échantillon de taille n de la même population), dans 100S cas sur 100 les intervalles obtenus (différents à chaque réalisation de l'expérience) recouvrent la vraie valeur du paramètre.

Remarques :

- L'intervalle ainsi défini est un intervalle aléatoire puisqu'avant l'expérience, les limites de l'intervalle sont des variables aléatoires (elles sont fonctions des observations de l'échantillon).
- Le niveau de confiance est toujours associé à l'intervalle et non au paramètre inconnu θ. θ n'est pas une variable aléatoire : il est ou n'est pas dans l'intervalle $[LI, LS]$.
- Le niveau de confiance doit être choisi *avant* que ne s'effectue l'estimation par intervalle. Il arrive souvent que le chercheur non averti calcule plusieurs intervalles d'estimation à des niveaux de confiance différents

et choisisse par la suite l'intervalle qui lui semble le plus approprié. Une telle approche constitue en réalité une interprétation inacceptable des données en ce qu'elle fait dire aux résultats échantillonnaux ce que l'on veut bien entendre.

— Il y a une infinité de solutions possibles pour déterminer l'intervalle $[LI, LS]$. On choisira de prendre des risques symétriques, c'est-à-dire de choisir LI et LS tels que

$$P(\theta \leq LI) = P(\theta \geq LS) = \frac{1-S}{2}.$$

Pour calculer l'intervalle de confiance, on doit connaître la distribution d'échantillonnage (distribution de probabilité) de l'estimateur correspondant, c'est-à-dire connaître de quelle façon sont distribuées toutes les valeurs possibles de l'estimateur obtenues à partir de tous les échantillons possibles de même taille prélevés de la même population. Ce travail a été effectué au chapitre précédent. Nous allons voir à présent comment déduire des distributions d'échantillonnage la construction des intervalles de confiance.

3.2 Estimation d'une moyenne par intervalle de confiance

On se propose d'estimer, par intervalle de confiance, la moyenne m d'un caractère mesurable d'une population. Il s'agit donc de calculer, à partir de la moyenne $\bar{x}$ (valeur prise par l'estimateur $\bar{X}$) de l'échantillon, un intervalle dans lequel il est vraisemblable que la vraie valeur de m se trouve. Cet intervalle se définit d'après l'équation $P(A \leq m \leq B) = S$. Les limites A et B de cet intervalle sont des quantités aléatoires et prendront, après avoir prélevé l'échantillon et calculé l'estimation $\bar{x}$, la forme $LI \leq m \leq LS$. Nous allons déterminer LI et LS en utilisant la distribution d'échantillonnage de $\bar{X}$. L'étude du chapitre 4 nous amène donc à distinguer deux cas, suivant la taille de l'échantillon.

3.2.1 a. On dispose d'un grand échantillon ($n \geq 30$) ou d'un petit échantillon ($n < 30$) dont la distribution est normale d'écart-type connu σ_{pop}

Dans ces conditions on considère que la variable aléatoire $\bar{X}$ suit une loi normale,

$$\bar{X} \rightharpoonup \mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}}).$$

Donc $\underset{\sim}{T} = \frac{\bar{X}-m}{\frac{\sigma_{pop}}{\sqrt{n}}}$ suit la loi $\mathcal{N}(0,1)$.

On cherche à déterminer A et B tels que $P(A \leq m \leq B) = S$.

Puisqu'on choisit des risques symétriques, on va déterminer dans la table de la loi normale centrée réduite la valeur $t_{\alpha/2}$ telle que $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = S$, ce qui peut s'écrire

$$P(\bar{X} - t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}) = S,$$

qui est bien de la forme cherchée en posant

$$A = \bar{X} - t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}, \quad B = \bar{X} + t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}.$$

Signification. Avant toute expérience, la probabilité que l'intervalle aléatoire $[\bar{X} - t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}]$ contienne la vraie valeur de m est S. Ces deux limites sont des variables aléatoires qui prendront des valeurs numériques particulières une fois que l'échantillon est choisi et qu'on a obtenu la valeur de $\bar{x}$ (réalisation de la variable aléatoire $\bar{X}$). On en déduit par la suite un intervalle d'extrémités fixes (et non plus un intervalle aléatoire) qui s'écrit $[\bar{x} - t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}]$, et on lui attribue, non pas une probabilité, mais un niveau de confiance de α de contenir la vraie valeur de m.

Conclusion. A partir d'un échantillon de grande taille ($n \geq 30$) ou à partir d'un échantillon de petite taille ($n < 30$), prélevé à partir d'une population normale de moyenne m (inconnue) et de variance σ^2_{pop} connue, on définit un intervalle de confiance ayant un niveau de confiance S de contenir la vraie valeur de m par

$$[\bar{x} - t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{\sigma_{pop}}{\sqrt{n}}].$$

Remarque

Dans le cas d'un grand échantillon, si la variance σ^2_{pop} de la population est inconnue, on peut l'estimer sans problème par la variance d'échantillon $s^2 = \frac{n}{n-1}\sigma^2_{ech}$ (voir chapitre 4).

3.2.2 b. On dispose d'un petit échantillon ($n < 30$) et la distribution de X est normale d'écart-type inconnu

Dans ces conditions, l'étude du chapitre précédent nous a appris que nous ne disposions pas directement de la loi de $\bar{X}$ mais de celle de

$$T = \frac{\bar{X} - m}{\Sigma_{ech}/\sqrt{n-1}}.$$

T suit une loi de Student à $n-1$ degrés de liberté : $T \rightharpoonup T_{n-1}$.

Pour trouver l'intervalle de confiance de m au risque α, nous allons procéder comme dans le cas précédent.

On détermine dans la table de la loi de Student la valeur $t_{\alpha/2,\nu}$ (où $\nu = n-1$) telle que $P(-t_{\alpha/2,\nu} \leq T \leq t_{\alpha/2,\nu}) = S$, ce qui peut s'écrire

$$P(\bar{X} - t_{\alpha/2,\nu}\frac{\Sigma_{ech}}{\sqrt{n-1}}, \bar{X} + t_{\alpha/2,\nu}\frac{\Sigma_{ech}}{\sqrt{n-1}}) = S.$$

Après avoir choisi l'échantillon, $\bar{X}$ a pris la valeur $\bar{x}$ et Σ_{ech} la valeur σ_{ech}. On en déduit par la suite un intervalle d'extrémités fixes (et non plus un intervalle aléatoire) qui s'écrit $[\bar{x}-t_{\alpha/2,\nu}\frac{\sigma_{ech}}{\sqrt{n-1}}, \bar{x}+t_{\alpha/2,\nu}\frac{\sigma_{ech}}{\sqrt{n-1}}]$ et on lui attribue, non pas une probabilité, mais un niveau de confiance de α de contenir la vraie valeur de m.

Conclusion. A partir d'un échantillon de petite taille ($n < 30$), prélevé à partir d'une population normale de moyenne m (inconnue) et de variance σ^2_{pop} inconnue, on définit un intervalle de confiance ayant un niveau de confiance S de contenir la vraie valeur de m par

$$[\bar{x} - t_{\alpha/2,\nu}\frac{\sigma_{ech}}{\sqrt{n-1}}, \bar{x} + t_{\alpha/2,\nu}\frac{\sigma_{ech}}{\sqrt{n-1}}],$$

où $\nu = n-1$ est le nombre de degrés de liberté de la distribution de Student.

On pourra bien sûr remplacer $\frac{\sigma_{ech}}{\sqrt{n-1}}$ par $\frac{s}{\sqrt{n}}$.

3.3 Remarques

1. L'intervalle de confiance pourra être numériquement différent chaque fois qu'on prélève un échantillon de même taille de la population puisque l'intervalle est centré sur la moyenne de l'échantillon qui varie de prélèvement en prélèvement.

2. Le niveau de confiance est associé à l'intervalle et non au paramètre m. Il ne faut pas dire que la vraie valeur de m a, disons 95 chances sur 100, de se trouver dans l'intervalle mais plutôt que l'intervalle de confiance a 95 chances sur 100 de contenir la vraie valeur de m ou encore que 95 fois sur 100, l'intervalle déterminé contiendra la vraie valeur de m. Une fois que l'intervalle est calculé, m est ou n'est pas dans l'intervalle (pour une population donnée, m est une constante et non une variable aléatoire).

3. Plus le niveau de confiance est élevé, plus l'amplitude de l'intervalle est grande. Pour la même taille d'échantillon, on perd de la précision en gagnant une plus grande confiance.

4. Dans le cas où la variance de la population est inconnue, des échantillonnages successifs de la population peuvent conduire pour une même taille d'échantillon et le même niveau de confiance, à des intervalles de diverses amplitudes parce que l'écart-type s variera d'échantillon en échantillon.

3.4 Estimation d'une variance par intervalle de confiance

On se propose d'estimer, par intervalle de confiance, la variance σ^2_{pop} d'un caractère mesurable d'une population. Il s'agit donc de déterminer, à partir de la variance de l'échantillon σ^2_{ech}, un intervalle dans lequel il est vraisemblable que la vraie valeur de σ^2_{pop} se trouve.

On cherche un intervalle $[A, B]$ vérifiant $P(A \leq \sigma^2_{pop} \leq B) = S$. Les limites de cet intervalle prendront, après avoir prélevé l'échantillon et calculé l'estimation les valeurs prises par les deux quantités aléatoires A et B, la forme $a \leq \sigma^2_{pop} \leq b$.

Nous allons déterminer A et B en utilisant la distribution d'échantillonnage de la variance d'échantillon S^2.

Nous supposerons par la suite que la population est "normale", c'est-à-dire que le caractère X suit une loi normale. L'étude du chapitre 4 nous amène donc à distinguer deux cas.

3.4.1 a. La population est "normale" et on dispose d'un grand échantillon ($n \geq 30$)

La variance d'échantillon $S^2 = \frac{n}{n-1}\Sigma^2_{ech}$ suit approximativement une loi normale (voir chapitre 4), $S^2 \rightharpoonup \mathcal{N}(\sigma^2_{pop}, \sigma^2_{pop}\sqrt{\frac{2}{n-1}})$, donc

$$T = \frac{S^2 - \sigma^2_{pop}}{\sigma^2_{pop}\sqrt{\frac{2}{n-1}}}$$

suit une loi normale centrée réduite.

On peut déterminer dans la table de la loi normale centrée réduite la valeur $t_{\alpha/2}$ telle que $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = S$, ce qui peut s'écrire

$$P(-t_{\alpha/2} \leq \frac{S^2 - \sigma^2_{pop}}{\sigma^2_{pop}\sqrt{\frac{2}{n-1}}} \leq t_{\alpha/2}) = 1 - \alpha.$$

Comme on a un grand échantillon, on peut estimer σ^2_{pop} par $s^2 = \frac{n}{n-1}\sigma^2_{ech}$. Soit encore

$$P(S^2 - t_{\alpha/2}s^2\sqrt{\frac{2}{n-1}} \leq \sigma^2_{pop} \leq S^2 + t_{\alpha/2}s^2\sqrt{\frac{2}{n-1}}) = 1 - \alpha,$$

qui est bien de la forme cherchée.

Ces deux limites sont des variables aléatoires qui prendront des valeurs numériques particulières une fois que l'échantillon est choisi et qu'on a obtenu

la valeur de s^2 (réalisation de la variable aléatoire S^2). On en déduit par la suite un intervalle d'extrémités fixes (et non plus un intervalle aléatoire) qui s'écrit $[s^2 - t_{\alpha/2}s^2\sqrt{\frac{2}{n-1}} \leq \sigma^2_{pop} \leq s^2 + t_{\alpha/2}s^2\sqrt{\frac{2}{n-1}}]$, et on lui attribue un niveau de confiance S de contenir la vraie valeur de σ^2_{pop}.

Conclusion. A partir d'un échantillon de grande taille ($n \geq 30$), prélevé à partir d'une population normale de variance σ^2_{pop} inconnue, on définit un intervalle de confiance ayant un niveau de confiance $1-\alpha$ de contenir la vraie valeur de σ^2_{pop} par

$$[s^2 - t_{\alpha/2}s^2\sqrt{\frac{2}{n-1}} \leq \sigma^2_{pop} \leq s^2 + t_{\alpha/2}s^2\sqrt{\frac{2}{n-1}}].$$

3.4.2 b. La population est "normale" et on dispose d'un petit échantillon ($n < 30$)

La variable

$$Y = \frac{n\Sigma^2_{ech}}{\sigma^2_{pop}} = \frac{(n-1)S^2}{\sigma^2_{pop}}$$

suit une loi du χ^2 à $n-1$ degrés de liberté (voir chapitre 4), $Y \rightharpoonup \chi^2_{n-1}$.

Nous allons chercher un intervalle $[\chi^2_a, \chi^2_b]$ de valeurs telles que $P(\chi^2_a \leq Y \leq \chi^2_b) = S$.

On choisit un intervalle correspondant à des risques symétriques, c'est-à-dire tel que

$$P(Y < \chi^2_a) = P(\chi^2_b < Y) = \frac{1-S}{2} = \frac{\alpha}{2}.$$

Les deux valeurs χ^2_a et χ^2_b se déterminent à l'aide des tables. On peut alors écrire que $P(\chi^2_a \leq \frac{n\Sigma^2_{ech}}{\sigma^2_{pop}} \leq \chi^2_b) = S$ et donc que

$$P(\frac{n\Sigma^2_{ech}}{\chi^2_b} \leq \sigma^2_{pop} \leq \frac{n\Sigma^2_{ech}}{\chi^2_a}) = S.$$

Ces deux limites sont des variables aléatoires qui prendront des valeurs numériques particulières une fois que l'échantillon est choisi et qu'on a obtenu la valeur de s^2 (réalisation de la variable aléatoire S^2). On en déduit par la suite un intervalle d'extrémités fixes qui s'écrit $[\frac{n\sigma^2_{ech}}{\chi^2_b}, \frac{n\sigma^2_{ech}}{\chi^2_a}]$ et on lui attribue un niveau de confiance S de contenir la vraie valeur de σ^2_{pop}.

Conclusion. A partir d'un échantillon de petite taille ($n < 30$), prélevé à partir d'une population normale de variance σ^2_{pop} inconnue, on définit un

intervalle de confiance ayant un niveau de confiance S de contenir la vraie valeur de σ^2_{pop} par

$$[\frac{n\sigma^2_{ech}}{\chi^2_b}, \frac{n\sigma^2_{ech}}{\chi^2_a}].$$

3.5 Estimation d'une proportion par intervalle de confiance

On se propose d'estimer, par intervalle de confiance, la proportion p d'un caractère quantitatif d'une population. Il s'agit donc de déterminer, à partir de la proportion de l'échantillon f, un intervalle dans lequel il est vraisemblable que la vraie valeur de p s'y trouve. On cherche un intervalle $[A, B]$ vérifiant $P(A \leq p \leq B) = S$. Les limites de cet intervalle prendront, après avoir prélevé l'échantillon et calculé les valeurs prises par les deux quantités aléatoires A et B, la forme $LI \leq p \leq LS$.

Nous allons déterminer A et B en utilisant la distribution d'échantillonnage de la proportion d'échantillon F.

Nous supposons que nous sommes en présence d'un *grand échantillon* ($n \geq 30$) et que p (que nous devons estimer) n'est pas trop petit ($np \geq 15$ et $nq \geq 15$). La fréquence d'échantillon F suit approximativement une loi normale (voir chapitre 4), $F \rightharpoonup \mathcal{N}(p, \sqrt{\frac{pq}{n}})$. Donc

$$T = \frac{F - p}{\sqrt{\frac{pq}{n}}}$$

suit approximativement une loi normale centrée réduite.

On peut déterminer dans la table de la loi normale centrée réduite la valeur $t_{\alpha/2}$ telle que $P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = S$. Ce qui peut s'écrire :

$$P(-t_{\alpha/2} \leq \frac{F - p}{\sqrt{\frac{pq}{n}}} \leq t_{\alpha/2}) = S.$$

Le problème est qu'on ignore la valeur de p et qu'elle intervient dans l'écart-type. Comme n est grand, il est correct d'estimer p par la valeur f (prise par l'estimateur F) trouvée dans l'échantillon. En effet, la grande taille de l'échantillon garantit que f ne fluctue pas trop d'échantillon en échantillon. Soit encore

$$P(F - t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} \leq p \leq F + t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}) = S.$$

qui est bien de la forme cherchée.

Ces deux limites sont des variables aléatoires qui prendront des valeurs numériques particulières une fois que l'échantillon est choisi et qu'on a obtenu la valeur de f (réalisation de la variable aléatoire F). On en déduit par la suite un intervalle d'extrémités fixes qui s'écrit $[f - t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}, f + t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}]$ et on lui attribue un niveau de confiance S de contenir la vraie valeur de p.

Conclusion. A partir d'un échantillon de grande taille ($n \geq 30$), prélevé à partir d'une population dont la proportion p d'un caractère qualitatif est inconnue mais pas trop petite, on définit un intervalle de confiance ayant un niveau de confiance S de contenir la vraie valeur de p par

$$[f - t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}, f + t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}].$$

Nous verrons en travaux dirigés comment vous vous procéder.