Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Batna 2
Faculté des Mathématiques et Informatique
Département de Statistique et Science des Données
3eme année Statistique et Analyse des Données
Année universitaire :2023-2024

وزارة التعـــليم العـــالي والبحث العلـــمي
جامعـــــة بـــاتنة 2
كلـــية الرياضيـــات والإعـــلام الآلـــي
قـــسم الإحصـــــاء و علم البيانا ت
السنة الثالثة إحصاء وتحليل المعطيات
السنة الجامعية : 2024-2023

Sondage

Dr. Allaeddine HADDARI

Sondage

Qu'est-ce qu'un sondage

Il existe deux approches pour connaître les caractéristiques statistiques d'un caractère sur une population.

1. **Le recensement** est l'approche descriptive. Il consiste à mesurer le caractère sur toute la population.
2. **Le sondage** est l'approche inférentielle. Lorsque le recensement n'est pas possible pour des raisons de coût, de temps ou à cause de certaines contraintes (test destructif par exemple), on a recours à un sondage, c'est-à-dire à l'étude statistique sur un sousensemble de la population totale, appelé échantillon. Si l'échantillon est constitué de manière correcte, les caractéristiques statistiques de l'échantillon seront proches de celles de la population totale.

Exemple :

Je désire connaître l'âge moyen de tous les étudiants de l'Université de Batna 2.

- Recensement : je demande l'âge à tous les étudiants et je calcule la moyenne... ça risque d'être long!!!
- Sondage : je choisis une partie des étudiants (échantillon), je calcule la moyenne des âges sur cette partie en espérant que cette moyenne soit "proche" de l'âge moyen de tous les étudiants.

Nous voyons sur cet exemple que la mise au point d'un sondage nécessite plusieurs choix pour le statisticien :

• comment choisir les étudiants ?

• combien d'étudiants doit-on choisir ?

• comment doit-on formuler la réponse :

— sous la forme d'une valeur, c'est à dire que l'on donne une estimation de l'âge moyen sous la forme d'un réel (23.6 ans par exemple) ;

— sous la forme d'un ensemble de valeurs. On pourra par exemple donner une fourchette ou un intervalle ([22.3 ;25.4] par exemple).

• est-ce que l'estimation est satisfaisante ? Dit autrement suis-je capable de donner une estimation de l'erreur commise par la prédiction. On pourra par exemple dire " l'âge moyen des étudiants de l'Université de Batna 2 se trouvent dans l'intervalle [22.3 ;25.4] avec un niveau de confiance de 95%.".

L'objectif de ce cours consiste à étudier des procédures de sondage pour lesquelles nous pourrons répondre à ces questions. Nous allons dans ce résumé présenter le contexte et proposer différentes méthodes de sondage permettant d'estimer des moyennes et proportions.

Echantillonnage Aléatoire Simple

1 Introduction

Avant d'envisager une méthode de sondage, on doit définir soigneusement la population à étudier et suivre correctement les étapes de la sélection d'un échantillon. Le choix de l'unité d'échantillonnage doit être bien étudié:

- Au cours d'une enquête sociale, des individus seront interrogés, le feront-ils pour eux ou pour le ménage?
- Au cours d'une enquête économique, des entreprises peuvent être étudiées, qui interroge-t-on? le responsable des ressources humaines, les employés, ...

Pour définir la base de sondage, on doit établir une liste d'individus qui répondent aux conditions précédentes. Deux cas se présentent soit on a la dite liste ou bien non. Dans le premier cas, les méthodes probabilistes sont généralisables, par contre dans le second cas, les conditions réunies sont insuffisantes et l'échantillonnage prendra un caractère approximatif et fondamentament non probabiliste.

L'existence de la base de sondage détermine en grande partie le choix de la méthode d'échantillonnage.

2 Approche probabiliste

L'échantillonnage probabiliste entraine la sélection d'un échantillon à partir d'une population, une sélection qui repose sur le hasard ou la chance. Comme les unités de la population sont sélectionnées au hasard et qu'il est possible de calculer la probabilité d'inclusion de chaque unité dans l'échantillon, on peut grâce à l'échantillonnage probabiliste, produire des estimations fiables, de même que des estimations de l'erreur d'échantillonnage et faire alors des inférences au sujet de la population.

Il existe plusieurs méthodes permettant de sélectionner un échantillon probabiliste. La méthode qu'on choisira dépendra de plusieurs facteurs: la base de sondage, la distribution de la population,... .

Le but est de réduire le plus possible l'erreur d'échantillonnage des estimations pour les variables d'enquête les plus importantes, en réduisant le plus possible le délai et le coût de réalisation de l'enquête. Les méthodes les plus courantes sont l'échantillonnage aléatoire simple avec et sans remise.

2.1 Echantillonnage aléatoire simple

Dans un échantillonnage aléatoire simple (EAS) ou (SAS), tous les échantillons possibles de même taille ont la même probabilité d'être choisis et tous les éléments de la population ont une chance égale de faire partie de l'échantillon. L'échantillonnage aléatoire simple peut s'effectuer avec ou sans remise.

1) Echantillonnage aléatoire simple avec remise (SASAR):

Chaque individu tiré est remis dans la base d'échantillonnage.

Dans ce cas, il y a équiprobabilité et indépendance entre chaque tirage. Cette méthode fournit une estimation satisfaisante de la moyenne réelle m de la population, et la probabilité p d'apparition d'un caractère dans la population. On peut donner un encadrement appelé intervalle de confiance de l'estimateur dès que la taille n de l'échantillon est suffisament grande.

L'intervalle de confiance pour la moyenne m de la population:

$$I_m = [\overline{X} - u_\alpha \frac{\sigma}{\sqrt{n}}; \overline{X} + u_\alpha \frac{\sigma}{\sqrt{n}}]$$

L'intervalle de confiance pour la probabilité p de la population:

$$I_p = [f - u_\alpha \sqrt{\frac{f(1-f)}{n}}; f - u_\alpha \sqrt{\frac{f(1-f)}{n}}]$$

2) Echantillonnage aléatoire simple sans remise (SASSR):

Chacun des individus tiré est sorti de la base sondage.

Dans le tirage sans remise, les expériences ne sont plus équiprobable, ni indépendantes, mais on montre qu'il n'y a pas de différences fondamentales dans les résultats statistiques.

De même que dans le SASAR, on peut déterminer les intervalles de confiance de m et p.

La distribution de la moyenne $\overline{X}$ est gaussienne de moyenne m et de variance $\frac{N-n}{N-1}\frac{\sigma^2}{n}$, d'où on obtient

$$I_m = [\overline{X} - u_\alpha \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}; \overline{X} + u_\alpha \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}]$$

Si l'échantillon n est petit par rapport à la taille de la population N, alors le facteur dit facteur d'exhaustivité $\frac{N-n}{N-1}$ est négilgé.

De même pour l'intervalle de confiance de la probabilité p d'apparition d'un caractère dans la population, on rajoute le facteur d'exhaustivité.

Remarque:
L'échantillonnage aléatoire simple est la méthode la plus facile à appliquer et la plus couramment utilisée. Ceci vient du fait qu'elle n'exige pas de données supplémentaires autres que la liste complète des membres de la population observée. Il existe pour cette méthode des formules types pour déterminer la taille de l'échantillon, les estimations et les erreurs d'échantillonnage.

3 Calcul de la taille n de l'échantillon:

3.1 Cas d'une proportion:

Soit P la proportion de l'apparition d'un caractère donné A dans une population. Soit f la fréquence estimée de ce caractère sur un échantillon aléatoire de taille n.

L'intervalle de confiance pour la proportion P pour un niveau de confiance $(1-\alpha)$ est donné par:

$$\left[f - u_\alpha \sqrt{\frac{f(1-f)}{n}}, \quad f + u_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$$

On note par d la précision de cette estimation:

$$d = \pm u_\alpha \sqrt{\frac{f(1-f)}{n}}$$

D'où la taille n de l'échantillon sera égale à:

$$n = \frac{f(1-f)}{d^2} u_\alpha^2$$

Remarque

Pratiquement on ne possède pas d'estimation de la fréquence, on ne peut pas faire le calcul a priori de la taille de l'échantillon sans ces informations. Dans ce cas, on considère le cas le plus défavorable, c'est-à-dire le cas où $f(1-f)$ est maximale d'où $f = 0,5$;

Ce qui donne:

$$n = \frac{1}{4d^2} u_\alpha^2$$

3.2 Cas d'une moyenne:

Dans le cas du calcul de la taille de l'échantillon pour estimer une moyenne m d'une population, la démarche est identique. A partir de l'intervalle de confiance, la précision est donnée par:

$$d = u_\alpha \frac{\sigma}{\sqrt{n}}$$

où σ est remplacé par s qui représente l'écart type de l'échantillon. Ce qui nous donne:

$$n = \frac{s^2}{d^2} u_\alpha^2$$

Remarque:

Etant donné que la variance est inconnue, alors on a deux possibilités pour l'estimer:

- Soit on utilise le fait que l'étendue e d'une série statistique est généralement inférieure à 6 fois l'écart type s donc:

$$e \leq 6s \Rightarrow n = u_\alpha^2 \frac{[(\max - \min)/6]^2}{d^2}$$

- Soit on procède à un échantillonnage en deux temps: le premier échantillonnage sera fait sur un échantillon de faible taille qui nous permet de faire une estimation des paramètres clés. Ceci nous permet de définir de manière plus ou moins précise la taille de l'échantillon de l'enquête principale.

4 Efficacité d'un sondage

En général, les estimateurs que l'on doit comparer sont en moyenne égaux aux paramètres à estimer. Ils ne diffèrent que par leur variance. On définit l'efficacité d'une méthode de sondage sur une autre par le rapport des variations d'échantillonnage. On utilise l'effet de sondage défini par:

$$D(m^*/m) = \frac{var(\widehat{\theta}^*)}{var(\widehat{\theta})}$$

Si $D(m^*/m) \prec 1$, alors la méthode m^* est plus précise que m.

Exemple:

Pour comparer les deux méthodes d'échantillonnage aléatoire simple avec et sans remise dans l'estimation de la moyenne de la population, on a:

SASAR: $var(\overline{X}) = \frac{\sigma^2}{n}$

SASSR: $var(\overline{X}^*) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

$$D(m^*/m) = \frac{N-n}{N-1}$$

D'où: si $n \succ 1$, $D(m^*/m) \prec 1$. L'estimateur obtenu par la méthode SASSR est plus précis que celui obtenu par la méthode SASAR.

Remarque:

Dans le cas d'un estimateur quelconque, on mesure la précision par l'écart quadratique moyen:

$$EQM(\widehat{\theta}) = var(\widehat{\theta}) + B^2(\widehat{\theta})$$

où $B(\widehat{\theta})$ est le biais de l'estimateur $\widehat{\theta}$.

5 Echantillonnage systématique

Ce type d'échantillonnage est appelé aussi échantillonnage par intervalles du fait qu'il existe un intervalle entre chaque unité sélectionnée qui est incluse dans l'échantillon.

5.1 Méthode de sélection d'un échantillon

Pour sélectionner un échantillon systématique de taille n, il faut suivre les étapes suivantes:

1) Numéroter de 1 à N les unités incluses dans la base de sondage (N est la taille de la population).

2) Déterminer l'intervalle d'échantillonnage k tel que $k = \frac{N}{n}$ où n est la taille de l'échantillon que l'on veut obtenir. Si k n'est pas entier, on prend la partie entière de ce nombre.

3) Sélectionner au hasard un nombre a entre 1 et k qu'on appellera origine. L'origine a sera choisie de manière aléatoire.

.

4) L'échantillon est alors constitué des unités : $\{a, a+k, a+2k, a+3k, ..., a+(n-1)k\}$.

Remarques:

1. Chaque membre de la population ne peut faire partie que de l'un des échantillons.

2. Chaque unité ne peut faire partie que d'un seul échantillon.

3. Cette méthode est souvent utilisée dans le contrôle de qualité en industrie, c'est-à-dire dans la sélection des unités pour des essais.

5.2 Avantages et inconvénients:

- Le plus gros avantage de l'échantillonnage systématique est que la sélection de l'échantillon est très facile, il suffit de choisir un seul nombre aléatoire qui constitue l'origine.

- Le plus gros inconvénient est que les échantillons choisis peuvent ne pas être représentatifs de la population s'il existe un cycle.

Sondage stratifié

1 Introduction:

Dans un sondage aléatoire simple, toutes les combinaisons de n unités de l'échantillon parmi N unités de la population ont la même probabilité d'être choisies, mais certaines d'entre elles peuvent être indésiables. Par exemple si on est en présence que des unités qui constituent les extrémités de la série statistique. Pour exclure les échantillons extrêmes et améliorer la précision des estimateurs, on a recours à l'échantillonnage stratifié.

Le but de ce type d'échantillonnage est de:

1. découper la population en sous-ensembles appelés "strates", les plus homogènes possibles.

2. Ces strates s'excluent mutuellement, c'est-à dire qu'une unité ne peut faire partie que d'une seule strate à la fois.

3. On sélectionne à partir de chaque strate des échantillons indépendants.

5. On utilise n'importe quelle méthode d'échantillonnage pour sélectionner l'échantillon à l'intérieur de la strate.

Remarque:
Le but d'un sondage stratifié est de réduire les coûts d'enquête et de l'optimiser. Donc, les échantillons sont sélectionnés selon des facteurs susceptibles d'expliquer les différences ou bien selon le découpage géographique. La stratification dépend de la population étudiée.

2 Notations:

Supposons que la population est subdivivée en k strates numérotées de 1 à k.

Pour toute strate h, on note:

- N_h l'effectif de la strate h,

- $\overline{X}_h$ la moyenne d'une variable d'intérêt X dans la strate h,

- la variance estimée de la variable X dans la strate h est égale à:

$$S^2_{h,c} = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} (X_k - \overline{X}_h)^2$$

- L'effectif de l'échantillon choisi dans la strate h est noté n_h.

- Le taux de sondage correspondant à la state h est $f_h = \frac{n_h}{N_h}$.

- La moyenne de l'échantillon dans la strate h est égale à:

$$\overline{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_i$$

- La variance corrigée dans la strate h est égale à:

$$s^2_{h,c} = \frac{1}{n_h - 1} \sum_{i=1}^{N_h} (x_i - \overline{x}_h)^2$$

3 Estimation

3.1 Estimation de la moyenne m de la population

L'estimateur de la moyenne m de la population dans un sondage stratifié est défini par:

$$\overline{X}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \overline{X}_h$$

On montre que cet estimateur est sans biais.

Sur un échantillon de taille $n = \sum_{h=1}^{k} n_h$, l'estimateur de la moyenne est donné par:

$$\overline{x}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \overline{x}_h$$

La variance de cet estimateur est donnée par:

$$var(\overline{x}_{st}) = \sum_{h=1}^{k} (\frac{N_h}{N})^2 (1 - f_h) \frac{s_{h,c}^2}{n_h}$$

D'où l'on peut déduire un intervalle de confiance pour la moyenne m de la population, pour un niveau de confiance $(1 - \alpha)$:

$$I_m = \left[\overline{x}_{st} \pm u_\alpha \sqrt{var(\overline{x}_{st})} \right]$$

3.2 Estimation de la proportion P de la population

De même que pour la moyenne, on définit un estimateur de la proportion P d'une population par:

$$\overline{p}_{st} = \sum_{h=1}^{k} \frac{N_h}{N} \overline{p}_h$$

de variance:

$$var(\overline{p}_{st}) = \sum_{h=1}^{k} (\frac{N_h}{N})^2 (1 - f_h) \frac{p_h(1 - p_h)}{n_h}$$

On en déduit un intervalle de confiance pour la proportion P de la population, pour un niveau de confiance $(1 - \alpha)$:

$$I_m = \left[\overline{p}_{st} \pm u_\alpha \sqrt{var(\overline{p}_{st})} \right]$$

4 Répartition des individus entre les states

Dans un sondage stratifié, il faut déterminer des strates les plus homogènes possibles pour obtenir une représentation fidèle à la population par rapport au sujet étudié. Pour cela, on tient compte des considérations suivantes pour choisir les critères de répartition des strates:

- Disponibilité des critères sur la base de sondage.

-.Pertinence des différents critères pour créer des strates homogènes ceci repose

- soit sur une connaissance intuitive,
- soit selon des études réalisées auparavant.

Les formules d'estimation sont valables quelque soit les nombres d'unités tirées par strate; ce qui fait que le taux de sondage f_h peut être variable d'une strate à une autre.

4.1 Sondage stratifié proportionnel

Cette répartition consiste à imposer le même taux de sondage f_h pour toutes les strates, c'est-à-dire on pose:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

Le sondage est alors appelé sondage stratifié proportionnel et l'échantillon est dit représentatif.

$$\overline{X}_{st} = \sum_{h=1}^{k} \frac{n_h}{n} \overline{x}_h = \frac{1}{n} \sum_{h=1}^{k} \sum_{i=1}^{h} x_i$$

où n est la taille de l'échantillon. Ainsi on obtient la moyenne simple calculée sur la population.

4.2 Répartition optimale

Cette répartition appelée répartition de Neyman consiste à imposer cette égalité:

$$\frac{n_h}{N_h S_{h,c}} = \frac{n}{\sum_{h=1}^{k} N_h S_{h,c}}$$

La théorie montre que c'est la répartition qui donne la variance la plus faible.

Remarque:
Cette stratification suppose connues a priori les valeurs des variances estimées. Ceci est possible dans le cas où nous avons des études antérieures réalisées sur le sujet.

En pratique, cette formule est utilisée quand le phénomène étudié a une distribution dissymétrique. Dans le cas contraire, un sondage stratifié proportionnel fournit des résultats satisfaisants.

Principes de stratification

1) La précision de l'estimateur dans un sondage stratifié est toujours meilleure ou égale à celle d'un sondage aléatoire simple.

2) Le principe d'un sondage stratifié est basé sur:
- Forcer le hasard
- Imposer à l'échantillon de représenter la population strate par strate.

3) On considère généralement comme critères de stratification:
- des critères d'une typologie (par exemple le type d'industrie, ...),
- des critères de taille (par exemple les chiffres d'affaires des entreprises, les tranche d'âges des individus).

4) Pour des sondages géographiques, on utilise la stratification selon la région par exemple ou selon les activités de la population, ou autres.