EPITECH
PROBABILITES ET STATISTIQUES
Cours208
Echantillonnage

Dominique Neveu

Année 2009-2010

Table des matières

1 Echantillonnage **3**

1.1 Méthodes d'échantillonnage 3

1.1.1 Présentation 3

1.1.2 Sondage par quotas 3

1.1.3 Sondage aléatoire simple 4

1.1.4 Sondage aléatoire stratifié 4

1.2 Paramètre et statistique 4

1.3 Distribution d'échantillonnage 5

1.3.1 Définition 5

1.3.2 Dénombrement des échantillons 5

1.4 Distribution d'échantillonnage d'une moyenne 6

1.4.1 Calcul de la moyenne $\mu_{\overline{X}}$ 6

1.4.2 Calcul de la variance $\sigma^2_{\overline{X}}$ 7

1.5 Distribution d'échantillonnage d'une proportion 8

1.5.1 Calcul de la moyenne μ_P 8

1.5.2 Calcul de la variance σ^2_P 9

Résumé du cours

La théorie de l'échantillonnage vise à étudier les relations qui existent entre la population et les échantillons tirés de cette population. L'étude des caractéristiques d'une population à partir d'échantillons tirés de celle-ci, ainsi que l'estimation de la précision de ces estimations grâce à la théorie des probabilités s'appelle l'**inférence statistique**.

On passera en revue les différentes méthodes de sélection d'un échantillon au sein d'une population. Puis, on verra comment estimer un paramètre par une statistique. La valeur inconnue d'une population, à estimer à partir d'un échantillon, est appelée un **paramètre**. La paramètre de la population est estimé à partir d'une **statistique** calculée sur la base d'un échantillon.

A titre d'exemple, on considérera des distributions d'échantillonnage de moyenne ou de proportion.

Chapitre 1

Echantillonnage

1.1 Méthodes d'échantillonnage

1.1.1 Présentation

La théorie de l'échantillonnage vise à étudier les relations qui existent entre la population et les échantillons tirés de cette population. La première étape de l'étude consiste à extraire un échantillon de la population. On distingue deux grandes catégories de méthodes :

- le sondage par **quotas**
- le sondage **aléatoire**

Dans le sondage par quotas, la sélection de l'échantillon n'est pas basée sur des méthodes aléatoires. Il est donc difficile d'évaluer objectivement à quel point l'échantillon est représentatif. Par conséquent, il n'est pas possible de connaître la marge d'erreur des résultats obtenus à partir de l'échantillon.

Le sondage aléatoire correspond à des méthodes de tirage de l'échantillon où chaque individu de la population a une probabilité connue d'être sélectionnée. Ces méthodes permettent non seulement d'estimer les paramètres de la population, mais aussi d'obtenir une mesure de l'erreur commise.

Les deux types de sondages aléatoire les plus courants sont :
- le sondage aléatoire simple,
- le sondage stratifié,

1.1.2 Sondage par quotas

Dans le sondage par quotas, l'enquêteur sélectionne les individus en fonction de quotas qui lui sont donnés. Par exemple, dans le cas d'une enquête auprès des ménages, ces quotas peuvent porter sur des critères socio-

démographiques tels que le sexe, l'âge ou la catégorie socio-professionnelle.

La méthode des quotas est très fréquemment utilisée par les entreprises privées en raison de ses avantages pratiques. Elle permet un gain de temps et elle est moins coûteuse que les sondages probabilistes.

1.1.3 Sondage aléatoire simple

Dans le sondage aléatoire simple, on suppose que tous les éléments de la population ont une probabilité égale de faire partie de l'échantillon. Il s'agit d'un échantillon **sans remplacement** si l'extraction est faite sans remettre les individus sélectionnés dans la population. Sinon, lorsque l'extraction est faite avec remise, l'échantillon est dit **avec remplacement**.

Pour effectuer un sondage aléatoire simple, il faut d'une part, avoir accès à une liste complète des éléments de la population et d'autre part, utiliser une méthode de tirage qui garantisse la même probabilité de sélection à tous les éléments de la liste. On utilise généralement les tables de nombres aléatoires ou des programme de génération de nombres aléatoires.

1.1.4 Sondage aléatoire stratifié

Le sondage stratifié consiste à découper la population en strates ou classes homogènes puis à réaliser dans chaque strate un sondage aléatoire simple. La méthode de sondage stratifié est souvent utilisée lorsque la population est hétérogène à certains égards.

Par exemple, si on a besoin d'obtenir des résultats pour différentes régions géographiques d'un pays, on considère chacune des régions comme une strate et on procède à un sondage aléatoire simple à l'intérieur de chaque strate.

1.2 Paramètre et statistique

Quand on cherche à estimer une valeur inconnue sur une population, on essaye d'estimer cette valeur sur un échantillon. La valeur inconnue sur la population est appelée un **paramètre**. La valeur estimée sur l'échantillon est appelée **statistique**. Souvent le paramètre est une moyenne, un écart-type, une variance, un pourcentage ou un total. Un paramètre est une caractéristique de la population alors qu'une statistique une caractéristique de l'échantillon.

Par exemple, le revenu moyen en France est un paramètre de la population, alors que le revenu moyen d'un échantillon représentatif des Français est une statistique.

Nous nous doutons que le paramètre de la population ne sera pas exactement évalué par la statistique d'un échantillon particulier. Toutefois, il peut en donner une idée approximative.

1.3 Distribution d'échantillonnage

1.3.1 Définition

On considère tous les échantillons de taille n tirés d'une population donnée. Pour chaque échantillon, on peut calculer une statistique (par exemple la moyenne ou une proportion) qui variera avec l'échantillon. On obtient alors une distribution de la statistique que l'on appelle **distribution d'échantillonnage**.

Par exemple, si on utilise la moyenne comme statistique, la distribution s'appelle distribution d'échantillonnage de la moyenne. De la même manière, on peut obtenir des distributions d'échantillonnage pour l'écart-type, les proportions...

On fera attention de bien distinguer la distribution de la population de la distribution d'échantillonnage. La distribution de la population est la distribution de la variable étudiée. Par exemple, le prix d'un article dans un magasin, le revenu d'un ménage dans une ville... La distribution d'échantillonnage est la distribution des statistiques obtenues en considérant tous les échantillons possibles de taille n issus d'une même population.

1.3.2 Dénombrement des échantillons

Considérer tous les échantillons composés de n individus dans une population de taille N revient à considérer toutes les combinaisons possibles de n éléments parmi N éléments. Le nombre d'échantillons est donc donné par la quantité :

$$C_N^n = \frac{N!}{n!(N-n)!}$$

Par exemple, si nous avons une population composée de 10 nombres, et que nous désirons prélever un échantillon aléatoire sans remise de 3 nombres, nous aurons :

$$C_{10}^3 = \frac{10!}{3!(10-3)!} = 120 \text{ échantillons possibles.}$$

N° du nombre	Nombre
1	10
2	15
3	7
4	12
5	11
6	6
7	9
8	13
9	14
10	8

TAB. 1.1 – Table de nombres

1.4 Distribution d'échantillonnage d'une moyenne

On étudie un caractère sur une population et on note μ la valeur moyenne de ce caractère. On choisit d'approcher le paramètre μ par la statistique sur un échantillon donné par la moyenne des valeurs. Pour un échantillon de numéro i, on note $\overline{x}_i$ la moyenne des valeurs de l'échantillon. On définit alors la variable aléatoire $\overline{X}$ égale à chaque moyenne $\overline{x}_i$ sur chaque échantillon.

Exemple 1 *On envisage une étude concernant 10 nombres. Les nombres sont donnés dans le tableau 1.1. Prenons un premier échantillon au hasard composé des nombres n°4, 6 et 7. La moyenne de ces nombres est :*

$$\overline{x}_{91} = \frac{12+6+9}{3} = 9$$

Prenons un autre échantillon au hasard composé des nombres n°1, 3 et 8. La moyenne de ces nombres est :

$$\overline{x}_{13} = \frac{10+7+13}{3} = 10$$

On définit la variable aléatoire $\overline{X}$ qui est la moyenne des nombres sur chaque échantillon. Les valeurs prises par $\overline{X}$ sont les valeurs : $\{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_{120}\}$.

1.4.1 Calcul de la moyenne $\mu_{\overline{X}}$

La moyenne respective de chaque échantillon étant notée $\overline{X}$, nous utiliserons le symbole $\mu_{\overline{X}}$ pour représenter la moyenne des valeurs de $\overline{X}$ sur l'ensemble des échantillons possibles de taille n. On a le résultat suivant :

$$\mu_{\overline{X}} = \mu$$

ce qui signifie que la moyenne de la distribution d'échantillonnage des moyennes est égale à celle de la population.

Exemple 2 (suite de l'exemple 1) *La moyenne μ de l'ensemble des 10 nombres est égale à :*

$$\mu = \frac{10 + 15 + ... + 8}{10} = 10,5$$

La valeur de $\mu_{\overline{X}}$ est obtenue en calculant la moyenne de la distribution des moyennes $\overline{X}$ obtenues à partir de l'ensemble des échantillons de taille 3 tirés parmi les 10 nombres.
Nous donnons le calcul des premiers et dernier échantillons.
Echantillon n°1, 2 et 3 :

$$\overline{x}_1 = \frac{10 + 15 + 7}{3} = 10,67$$

Echantillon n°1, 2 et 4 :

$$\overline{x}_2 = \frac{10 + 15 + 12}{3} = 12,33$$

Echantillon n°8, 9 et 10 :

$$\overline{x}_{120} = \frac{13 + 14 + 8}{3} = 11,67$$

La moyenne de toutes les moyennes donne :

$$\mu_{\overline{X}} = \frac{10,67 + 12,33 + ... + 11,67}{120} = 10,5$$

On vérifie donc que la valeur obtenue $\mu_{\overline{X}}$ est bien égale à la valeur moyenne de la population μ.

1.4.2 Calcul de la variance $\sigma^2_{\overline{X}}$

Nous noterons $\sigma^2_{\overline{X}}$ la variance des différentes valeurs de $\overline{X}$. On peut démontrer que la variance d'échantillonnage est égale à l'expression suivante :

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} . \frac{N - n}{N - 1},$$

où σ^2 est la variance de la population, N est la taille de la population, la fraction

$$\frac{N - n}{N - 1}$$

est un facteur correctif à utiliser pour une population finie.
Dans le cas où la population est infinie, le facteur correctif tend vers 1, et nous obtenons la relation simple :

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Ce résultat peut aussi être utilisé pour une population finie quand la taille de la population est suffisamment grande.

Exemple 3 (suite de l'exemple 1 et 2) *Calculons la variance de la population σ d'après la définition :*

$$\sigma^2 = \frac{(10 - 10,5)^2 + (15 - 10,5)^2 + ... + (8 - 10,5)^2}{10} = 8,25$$

Calculons maintenant la variance de l'ensemble des moyennes d'échantillonnage :

$$\sigma^2_{\overline{X}} = \frac{(10,67 - 10,5)^2 + (12,33 - 10,5)^2 + ... + (11,67 - 10,5)^2}{120} \simeq 2,14$$

En utilisant la formule, on vérifie que :

$$\sigma^2_{\overline{X}} = \frac{8,25}{3} . \frac{10 - 3}{10 - 1} \simeq 2,14$$

On constate que la valeur obtenue est égale à l'arrondi près à celle calculée précédemment directement à partir de l'ensemble des 120 échantillons possibles.

1.5 Distribution d'échantillonnage d'une proportion

On utilise le symbole π pour représenter la proportion d'unités possédant un certain attribut au sein d'une population. Le symbole P est utilisé pour représenter la proportion correspondante au sein de l'échantillon, c'est à dire le nombre d'unités de l'échantillon possédant le caractère étudié, rapporté au nombre total d'unités de l'échantillon.

La valeur de P donne une estimation de la valeur inconnue π. Les propriétés de l'estimateur P s'étudient à partir de la moyenne μ_P et de la variance σ^2_P de la distribution d'échantillonnage.

1.5.1 Calcul de la moyenne μ_P

On peut démontrer que le pourcentage de la population est égale à la moyenne de la distribution d'échantillonnage des proportions. On a donc :

$$\mu_P = \pi$$

Cela indique que le résultat obtenu à partir d'un échantillon aléatoire sera en moyenne égal à la valeur recherchée de la population.

Exemple 4 (suite de l'exemple 1) *Dans l'exemple de la population représentée par les 10 nombres, examinons la proportion de nombres ayant une valeur supérieure ou égale à 10. Le résultat est présenté dans le tableau 1.2.*

N° du nombre	Nombre	Nombre ≥ 10
1	10	vrai
2	15	vrai
3	7	faux
4	12	vrai
5	11	vrai
6	6	faux
7	9	faux
8	13	vrai
9	14	vrai
10	8	faux

TAB. 1.2 – Proportion de nombres ≥ 10

Le pourcentage de la population est donné par :

$$\pi = \frac{\textit{nombre de vrai}}{10} = \frac{6}{10} = 0,6$$

D'autre part, la moyenne des 120 proportions d'échantillons se calcule à partir de la liste des échantillons. Donnons la valeur de la proportion pour les premiers et dernier échantillon.
Echantillon n°1, 2 et 3 :

$$p_1 = \frac{2}{3} = 0,67$$

Echantillon n°1, 2 et 4 :

$$p_2 = \frac{3}{3} = 1$$

Echantillon n°8, 9 et 10 :

$$p_{120} = \frac{2}{3} = 0,67$$

La moyenne des proportions des 120 échantillons est donnée par :

$$\mu_P = \frac{0,67 + 1 + ... + 0,67}{120} = \frac{72}{120} = 0,6$$

On vérifie bien l'égalité de la proportion de la population avec la moyenne de la distribution d'échantillonnage des proportions.

1.5.2 Calcul de la variance σ_P^2

Dans le cas d'une proportion, la variance d'une loi de Bernouilli est donnée par :

$$\sigma_\pi^2 = \pi.(1 - \pi)$$

On peut démontrer la formule suivante liant la variance σ_P^2 de la distribution d'échantillonnage de la population originale :

$$\sigma_P^2 = \frac{\sigma_\pi^2}{n}.\frac{N-n}{N-1}$$

De même que pour le calcul de l'écart-type de la distribution d'échantillonnage des moyennes, le facteur correctif présenté ci-dessus n'est significatif que pour le cas d'une population finie. Il peut être supprimé lorsque la population est infinie ou suffisamment grande.

Exemple 5 (suite de l'exemple 1 et 4) *Reprenons l'exemple des 10 nombres traité précédemment, et effectuons le calcul de* σ_π^2 *:*

$$\sigma_\pi^2 = 0,6.(1-0,6) = 0,24$$

La formule donne alors la valeur de σ_P^2 *:*

$$\sigma_P^2 = \frac{0,24}{3}.\frac{10-3}{10-1} \simeq 0,062$$

Si on avait effectué le calcul direct d'après la liste des 120 proportions des échantillons, on aurait obtenu :

$$\sigma_P^2 = \frac{(0,67-0,6)^2 + (1-0,6)^2 + ... + (0,67-0,6)^2}{120} \simeq \frac{7,47}{120} \simeq 0,062$$

Ce résultat correspond bien à celui obtenu précédemment.