Echantillonnage et estimations

F. Wlazinski

Licence d'économie

1 Introduction

Remarque 1.1

Dans de nombreux domaines, soit parce que l'effectif total est trop important, soit parce que les modalités ne peuvent pas être toutes connues simultanément, il est impossible d'étudier une variable sur tous les individus d'une population. On procède alors à un échantillonnage c'est-à-dire une ou des sélections d'échantillons. Cet échantillonnage suit différentes stratégies (échantillon aléatoire, échantillon représentatif,...). Le but est d'extrapoler correctement à toute la population les résultats obtenus sur le ou les échantillons. Un des problèmes principaux concerne la fiabilité des conclusions que l'on a tirées à partir de l'échantillon : N'a-t-on pas exclu une partie significative des modalités de la population (totale)? N'a-t-on pas fait une erreur systématique? Peut-on estimer la confiance des conclusions?

Inversement, et même si le problème ne nous concerne pas dans ce chapitre, quels résultats peut-on déduire sur un échantillon d'une population sachant connu le même résultat sur l'ensemble de cette population?

Rappelons aussi, qu'en termes statistiques, les individus peuvent certes être des personnes physiques mais aussi des objets ou des notions immatérielles (machine, produit, usine, situation d'un compte bancaire, temps,...)

Le cadre de travail de ce chapitre est le suivant : Nous nous plaçons dans une situation où le recensement des modalités d'une certaine variable sur toute la population n'est pas faisable. On sélectionne (peu importe la méthode) un échantillon de taille n appelé n -échantillon dont on évalue les modalités. Ensuite, on fera une estimation d'une fonction de la variable sur l'ensemble de la population.

Exemples 1.2

2 Estimation ponctuelle

Notations

Dans tout ce chapitre, on s'intéresse à un caractère C (qualitatif ou quantitatif) des N individus d'une population et on utilise les notations suivantes.

Si C est quantitatif, alors X est la v.a.r. égale à cette valeur. Son espérance (inconnue ou théorique) est notée μ et son écart-type est noté σ . La variance de X sera notée σ^2 .

Si C est qualitatif à deux modalités, alors X est la v.a.r. qui prend la valeur 1 pour la première modalité et 0 pour la deuxième. Son espérance (inconnue ou théorique) est notée p . Autrement dit, X suit une loi de Bernouilli $\mathcal{B}(p)$.

On note $(x_1, x_2, \dots, x_n)$ une observation du n -échantillon.

Remarques 2.1

Définition 2.2

Lorsque l'on choisit des individus dans une population et si cette sélection se fait avec remise, on dit que l'échantillon est *non-exhaustif*. Il est *exhaustif* dans le cas contraire.

Remarque 2.3**Définition 2.4**

On suppose que C est un caractère quantitatif et X est la v.a.r. (d'espérance μ et de variance σ^2) égale à sa valeur.

On dit que la moyenne de l'observation $(x_1, x_2, \dots, x_n)$ c'est-à-dire $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$ est une *estimation ponctuelle* de l'espérance μ .

On dit que la *variance corrigée* de l'observation $(x_1, x_2, \dots, x_n)$ c'est-à-dire $s_c^2 = \frac{n}{n-1} (\overline{x^2} - \bar{x}^2)$ est une *estimation ponctuelle* de la variance σ^2 .

Définition 2.5

On suppose que C est un caractère qualitatif et $X \sim \mathcal{B}(p)$.

On dit que la fréquence de l'observation $(x_1, x_2, \dots, x_n)$ c'est-à-dire $f = \frac{x_1 + x_2 + \dots + x_n}{n}$ est une *estimation ponctuelle* de p .

Remarque 2.6**3 Intervalles de confiance**

Dans toute cette partie, on note α le niveau de risque et $c = 1 - \alpha$ le niveau de confiance. En pratique, on prend très souvent $\alpha = 0,05 = 5\%$ et $c = 1 - \alpha = 0,95 = 95\%$. Et, hormis précision, ce sera la valeur par défaut.

Propriété 3.1

Si X suit une loi normale et si σ est connu, alors on peut estimer, au niveau de confiance $c = 1 - \alpha$ (ou au seuil α) que :

$$\mu \in [\bar{x} - d_\alpha; \bar{x} + d_\alpha]$$

où $d_\alpha = \frac{\sigma}{\sqrt{n}} \times z_\alpha$ et z_α est lu dans la table de la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Exemple 3.2

On dispose d'un échantillon de 25 individus dont la moyenne d'un caractère est 6. De plus, on sait que le caractère suit une loi normale dont l'écart type est 4. On cherche un intervalle de confiance au seuil 0,05 de la moyenne de la population.

Remarque 3.3**Remarque 3.4**

En pratique, les valeurs suivantes (surtout la première) sont souvent utilisées :

Pour $c = 95\% = 0,95$, on a $\alpha = 0,05$, $1 - \frac{\alpha}{2} = 0,975$ et $z_\alpha = 1,96$.

Pour $c = 99\% = 0,99$, on a $\alpha = 0,01$, $1 - \frac{\alpha}{2} = 0,995$ et $z_\alpha = 2,58$.

Pour $c = 90\% = 0,9$, on a $\alpha = 0,1$, $1 - \frac{\alpha}{2} = 0,95$ et $z_\alpha = 1,65$.

Propriété 3.5

Si $n > 30$, alors on peut estimer, au niveau de confiance $c = 1 - \alpha$ (ou au seuil α) que :

$$\mu \in [\bar{x} - d_\alpha; \bar{x} + d_\alpha]$$

où $d_\alpha = \frac{s_c}{\sqrt{n}} \times z_\alpha$ et z_α est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Exemple 3.6

On dispose d'un échantillon de 100 individus dont la moyenne d'un caractère est 6 et l'écart-type est 4. On cherche un intervalle de confiance au seuil 0,05 de la moyenne de la population.

Remarque 3.7

Propriété 3.8

Si X suit une loi normale, si σ est inconnu et si $n \leq 30$, alors on peut estimer, au niveau de confiance $c = 1 - \alpha$ (ou au seuil α) que :

$$\mu \in [\bar{x} - d_\alpha; \bar{x} + d_\alpha]$$

où $d_\alpha = \frac{s_c}{\sqrt{n}} \times t_\alpha$ et t_α est lu dans la table bilatérale de Student avec $d.d.l. = n - 1$.

Exemple 3.9

On dispose d'un échantillon de 25 individus dont la moyenne d'un caractère est 6 et l'écart type est 4. De plus, on sait que le caractère suit une loi normale. On cherche un intervalle de confiance au seuil 0,05 de la moyenne de la population.

Propriété 3.10

On peut estimer, au niveau de confiance $c = 1 - \alpha$ (ou au seuil α) que :

$$\sigma^2 \in \left[\frac{n-1}{\chi_{\text{sup}}^2} s_c^2; \frac{n-1}{\chi_{\text{inf}}^2} s_c^2 \right]$$

où χ_{sup}^2 est lu dans la table du χ^2 avec $\frac{\alpha}{2}$ et $n - 1$
et χ_{inf}^2 lu dans la table du χ^2 avec $1 - \frac{\alpha}{2}$ et $n - 1$.

Remarque 3.11

Exemple 3.12

Propriété 3.13

Dans le cas où $X \sim \mathcal{B}(p)$, si $np \geq 10$ et $n(1-p) \geq 10$, on peut estimer, au niveau de confiance $c = 1 - \alpha$ (ou au seuil α) que :

$$p \in I_c = [f - d_\alpha, f + d_\alpha]$$

où $d_\alpha = z_\alpha \times \sqrt{\frac{f(1-f)}{n-1}}$ et z_α est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Exemple 3.14

On dispose d'une urne contenant des boules rouges et des boules blanches. Mais on ne connaît pas la composition exacte de l'urne. En 100 tirages avec remise, on obtient 60 boules rouges et 40 boules blanches. A combien peut-on estimer la proportion de boules rouges dans l'urne?

Soit X la v.a.r. qui vaut 1 si on tire une boule rouge et 0 sinon. On sait que X suit une loi de Bernoulli de paramètre p où p est la probabilité de tirer une boule rouge, c'est-à-dire que p est la proportion de boules rouges. Dans cette expérience, on cherche à connaître la valeur de p .

4 Les tests de conformité

Principe

Les tests de conformité permettent de juger de la recevabilité d'une hypothèse concernant un ou plusieurs échantillons. Il s'agit d'une démarche consistant à rejeter ou pas une hypothèse statistique, appelée hypothèse nulle et notée H_0 à un niveau de risque α . Cette hypothèse est que l'observation du ou des échantillons concorde avec la valeur attendue ou théorique. L'hypothèse alternative est notée H_1 .

On suppose que la variable X est distribuée suivant une loi normale de moyenne μ et d'écart-type σ . On dit aussi que X est une variable *gaussienne*.

Rappelons que, si C est quantitatif, la moyenne de $(x_1, x_1, \dots, x_n)$ une observation d'un n -échantillon est notée $\bar{x}$ et son écart-type corrigé est noté s_c . Si C est qualitatif, X suit une loi de Bernouilli $\mathcal{B}(p)$.

Remarque 4.1

Dans ce cours, nous ne ferons que des tests bilatéraux. Autrement dit, l'hypothèse alternative sera toujours la négation de l'hypothèse nulle.

Remarque 4.2

Propriété 4.3

Soit μ_0 une valeur moyenne attendue.

Si X suit une loi normale et si σ est connu, on pose $(z_{obs} =) z_{calcul} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma}$
et $z_{critique}$ est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Si $|z_{calcul}| > z_{critique}$, on rejette l'hypothèse $H_0 : \mu = \mu_0$ au risque d'erreur α .

Exemple 4.4

Propriété 4.5

Soit μ_0 une valeur moyenne attendue.

Si $n > 30$, on pose $(z_{obs} =) z_{calcul} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s_c}$

et $z_{critique}$ est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Si $|z_{calcul}| > z_{critique}$, on rejette l'hypothèse $H_0 : \mu = \mu_0$ au risque d'erreur α .

Exemple 4.6

Propriété 4.7

Soit μ_0 une valeur moyenne attendue.

Si $n \leq 30$ et si σ est inconnu, on pose $(t_{obs} =) t_{calcul} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s_c}$

et $t_{critique}$ est lu dans la table bilatérale de Student avec *d.d.l.* = $n - 1$.

Si $|t_{calcul}| > t_{critique}$, on rejette l'hypothèse $H_0 : \mu = \mu_0$ au risque d'erreur α .

Exemple 4.8

Propriété 4.9

Soit p_0 une valeur attendue.

Si $n \geq 30$, $nf \geq 10$ et $n(1-f) \geq 10$, on pose $z_{calcul} = \frac{f - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

et z_{cri} est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Si $|z_{calcul}| > z_{critique}$, on rejette l'hypothèse $H_0 : p = p_0$ au risque d'erreur α .

Exemple 4.10

La fréquence d'apparition d'une modalité est 0,4 sur un échantillon de 50 individus.

On cherche à savoir au seuil de 5% si l'échantillon pourrait être celui d'une population où $p_0 = 0,5$.

Remarque 4.11

5 Les tests de comparaison

Principe

Les tests de comparaison sont aussi appelés tests d'égalité ou tests d'homogénéité.

On dispose de deux échantillons d'effectifs n_1 et n_2 de deux populations.

On cherche à estimer si un caractère de ces populations pourraient avoir la même valeur moyenne ou la même fréquence (hypothèse nulle) à un niveau de risque α .

Si le caractère est quantitatif, on note μ_1 la moyenne (inconnue) de la première population et μ_2 (inconnue) la moyenne de la seconde population. Ainsi que $\bar{x}_1$ et $\bar{x}_2$ les moyennes respectives des échantillons correspondants. De même, $s_{c,1}^2$ et $s_{c,2}^2$ sont les variances corrigées respectives des échantillons.

Si le caractère est qualitatif, on suppose que la variable de la première population suit une loi $\mathcal{B}(p_1)$ et que la variable de la deuxième population suit une loi $\mathcal{B}(p_2)$ avec p_1 et p_2 inconnus. On note f_1 et f_2 les fréquences respectives de la première modalité dans les échantillons correspondants.

Propriété 5.1

On suppose que $s_{c,1}^2 \geq s_{c,2}^2$ et on pose $F_{calcul} = \frac{s_{c,1}^2}{s_{c,2}^2}$.

On trouve F_{cri} dans la table de Snédécour avec $n_1 - 1$, $n_2 - 1$ et $\frac{\alpha}{2}$.

Si $F_{calcul} > F_{cri}$, on rejette l'hypothèse $H_0 : \sigma_1^2 = \sigma_2^2$ au risque d'erreur α .

Exemple 5.2

On cherche à connaître l'égalité des variances de deux populations. On dispose de deux échantillons respectifs avec les données suivantes :

Echantillon 1 : $n_1 = 10$, $\sum_{i=1}^{10} x_i = 50$ et $\sum_{i=1}^{10} x_i^2 = 290$

Echantillon 2 : $n_2 = 11$, $\sum_{i=1}^{11} x_i = 60,5$ et $\sum_{i=1}^{11} x_i^2 = 387,75$

Propriété 5.3

Si $n_1 > 30$ et $n_2 > 30$, on pose $(z_{obs} =) z_{calcul} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{c,1}^2}{n_1} + \frac{s_{c,2}^2}{n_2}}}$

et $z_{critique}$ est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Si $|z_{calcul}| > z_{critique}$, on rejette l'hypothèse $H_0 : \mu_1 = \mu_2$ au risque d'erreur α .

Exemple 5.4

On cherche à connaître l'égalité des moyennes de deux populations. On dispose de deux échantillons respectifs avec les données suivantes :

$$\text{Echantillon 1 : } n_1 = 35, \sum_{i=1}^{35} x_i = 297,5 \text{ et } \sum_{i=1}^{35} x_i^2 = 2660$$

$$\text{Echantillon 2 : } n_2 = 50, \sum_{i=1}^{50} x_i = 400 \text{ et } \sum_{i=1}^{50} x_i^2 = 3400$$

Propriété 5.5

Si $n_1 \leq 30$ ou si $n_2 \leq 30$ et si σ_1 et σ_2 sont inconnus mais supposés égaux, on pose :

$$s_{c12}^2 = \frac{(n_1 - 1)s_{c,1}^2 + (n_2 - 1)s_{c,2}^2}{n_1 + n_2 - 2} \text{ et } (t_{obs} =) t_{calcul} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{c12}^2}{n_1} + \frac{s_{c12}^2}{n_2}}}$$

et $t_{critique}$ est lu dans la table bilatérale de Student avec $d.d.l. = n_1 + n_2 - 2$.

Si $|t_{calcul}| > t_{critique}$, on rejette l'hypothèse $H_0 : \mu_1 = \mu_2$ au risque d'erreur α .

Exemple 5.6**Propriété 5.7**

Si $n_1 \leq 30$ ou si $n_2 \leq 30$ et si σ_1 et σ_2 sont connus, on pose $(z_{obs} =) z_{calcul} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

et $z_{critique}$ est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Si $|z_{calcul}| > z_{critique}$, on rejette l'hypothèse $H_0 : \mu_1 = \mu_2$ au risque d'erreur α .

Exemple 5.8**Propriété 5.9**

Si $X_1 \sim \mathcal{B}(p_1)$ et $X_2 \sim \mathcal{B}(p_2)$ et si $n_1 f_1$, $n_1(1 - f_1)$, $n_2 f_2$ et $n_2(1 - f_2)$ sont tous plus grand que 5, alors on pose $f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ et $(z_{obs} =) z_{calcul} = \frac{f_1 - f_2}{\sqrt{f(1-f) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$.

Et $z_{critique}$ est lu dans la table la loi normale centrée réduite avec $1 - \frac{\alpha}{2}$.

Si $|z_{calcul}| > z_{critique}$, on rejette l'hypothèse $H_0 : p_1 = p_2$ au risque d'erreur α .

Exemple 5.10

On cherche à connaître l'égalité des proportions d'un caractère qualitatif A dans deux populations. On dispose de deux échantillons respectifs avec les données suivantes :

Echantillon 1 : $n_1 = 100$ dont 63 ayant le caractère A .

Echantillon 2 : $n_2 = 150$ dont 87 ayant le caractère A .