Chapitre 5

Echantillonnage et Estimation

1

Echantillonnage

2

Population

- Une Population est toute collection d'objets à étudier ayant des propriétés communes appelés des individus ou unités statistiques
- Une population peut être infinie, ou finie de taille N.
- La statistique traite des propriétés des populations plus que celles d'individus particuliers.

3

Échantillon

- L'étude de tous les individus d'une population finie s'appelle un recensement. Lorsque l'on observe qu'une partie de la population, on parle de sondage.
- La partie étudiée s'appelle l'échantillon.
- Il existe plusieurs méthode de construction d'un échantillon, dont la plus simple est celle de l'échantillonnage aléatoire simple correspondant à des tirages équiprobables et indépendants les uns des autres.
- Sa taille est notée n<<N
- Dans ces conditions, les observations deviennent des v.a. ainsi que les résumés numérique usuels: ils convient donc d'en chercher les lois de probabilité avant de tenter d'extrapoler (inférés) à la population.

4

Échantillon

- Si on prélève au hasard n individu dans une population finie de taille N et on veut étudier une caractéristique X de la population.
- X est une v.a. appelée v.a. mère ou parente.
- À chaque individu i tiré, on associe une v.a. X_i dont on observe une seule réalisation x_i. Alors les X_i sont des v.a. ayant toutes la même distribution, celle de X.
- On suppose que les X_i sont mutuellement indépendantes (ou au moins, indépendantes deux à deux).

5

Échantillon

- On a donc la double conception suivante: Les valeurs observées $(x_1,x_2,\ldots,x_n)$ constituent n réalisations indépendantes d'une v.a. X ou encore, une réalisation unique du n-uple $(X_1,X_2,\ldots,X_n)$ où les X_i sont n v.a. indépendantes et de même loi.
- On note par la suite un échantillon le n-uple $(X_1,X_2,\ldots,X_n)$.

6

Les statistiques

- La théorie de l'échantillonnage se propose d'étudier les propriétés du n-uple $(X_1,X_2,\ldots,X_n)$ et des caractéristiques le résumant, les statistiques, à partir de la distribution supposée connue de la variable parente X, et d'étudier en particulier ce qui se passe lorsque la taille de l'échantillon est élevée.

7

Les statistiques

- Il est d'usage dans la pratique de résumer les n valeurs d'un échantillon $x_1,x_2,\ldots,x_n$ par quelques caractéristiques simples telles que moyenne, plus grande valeur, etc.
- Ces caractéristiques sont elles-mêmes des réalisations de v.a. issues de $X_1,X_2,\ldots,X_n$.
- Une statistique T est une v.a. fonction mesurable de $X_1,X_2,\ldots,X_n$

$$T=f(X_1,X_2,\ldots,X_n)$$

8

Les statistiques

- **Exemples:**
- La moyenne empirique d'un échantillon $(X_1,X_2,\ldots,X_n)$ est:

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

- Sa variance empirique est:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

9

La distribution de la moyenne

- Pour une réalisation $(x_1,x_2,\ldots,x_n)$, la statistique $\overline{X}$ prendra la valeur $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Cette valeur est la moyenne arithmétique.
- Pour une autre réalisation, dans les mêmes conditions, un deuxième échantillon donnera pour réalisation $(x'_1, x'_2, \ldots, x'_n)$ et $\overline{X}$ prendra alors la valeur $\overline{x'} = \frac{1}{n}\sum_{i=1}^{n} x'_i$

10

Propriétés

1. L'espérance mathématique, notée $\mu_{\bar{X}}$, de $\bar{X}$ est égale à la moyenne m de la population:

$$\mu_{\bar{X}} = m$$

En effet, on a:

$$\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n}(m + m + \ldots + m) = \frac{1}{n} nm = m$$

11

Propriétés

2. La variance de $\bar{X}$, notée $\sigma^2_{\bar{X}}$, est égale à $\frac{\sigma^2}{n}$ où σ^2 est la variance de la population et n la taille de l'échantillon.

En effet, on a:

$$\sigma^2_{\bar{X}} = Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

où les v.a. X_i sont indépendantes

12

Remarques

- La moyenne et la variance de $\overline{X}$ sont calculées pour le cas d'un échantillon de variables aléatoires indépendantes et identiquement distribuées (échantillon tiré avec remise d'une population finie ou échantillon tiré avec ou sans remise d'une population infinie).
- Si l'échantillon est tiré sans remise d'une population finie, les variables ne sont plus indépendantes. Dans ce cas, on a toujours

$$\mu_{\overline{X}} = E(\overline{X}) = m$$

13

mais on trouve un autre résultat pour la variance

$$Var(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

En effet, la population étant de taille N, il y a C_N^n échantillons de taille n et

$$\sigma_{\overline{X}}^2 = Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} Var(X_i) + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} Cov(X_i, X_j)\right]$$

avec $Var(X_i) = \sigma^2$ et $Cov(X_i, X_j) = E[(X_i - m)(X_j - m)]$

14

$$Cov(X_i, X_j) = E\left[(X_i - m)(X_j - m)\right]$$

$$= \sum_{l=1}^{N}\sum_{k=1}^{N}(x_l - m)(x_k - m)P(X_i = x_l; X_j = x_k)$$

$$= \sum_{l=1}^{N}\sum_{k=1}^{N}(x_l - m)(x_k - m)P(X_i = x_l)P(X_j = x_k / X_i = x_l)$$

$$= \sum_{l=1}^{N}\sum_{k=1}^{N}(x_l - m)(x_k - m)\frac{1}{N}P(X_j = x_k / X_i = x_l)$$

$$= \begin{cases} \sum_{l=1}^{N}\sum_{k=1}^{N}(x_l - m)(x_k - m)\frac{1}{N}\frac{1}{N-1} & pour\ k \neq l \\ 0 \quad pour\ \ k = l \end{cases}$$

On a donc: $Cov(X_i, X_j) = \frac{1}{N}\frac{1}{N-1}\sum_{\substack{l,k=1 \\ l\neq k}}^{N}(x_l - m)(x_k - m)$

15

Comme $\left[\sum_{i=1}^{N}(x_i - m)\right]^2 = \sum_{i=1}^{N}(x_i - m)^2 + \sum_{\substack{l,k=1 \\ l\neq k}}^{N}(x_l - m)(x_k - m),$

$$\left[\sum_{i=1}^{N}(x_i - m)\right]^2 = 0$$

et $\sum_{i=1}^{N}(x_i - m)^2 = N\sigma^2,$

on obtient $Cov(X_i, X_j) = \frac{1}{N}\frac{1}{N-1}\left(-N\sigma^2\right)$

et donc $Var(\bar{X}) = \frac{1}{n^2}\left[n\sigma^2 + \sum_{\substack{l,k=1 \\ l\neq k}}^{N}\left(\frac{-\sigma^2}{N-1}\right)\right] = \frac{1}{n^2}\left[n\sigma^2 - \frac{\sigma^2}{N-1}n(n-1)\right]$

D'où $Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

$\frac{N-n}{N-1}$ s'appelle facteur d'exhaustivité ■

16

Propriété:

- On peut affirmer, en vertu des propriétés de la loi normale, que lorsque la population a une distribution normale, la distribution d'échantillonnage de la moyenne est aussi normale.
- Le théorème de la limite centrale nous permet d'affirmer, d'autre part, que quelle que soit la distribution de la population, la distribution de $\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ est normale $N(0,1)$ lorsque n est grand (en pratique ceci est vrai dès que $n>30$).

17

La distribution de la variance

- La variance empirique d'un échantillon aléatoire $(X_1,X_2,\ldots,X_n)$ est défini par:

$$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

- Pour une réalisation $(x_1,x_2,\ldots,x_n)$, la statistique S^2 prendra la valeur $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$. Pour une autre réalisation, dans les mêmes conditions, un deuxième échantillon donnera pour réalisation $(x'_1,x'_2,\ldots,x'_n)$ et S^2 prendra alors la valeur $\frac{1}{n}\sum_{i=1}^{n}\left(x'_i-\overline{x}'\right)^2$ où $\overline{x}'=\frac{1}{n}\sum_{i=1}^{n}x'_i$

18

Propriétés

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

2- $Var(S^2) = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^4\right]$

3- $Cov(\bar{X}, S^2) = \frac{n-1}{n^2}\mu_3$

19

Propriétés

4. Si la distribution de la population est normale, la variable aléatoire $\frac{nS^2}{\sigma^2}$ suit une loi du χ^2 à *n*-1 degrés de liberté: $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$

En effet, on a:

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

$$= \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 - \left(\frac{\bar{X} - m}{\sigma/\sqrt{n}}\right)^2$$

20

Propriétés (suite de 4°)

Le premier terme est une somme de n carrés de variables $N(0,1)$ et suit donc une loi du χ^2 à n degrés de liberté. Le second terme est une variable qui suit une loi du χ^2 à 1 degré de liberté. Donc, le degré de $\frac{nS^2}{\sigma^2}$ est n-1

(on a une relation entre $\overline{X}$ et X_i: $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$).

21

Propriétés (suite de 4°)

- On vérifié que $E(S^2) = \frac{n-1}{n}\sigma^2$

En effet, on a

$$E(S^2) = E\left(\frac{\sigma^2}{n}\frac{nS^2}{\sigma^2}\right) = \frac{\sigma^2}{n}E\left(\frac{nS^2}{\sigma^2}\right) = \frac{\sigma^2}{n}k = \frac{\sigma^2}{n}(n-1)$$

où k est l'espérance mathématique d'une variable aléatoire qui suit une loi du χ^2 à k degrés de liberté (dans ce cas $k=n-1$).

22

Propriétés (suite de 4°)

- De même, on trouve:

$$Var\left(S^2\right)=Var\left(\frac{\sigma^2}{n}\frac{nS^2}{\sigma^2}\right)=\frac{\sigma^4}{n^2}Var\left(\frac{nS^2}{\sigma^2}\right)=\frac{\sigma^4}{n^2}2k=\frac{\sigma^4}{n^2}2(n-1)$$

où $2k$ est la variance d'une variable aléatoire qui suit une loi du χ^2 à k degrés de liberté (dans notre cas $k=n$-1).

23

Propriétés (suite de 4°)

- On peut affirmer de plus que la v.a. $\frac{\overline{X}-m}{\sqrt{\frac{S^2}{n-1}}}$ suit une loi de Student à n-1 degrés de liberté.

En effet, comme $\frac{\overline{X}-m}{\sigma/\sqrt{n}}\sim N(0,1)$ et $\frac{nS^2}{\sigma^2}\sim\chi^2(n-1)$, le rapport $\frac{\frac{\overline{X}-m}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS^2}{(n-1)\sigma^2}}}=\frac{\overline{X}-m}{\sqrt{\frac{S^2}{n-1}}}\sim T(n-1)$.

24

La distribution des fréquences

- La probabilité de réalisation d'un événement est supposée être égale à p.
- On considère les échantillons de taille n extraits, avec remise, d'une population de taille N.
- A chaque échantillon extrait correspond une fréquence f_n de réalisation de l'événement considéré.

25

Propriétés

1. µfn=$E(f_n)=p$

En effet, la variable aléatoire $X=n\ f_n \sim B(n,p)$ et

$$\mu_{f_n} = E(f_n) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

2. $Var(f_n) = \frac{p(1-p)}{n}$

En effet,

$$Var(f_n) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

26

Propriétés

3. Si le tirage se fait sans remise, on a toujours $E(f_n)=p$. Mais la variance dans ce cas, vaut:

$$Var\left(f_n\right)=\frac{p(1-p)}{n}\frac{N-n}{N-1}$$

- **En effet**,

$$Var\left(f_n\right)=Var\left(\frac{X}{n}\right)=\frac{1}{n^2}Var(X).$$

Comme X suit une loi hypergéométrique et

$$Var(X)=\frac{N-n}{N-1}np(1-p)$$

On en déduit la formule de $Var(f_n)$ au dessus.

27

Propriétés

4. Pour une taille n de l'échantillon assez grande (en pratique $n\geq 30$), on a

$$\frac{f_n-\mu_{f_n}}{\sigma_{f_n}}=\frac{f_n-p}{\sqrt{\frac{p(1-p)}{n}}}\sim N(0,1)$$

28

La distribution des différences de moyennes

- On considère 2 populations P_1 et P_2 de taille N_1 et N_2, de moyennes m_1 et m_2 et de variances σ^2_X et σ^2_Y respectivement.
- On s'intéresse, dans de nombreux problèmes à la différence m_1-m_2.
- On extrait de la population P_1 un échantillon $(x_1,x_2,\ldots,x_{n_1})$ de taille n_1 et de la population P_2 un échantillon $(y_1,y_2,\ldots,y_{n_2})$ de taille n_2.
- On note $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ et $\overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$

29

Propriétés

1. $\mu_{\overline{X}-\overline{Y}} = m_1 - m_2$

En effet,

$$\mu_{\overline{X}-\overline{Y}} = E\left(\overline{X}-\overline{Y}\right) = E\left(\overline{X}\right) - E\left(\overline{Y}\right) = m_1 - m_2$$

2. $\sigma^2_{\overline{X}-\overline{Y}} = \frac{1}{n_1}\sigma^2_{\overline{X}} + \frac{1}{n_2}\sigma^2_{\overline{Y}}$

En effet,

$$\sigma^2_{\overline{X}-\overline{Y}} = Var\left(\overline{X}-\overline{Y}\right) = Var\left(\overline{X}\right) + Var\left(\overline{Y}\right)$$

car $\overline{X}$ et $\overline{Y}$ sont indépendantes. On a donc le résultat pour un tirage exhaustif (avec remise)

30

Propriétés

- (suite 2°) Dans le cas d'un tirage non exhaustif (sans remise), il faut tenir compte du coefficient d'exhaustivité car

$$Var(\overline{X}) = \frac{\sigma_X^2}{n_1}\frac{N_1 - n_1}{N_1 - 1} \quad \text{et} \quad Var(\overline{Y}) = \frac{\sigma_Y^2}{n_2}\frac{N_2 - n_2}{N_2 - 1}$$

3. En supposant n_1 et n_2 grands, on peut dire que $\overline{X}$ et $\overline{Y}$ suivent toutes deux des lois normales. Comme elles sont indépendantes, $\overline{X} - \overline{Y}$ suit aussi une loi normale. On conclut donc que $\dfrac{(\overline{X} - \overline{Y}) - \mu_{\overline{X}-\overline{Y}}}{\sigma_{\overline{X}-\overline{Y}}} = \dfrac{(\overline{X} - \overline{Y}) - (m_1 - m_2)}{\sqrt{\dfrac{\sigma_{\overline{X}}^2}{n_1} + \dfrac{\sigma_{\overline{Y}}^2}{n_2}}} \sim N(0,1)$

31

Exercice

On choisit au hasard six nombres parmi les nombres entiers de 1 à 9, chacun de ces nombres a la même probabilité d'être choisi.

Calculer la moyenne et l'écart-type de la distribution d'échantillonnage des moyennes dans les 2 cas:

1. Tirage sans remise.
2. Tirage avec remise.

32

Solution

1. La moyenne de la population est $m = \frac{1+2+\cdots+9}{9} = 5$

Sa variance σ^2 vaut: $\sigma^2 = \frac{1}{9}\left[(1-5)^2 + (2-5)^2 + \cdots + (9-5)^2\right] = 6{,}67$

L'écart-type est $\sigma = 2{,}58$.

Il y a $C_9^6 = 84$ façons de choisir six nombres parmi les 9.

Chacun de ces 84 échantillons possibles a une moyenne $\bar{x} = \frac{1}{n}\sum_{i=1}^{6} x_i$ où x_i (i=1,2,…,6) représente un des 9 nombres.

33

- Par exemple l'échantillon (3, 8, 7, 2, 5, 1) a pour moyenne $\bar{x} = 4{,}33$.
- On obtient ainsi 84 moyennes et la moyenne de la distribution d'échantillonnage des moyennes $\mu_{\bar{X}}$ vaut $\mu_{\bar{X}} = m = 5$.
- La variance de la distribution d'échantillonnage des moyennes est

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{6{,}67}{6}\left(\frac{9-6}{9-1}\right) = 0{,}417$$

- D'où $\sigma_{\bar{X}} = 0{,}645$

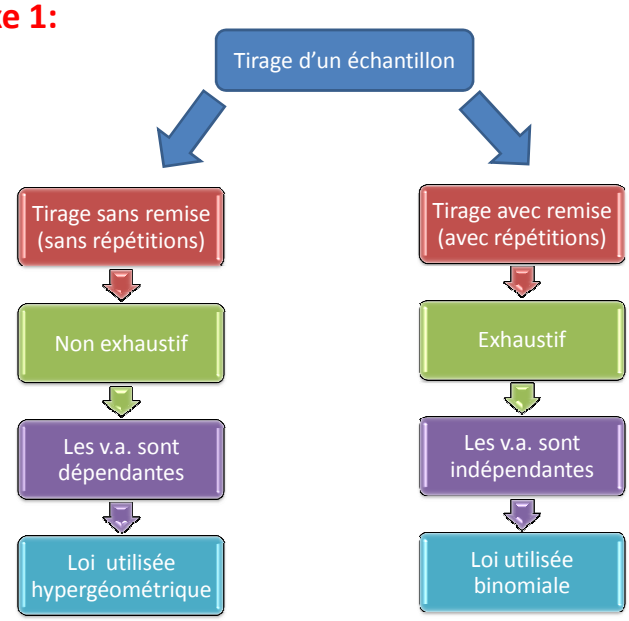
34

2. Il y a 9^6= 531441 façons de choisir six nombres parmi les 9. Chacun de ces échantillons a une moyenne $\overline{x} = \frac{1}{n}\sum_{i=1}^{6} x_i$ où x_i (i=1,2,…,6) représente, comme précédemment, un des 9 nombres.

- Par exemple, l'échantillon (4, 3, 4, 5, 7, 8) donne pour moyenne $\overline{x} = 5{,}17$. On obtient de cette manière 531441 moyennes et la moyenne distribution d'échantillonnage des moyennes $\mu_{\overline{X}}$ vaut $\mu_{\overline{X}} = m = 5$.
- La variance de la distribution d'échantillonnage des moyennes est: $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{6{,}67}{6} = 1{,}11$
- D'où $\sigma_{\overline{X}} = 1{,}05$

35

Annexe 1:



36

Annexe 2: Autre écriture des caractéristiques de la hypergéométrique

On a vu que pour une loi hypergéométrique $H(n,a,b)$

$$E(X)=n\frac{a}{a+b} \quad \text{et} \quad Var(X)=\frac{nab(a+b-n)}{(a+b)^2(a+b-1)}$$

Mais, $a+b=N$ alors,

$$E(X)=n\frac{a}{N} \quad \text{et} \quad Var(X)=\frac{nab(N-n)}{N^2(N-1)}$$

La probabilité de tirer une boule blanche sera:

$$p=\frac{a}{N} \quad \text{et} \quad 1-p=q=1-\frac{a}{N}=\frac{N-a}{N}=\frac{b}{N}$$

$$\Rightarrow \boxed{E(X)=np} \quad \text{et} \quad \boxed{Var(X)=npq\frac{(N-n)}{(N-1)}=np(1-p)\frac{(N-n)}{(N-1)}}$$

37