L'inférence statistique, estimation et sondages

MICHEL HENRY*

Résumé
Dans cet article on propose une initiation aux démarches de la statistique inférentielle: échantillons aléatoires et propriétés des statistiques standard, notamment des moyennes et des variances. Ces propriétés sont ensuite appliquées à une introduction à la théorie de l'estimation et aux tests d'hypothèses.
Mots-clés: statitsque inférentielle; enseignement-aprendizagem; test d'hypothèse; population; échantillon.

Resumo
Neste artigo discute-se a iniciação à estatística inferencial: amostragens aleatórias e propriedades de estatísticas standard, em especial, médias e variâncias. Essas propriedades são, depois, aplicadas a uma introdução à teoria da estimação e aos testes de hipóteses.
Palavras-chave: *estatitística inferencial; ensino-aprendizagem; teste de hipótese; população; amostragem.*

Abstract
In this paper, we discuss initiation into inferential statistics: random sampling and properties of standard statistics, specially means and variance. Those properties are, afterwards, applied to the introduction to estimation theory and to hypotheses testing.
Key-words: *inferential statistics; teaching and learning; hypothesis test; population; sample.*

Principes de l'inférence statistique

Populations statistiques et échantillons

On désire mieux connaître une **population** *P* (d'êtres vivants ou d'objets, comme par exemple une production de marchandises, ou tout autre ensemble relevant d'études statistiques) et notamment certaines de ses caractéristiques.

* IREM de Franche-Comté. E-mail: michel.henry@univ-fcomte.fr

L'investigation exhaustive n'est généralement pas possible: coûts exorbitants, inaccessibilité de l'ensemble *P*, temps disponible pour obtenir les résultats de l'étude...

On peut penser par exemple à la différence de coût et de méthode entre un recensement de la population française (périodiquement nécessaire) et un sondage dont la pratique dépend d'hypothèses générales basées sur un recensement antérieur.

L'étude de la population conduit à définir des **caractères** qui attribuent aux éléments de *P* certaines qualités ou valeurs qui résument les caractéristiques étudiées. L'ensemble des valeurs prises par un caractère χ fournit une **série statistique** qui peut être répartie en **classes** et représentée par divers **diagrammes**.

Quand le caractère χ est **quantitatif**, la série peut à son tour être résumée par certains **paramètres**, notamment la **moyenne** μ de χ et son **écart-type** σ. Ces valeurs sont en général inconnues lorsque l'on entreprend une étude statistique. L'un de ses objets est d'en donner des approximations suffisantes pour prendre des décisions.

Quand χ est **qualitatif**, on s'intéresse aux proportions dans *P* de ses diverses modalités. Par exemple, on désire connaître le poids p de l'une d'entre elles (situation des sondages).

Le principe de l'inférence statistique est d'obtenir des informations sur la population *P* (population «mère») à partir de la connaissance d'un échantillon *E*.

Précisons d'abord ce que l'on entend par «**échantillon**».

Dans la réalité statistique, un n-échantillon est un ensemble de n objets prélevés dans la population. Quand le prélèvement se fait au hasard, on obtient un **échantillon aléatoire**.

Quand la population *P* est partagée en strates, dont les poids respectifs sont connus (à partir d'un recensement), on peut décider a priori de composer l'échantillon proportionnellement aux poids des strates dans la population. On dit alors qu'on a un «**échantillon représentatif**» de *P.* Ce n'est pas toujours le meilleur quant aux résultats que l'on veut en tirer[1].

---

1 Car les valeurs du caractère sur certaines strates peuvent présenter une trop grande dispersion, ce qui peut faire surestimer le poids de la strate et augmenter la dispersion résultante. Or cette dispersion intervient pour déterminer un encadrement de la valeur pour la population P du paramètre étudié.

Pour appliquer certains théorèmes de probabilité, il convient de faire l'hypothèse que les éléments de l'échantillon *E* sont prélevés indépendamment les uns des autres. Pour cela, afin que les prélèvements successifs ne modifient pas la composition de la population, il faudrait les effectuer *avec remises*, ce qui n'est pas toujours possible. Dans la pratique, les échantillons sont plutôt prélevés *sans remises*. Lorsque la population est vaste par rapport à la taille n de l'échantillon (au moins 1000 fois plus grande), cet inconvénient est mineur, entachant les probabilités en jeu d'une erreur négligeable.

En inférence statistique, quand on s'intéresse à un caractère quantitatif χ, on peut **estimer** (évaluation statistique) les valeurs des paramètres qui résument la distribution du caractère χ sur la population *P*, à partir des valeurs de χ observées sur l'échantillon *E*. On peut aussi estimer la valeur de la proportion p des éléments de P qui appartiennent à une certaine modalité d'un caractère qualitatif χ, à partir de la fréquence f de cette modalité dans *E*. C'est notamment la situation des sondages aléatoires simples.

On peut aussi émettre des **hypothèses** concernant *P* et tester la validité de ces hypothèses à partir des renseignements que fournit l'échantillon (tests statistiques).

Quand l'échantillon est aléatoire, **cette inférence est en partie déterminée par le hasard** du prélèvement. Ainsi, toute déclaration à propos de la population, issue de l'observation d'un échantillon, est entachée d'une certain **risque** (probabilité) de se tromper. Le problème de l'inférence statistique est de pouvoir donner des résultats suffisamment précis avec un risque minimisé, deux contraintes qui varient en sens contraire. Pour améliorer à la fois la **précision** des résultats et la **fiabilité** des déclarations relatives à ces résultats, le statisticien ne peut qu'agrandir la taille de son échantillon, ce qui coûte plus cher.

Modèle probabiliste de l`échantillonnage

Nous allons maintenant construire le modèle mathématique général pour décrire une situation d'échantillonnage en précisant le problème:

Soit E un échantillon aléatoire de taille n, extrait de la population P.
Le caractère χ étudié a pour moyenne μ et pour écart type σ dans P.

Les éléments de P qui donnent à χ *une certaine modalité A sont en proportion p.*
La moyenne de χ *sur E est* m_n *et son écart type est* s_n.
La fréquence de la modalité A dans E est f_n.
Quel lien y a-t-il entre les valeurs inconnues μ, σ, *p et les valeurs observées* m_n, s_n *et* f_n *?*

Dans le modèle probabiliste, on représente le *prélèvement au hasard* d'un élément de la population *P* par un élément ω d'un **ensemble référentiel** Ω, que l'on supposera fini. Ω symbolise donc à la fois les éléments de *P* et le fait qu'ils peuvent être issus d'un tirage aléatoire dans lequel les éléments de *P* ont la même "chance" d'être choisis.

Pour représenter un caractère quantitatif χ étudié (pour simplifier, on le supposera de dimension 1), on considère une **variable aléatoire** (v.a.) X_0, définie sur Ω, prenant pour chaque **éventualité** ω, la valeur du caractère χ pour l'élément de *P* dont ω représente le tirage au hasard. Pour un caractère qualitatif, X_0 prendra les valeurs 1 ou 0 suivant que pour cet élément, χ prend ou non la modalité *A*.

La variable X_0 est censée décrire la répartition des valeurs de χ, avec leurs poids respectifs. Sa **loi** (ensemble des probabilités des événements associés aux valeurs possibles de χ) modélise cette répartition.

Dans le cas d'un caractère quantitatif, on a noté μ la moyenne de χ sur la population *P* et σ^2 sa variance. La loi de X_0 est inconnue, mais on fait l'hypothèse que ces valeurs μ et σ^2 sont les valeurs de l'**espérance mathématique** et de la **variance** de cette loi: $E(X_0) = \mu$ et $Var(X_0) = \sigma^2$.

Dans le cas d'une proportion p, la loi de X_0 est donnée par $IP(X_0 = 1) = p$ (probabilité de tirer au hasard de *P* un élément présentant la modalité *A*) et on a:

$$E(X_0) = p \text{ et } Var(X_0) = p(1 - p).$$

La réalisation de l'échantillon *E* fournit donc un n-uplet de valeurs observées de χ: $x = (x_1.., x_i.., x_n)$, interprétées comme images d'événements $\{\omega_i \in \Omega\}$ par l'application répétée n fois de X_0. On considère que ces x_i sont les réalisations (observations) de n v.a. X_i qui représentent cette réplique successive de X_0.

On se place donc dans une *hypothèse de travail*: les éléments de *E* sont prélevés au hasard et indépendamment les uns des autres (notamment la taille de *P* est assez grande par rapport à celle de *E*).

Cette hypothèse de travail se traduit par une *hypothèse de modèle*: les X_i sont des variables aléatoires indépendantes (au sens probabiliste) et de même loi[2] que X_0. Notamment, on a pour tous les i: $E(X_i) = E(X_0)$ et $Var(X_i) = Var(X_0)$.

L'échantillonnage est donc représenté par un **vecteur aléatoire** $X = (X_1,..., X_i,..., X_n)$ vérifiant cette hypothèse de modèle. Le résultat effectif de cet échantillonnage est donc un échantillon *E*, représenté par les valeurs observées $(x_1..., x_i..., x_n)$. X est encore appelé un «**échantillon de la v.a. X_0**» et X_0 est appelée «**variable parente**» de l'échantillon X.

Par définition, dans le modèle de la statistique:

> ***Un échantillon est un n-uplet de variables aléatoires, définies sur le même Ω, indépendantes et de même loi.***

Modélisation des paramètres standards

Pour un caractère quantitatif, la moyenne m_n de χ sur l'échantillon *E* est obtenue dans le modèle probabiliste par la moyenne arithmétique $\bar{x} = \frac{1}{n}\sum x_i$ (les probabilistes ont l'habitude de représenter cette moyenne par le symbole $\bar{x}$). $\bar{x}$ est donc la valeur observée d'une variable aléatoire $\overline{X} = \frac{1}{n}\sum X_i$ définie sur Ω^n, par l'intermédiaire de X.

La variance sur *E* du caractère χ est aussi un paramètre important. Sa valeur est donnée par: $\sigma_n^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$. C'est la valeur observée[3] de la variable aléatoire $V = \frac{1}{n}\sum (X_i - \overline{X})^2$ également définie sur Ω^n, par l'intermédiaire de X.

Pour l'étude d'une proportion, $\bar{x}$ représente aussi la fréquence observée f de la modalité *A* dans l'échantillon *E*, puisque le $\sum$ contient

---

2 X_i est de même loi que X_0 si pour tout x, $P(X_i = x) = P(X_0 = x)$. Les X_i sont indépendantes si pour tout n-uplet $(x_1.., x_i.., x_n)$, on a $P[(X_1,..., X_i,..., X_n) = (x_1.., x_i.., x_n)] = P[X_1 = x_1].P[X_2 = x_2]...P[X_n = x_n]$.

3 Rappelons que les valeurs $\overline{X}$ et σ_n^2 sont issues d'un échantillonnage aléatoire et ne coïncident pas avec les moyenne et variance inconnues m et σ^2 de χ sur la population P.

autant de nombres 1 qu'il y a d'éléments dans E de modalité A. Comme Var $(X_0) = p\,(1 - p)$, p est le seul paramètre inconnu pour caractériser la population du point de vue de la modalité A. L'introduction de la variable V est alors inutile.

Ces variables $\overline{X}$ et V sont aussi appelées des «résumés statistiques» (on dit plus brièvement «statistiques»[4]), car leurs réalisations combinent les valeurs de χ sur E.

L'étude des lois des variables $\overline{X}$ et V, en fonction éventuellement d'hypothèses supplémentaires sur la répartition de χ dans P, est une partie importante de la statistique inférentielle. Elles jouissent de propriétés générales obtenues à partir de certains théorèmes puissants de la théorie des probabilités. Nous utiliserons les suivants:

Espérance mathématique

i – L'espérance mathématique est une forme linéaire sur l'espace des v.a. définies sur le même Ω. Si Ω est un ensemble fini sur lequel une v.a. Y prend les valeurs y_k et dont la loi est donnée par les probabilités $p_k = IP(Y = y_k)$, on a $E(Y) = \sum p_k y_k$.

ii – Si Y et Z, définies sur Ω, sont deux variables indépendantes (tout événement lié à l'une est indépendant de tout événement lié à l'autre), on a: $E(Y.Z) = E(Y).E(Z)$.

Variance

Avec les hypothèses et notations précédentes, la variance de Y est définie par $Var(Y) = \sum p_k (y_k - E(Y))^2$. On a pour α réel: $Var(\alpha Y) = \alpha^2$ Var(Y).

Si Y et Z sont indépendantes, on a: $Var(Y+Z) = Var(Y) + Var(Z)$.

Inégalité de Bienaymé-Tchebychev

On montre assez facilement l'inégalité suivante: $P(|Y - E(Y)| > \varepsilon) \leq \dfrac{Var(Y)}{\varepsilon^2}$.

4 Une statistique est donc une application réelle définie sur Ω^n, composée de X et d'une application de IR^n dans IR.

Cette inégalité permet de majorer la probabilité de se tromper (le risque que l'on prend) quand, à partir d'une observation y de Y, on affirme que la valeur moyenne E(Y) est dans l'intervalle]y-ε, y+ε[. Elle permet donc d'avoir une idée de la fiabilité de cette affirmation. Notons que la majoration de cette probabilité varie comme la dispersion (variance) de Y.

Propriétés des statistiques $\overline{X}$ et V

– $\overline{X}$, **moyenne de l'échantillon**, a des propriétés sympathiques. La linéarité de l'espérance mathématique donne immédiatement:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}.n.E(X_0) = E(X_0)$$. D'où le résultat:

$$E(\overline{X}) = \begin{cases} \mu & \text{pour un caractère quantitatif} \\ p & \text{pour un caractère qualitatif} \end{cases}$$

- Les propriétés de la variance, compte tenu de l'indépendance des X_i, donnent aussi:

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{Var(X_0)}{n}$$.

D'où:

$$Var(\overline{X}) = \begin{cases} \dfrac{\sigma^2}{n} & \text{pour un caractère quantitatif} \\ \dfrac{p(1-p)}{n} & \text{pour un caractère qualitatif} \end{cases}$$

– Application à la précision de mesures physiques.

Anticipons un peu sur le paragraphe consacré à l'estimation par intervalle de confiance, pour montrer comment ces résultats justifient une intuition forte: sans changer d'appareil de mesure, on peut améliorer notablement sa **précision** (marge d'erreur sur le résultat annoncé) en faisant la moyenne de plusieurs mesures indépendantes.

Supposons que la variable parente X_0 représente la mesure d'une grandeur (physique par exemple) inconnue μ. L'inégalité de Bienaymé-Tchebychev, appliquée à X_0 avec $\varepsilon = \varepsilon'\sigma$, donne:

$$P[X_0 - \varepsilon'\sigma \leq E(X_0) \leq X_0 + \varepsilon'\sigma] \geq 1 - \frac{1}{\varepsilon'^2}.$$

L'intervalle $[x_0 - \varepsilon'\sigma\ ;\ x_0 + \varepsilon'\sigma]$ obtenu pour encadrer $\mu = E(X_0)$ à partir d'une simple observation x_0 de X_0, a une longueur proportionnelle à l'écart type σ de X_0, pour une **fiabilité** $1 - \frac{1}{\varepsilon'^2}$ donnée (minorant de la probabilité d'annoncer un bon encadrement).

Appliquée à $\overline{X}$, avec $\varepsilon = \frac{\varepsilon'\sigma}{\sqrt{n}}$, on a: $P[\overline{X} - \frac{\varepsilon'\sigma}{\sqrt{n}} \leq E(\overline{X}) \leq \overline{X} + \frac{\varepsilon'\sigma}{\sqrt{n}}] \geq 1 - \frac{1}{\varepsilon'^2}$.

La «**fourchette**» $]\bar{x} - \frac{\varepsilon'\sigma}{\sqrt{n}}\ ;\ \bar{x} + \frac{\varepsilon'\sigma}{\sqrt{n}}[$ proposée cette fois-ci pour l'encadrement de cette valeur inconnue $\mu = E(\overline{X})$, montre que l'utilisation de $\overline{X}$ plutôt que X_0 divise la marge d'erreur par $\sqrt{n}$. On voit aussi que, pour un n donné, on améliore la précision (on resserre la fourchette en diminuant ε') en consentant une perte de fiabilité (en $1 - \frac{1}{\varepsilon'^2}$): on ne peut pas avoir le beurre et l'argent du beurre ! Même remarque pour l'estimation d'une proportion.

- Espérance de V

On est dans le cas où χ est un caractère quantitatif.

La variance $V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ de l'échantillon est moins agréable que sa moyenne. Pour l'étudier, il est commode d'introduire la variance de χ restreinte à l'échantillon. Pour cela, on considère la statistique notée $\sum^2$, définie par:

$$\sum^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

On a la relation[5]: $V = \sum^2 - (\overline{X} - \mu)^2$.

---

5 On trouve le lien entre V et $\sum^2$ en ramenant dans V les variables Xi et $\overline{X}$ à leur espérance m, en écrivant:

$Xi - \overline{X} = (X_i - m) + (m - \overline{X})$ et en développant le carré:

$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - m)^2 + 2(X_i - m)(m - \overline{X}) + (\overline{X} - m)^2]$, d'où

$V = \sum^2 + \frac{2}{n}(n\overline{X} - nm)(m - \overline{X}) + (\overline{X} - m)^2 = \sum^2 - (\overline{X} - m)^2$.

Pour calculer E(V), on utilise les propriétés de linéarité de l'espérance mathématique:

$$E(V) = E(\sum^2) - E[(\overline{X} - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n} E[(X_i - \mu)^2] - E[(\overline{X} - \mu)^2] = Var(X_i) - Var(\overline{X})$$

D'où:

$$E(\sum^2) = \sigma^2 \text{ et } E(V) = \frac{n-1}{n}\,\sigma^2$$

Estimation

Principe de l`estimation

L'estimation d'un paramètre résumant les valeurs d'un caractère dans une population, consiste à donner pour la valeur τ prise par ce paramètre pour la population, une valeur approchée t_n, calculée à partir d'un échantillon, avec autant de précision que possible.

On utilise pour cela une statistique T appropriée, en retenant pour τ la valeur $t_n = T(x)$, où x est l'observation de l'échantillon X.

Par exemple, pour estimer une moyenne μ ou une proportion p, on utilisera $\overline{X}$. Pour estimer une dispersion, on peut penser à la statistique V (qui n'est pas la meilleure!). Ces variables prennent alors le nom d'«**estimateurs**».

Pour une **estimation ponctuelle**, on se contente des valeurs observées sur l'échantillon: $\overline{x}$ valeur de $\overline{X}$ pour estimer des valeurs théoriques m_n ou f_n (moyenne μ d'un caractère quantitatif ou fréquence p d'un caractère qualitatif), σ_n^2 de V pour estimer la dispersion σ^2 du caractère dans la population.

Le contrôle de la marge d'erreur renvoie à une **estimation par intervalle**. Par exemple, comment majorer la probabilité de se tromper en donnant pour une moyenne inconnue m un encadrement issu de la valeur observée $\overline{x}$ sur l'échantillon: $P(|\overline{X} - \mu| > \varepsilon) \leq \alpha$?

Un niveau de confiance $1 - \alpha$ étant donné (0,9 ou 0,95 suivant les degrés de fiabilité que l'on souhaite), si l'on sait calculer cette probabilité, on peut obtenir une valeur minimale pour ε (le 1/2 écartement de la **fourchette**), telle qu'avec une probabilité meilleure que $1 - \alpha$, ou puisse affirmer que μ est dans l'intervalle $]\,\overline{x} - \varepsilon\,;\,\overline{x} + \varepsilon\,[$.

La détermination de cet ε dans diverses situations concrètes est le problème du calcul des **intervalles de confiance**.

Estimation ponctuelle

Quand on donne une valeur expérimentale, issue de l'observation d'un échantillon aléatoire pour un paramètre inconnu, on peut se demander:

En quoi cette estimation ponctuelle est-elle une «bonne» estimation?

On aimerait que cette inférence statistique jouisse des deux propriétés suivantes:

i – Si, pour chaque échantillon, chaque valeur observée t peut être différente de la valeur τ à estimer (cela dépend du hasard de l'échantillonnage), dans l'ensemble des échantillons de taille n possibles, on aimerait que ces valeurs se répartissent autour de τ de telle sorte qu'en moyenne, elles donnent τ. Autrement dit, on souhaite que $E(T) = \tau$. On dit alors que l'estimateur T est «**sans biais**».

ii – On aimerait aussi que la précision et la fiabilité de l'estimation s'améliorent en augmentant la taille de l'échantillon (cf. l'exemple ci-dessus des mesures physiques). On dit alors que T est un estimateur **convergent**: plus on prend de grands échantillons, plus les valeurs observées t sont proches de τ, ou du moins la probabilité qu'elles s'en rapprochent tend vers 1.

Pour vérifier ces propriétés, on doit faire appel aux résultats théoriques en probabilités que l'on a rappelés et aux hypothèses de modèle. Voyons de plus près ce qu'il en est pour les deux estimateurs $\overline{X}$ et V introduits.

a) Estimation d'une moyenne ou d'une proportion

Dans le cas d'un caractère χ quantitatif, pour estimer la **moyenne** μ de χ, prenons l'estimateur $\overline{X}$. Comme $E(\overline{X}) = \mu$, $\overline{X}$ est un *estimateur sans biais* de μ.

De plus, $Var(\overline{X}) = \frac{\sigma^2}{n}$. L'inégalité de Bienaymé-Tchebychev s'écrit:

$$P(|\overline{X} - m| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2},$$

ce qui montre que $\overline{X}$ est un *estimateur convergent*.

De même, pour estimer une **proportion** p, l'estimateur $\overline{X}$ est performant:

comme $E(\overline{X}) = p$, $\overline{X}$ est encore un *estimateur sans biais*, et $Var(\overline{X}) = \frac{p(1-p)}{n}$ montre qu'il est convergent.

On retrouve un résultat attendu: pour estimer la probabilité d'un événement A, on peut prendre la **fréquence** des réalisations de A lors de la répétition d'un (assez) grand nombre d'expériences dont les issues constituent l'échantillon *E*.

Ce résultat n'est pas une démonstration de la loi des grands nombres, mais une confirmation de l'efficacité du modèle probabiliste qui permet notamment de démontrer l'inégalité de Bienaymé-Tchebychev.

Historiquement ce résultat a été le premier théorème important du calcul des probabilités, démontré par Jacques Bernoulli dans *Ars Conjectandi* (1713), connu aujourd'hui sous le nom de "loi faible des grands nombres". Le théorème de Bernoulli montre que la fréquence observée est un estimateur sans biais convergeant pour la probabilité d'un événement.

Mais il ne permet pas évaluer assez finement l'erreur commise par cette estimation ponctuelle, c'est à dire de donner un encadrement de confiance pour cette probabilité. Un tel encadrement (intervalle de confiance) repose sur un théorème dû à Moivre et Laplace (démontré en 1814), généralisé aujourd'hui sous le nom de "théorème limite central". C'est ce théorème qui fait de la statistique inférentielle un outil puissant.

b) Estimation d'une variance

Par contre, $E(V) = \frac{n-1}{n}\sigma^2$, V est donc un estimateur biaisé de σ^2.

Mais $E\left(\frac{n}{n-1}V\right) = \sigma^2$, et par conséquent la statistique $S^2 = \frac{n}{n-1}V$, qui s'écrit aussi:

$$S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2,$$

est un estimateur sans biais de σ^2.

Habituellement, on désigne par s^2 ou σ^2_{n-1} la valeur observée de S^2 sur l'échantillon et on l'appelle la **variance estimée** de la population (les calculettes statistiques présentent les deux valeurs σ_n et σ_{n-1}, ça fait

plus riche). $s^2 = \sigma^2_{n-1}$ est donc une estimation ponctuelle sans biais de la variance σ^2 du caractère χ dans la population. On démontre que S^2 est aussi un estimateur convergent de σ^2.

Remarque: $E(S^2) = \sigma^2$ n'entraîne pas $E(S) = \sigma$, S n'est pas un estimateur sans biais de σ!

Estimation par intervalle de confiance

Niveau de confiance

On veut estimer la valeur d'un paramètre τ relatif à un caractère χ défini sur une population *P*. Une estimation ponctuelle à partir d'un échantillon ne renseigne pas sur la précision de l'approximation de τ. On voudrait donc obtenir un «intervalle aléatoire», pas trop grand, à partir de l'échantillon prélevé, tel que la probabilité qu'il contienne τ soit acceptable.

Cette probabilité sera appelée «**niveau de confiance**» de l'estimation, on la désigne par $1 - \alpha$. Le nombre α est le risque que l'on prend de se tromper en affirmant que τ est bien dans l'intervalle proposé.

Pour préciser cela, prenons un niveau de confiance de 90%. A chaque échantillon correspond la valeur observée t de l'estimateur T utilisé.

On considère l'intervalle centré en t: $]t-\varepsilon, t+\varepsilon[$, où ε est choisi de sorte qu'en moyenne, pour 9 échantillons sur 10, τ soit dans $]t-\varepsilon, t+\varepsilon[$.

Autrement dit, on désire trouver ε tel que $IP(\tau \in]T-\varepsilon ; T+\varepsilon[) \geq 0{,}9$. On a rencontré cette situation dans le cas où τ est l'espérance mathématique de la variable parente. On a vu que l'inégalité de Bienaymé-Tchebychef donne alors une solution, mais celle-ci se révèle peu performante. Pour avoir un bon résultat, le calcul de cette probabilité fait nécessairement intervenir la loi de T. L'intervalle aléatoire $]T - \varepsilon, T + \varepsilon[$ est appelé **intervalle de confiance** pour τ de niveau $1 - \alpha$.

L'intervalle réel $]t - \varepsilon, t + \varepsilon[$ est «l'observation de l'intervalle de confiance»sur l'échantillon considéré, ou la «**fourchette**». On ne sait pas avec certitude si τ est dedans.

Estimation d'une proportion p.

Soit une population dans laquelle une modalité *A* d'un caractère qualitatif est en proportion p. Dans un prélèvement au hasard d'un élément de cette population, la probabilité qu'il présente la modalité *A* est donc p, valeur que l'on désire estimer.

De manière perceptive, on sait que lorsqu'on répète cette expérience un grand nombre n de fois, la fréquence f_n de réalisations de *A* "tend à se stabiliser", et f_n peut être observée aussi proche de p que l'on veut, pourvu que n soit assez grand.

Il s'agit de donner un sens plus précis à cette affirmation et de montrer comment on peut contrôler l'erreur $|f_n - p|$, commise en prenant f_n comme valeur pour p que l'on cherche ainsi à estimer expérimentalement.

La situation des sondages aléatoires simples est une application la plus directe de ce problème. Si, dans une population statistique, une proportion p d'éléments sont d'une certaine modalité *A*, un prélèvement aléatoire d'un échantillon de taille n dans cette population peut renseigner sur p (on suppose que la population est assez vaste pour pouvoir considérer ce prélèvement comme non exhaustif, i.e. "avec remises"). Par exemple, *A* peut être le choix préférentiel d'un consommateur ou d'un électeur.

Le fait pour chaque élément observé e_i de cet échantillon, prélevé au hasard, d'appartenir à la modalité *A* est un événement E de probabilité p. Cette situation est modélisée par une variable de Bernoulli X_i, qui prend la valeur 1 si E est réalisé, avec probabilité p, et 0 avec probabilité 1 – p sinon, d'espérance $E(X_i) = p$ et de variance $Var(X_i) = p\,(1–p)$.

La variable $\sum X_i$ est égale au nombre des partisans de *A* dans l'échantillon. C'est une variable binomiale *B*(n; p).

La fréquence $F_n = \frac{1}{n}\sum X_i$ des éléments de modalité *A* dans l'échantillon, dépend de l'aléa du prélèvement. C'est une variable aléatoire d'espérance:

$$E(F_n) = \frac{1}{n}\sum E(X_i) = p \text{ et de variance } Var(F_n) = \frac{1}{n^2}\sum Var(X_i) = \frac{p(1-p)}{n}$$

(si l'on suppose que les X_i sont indépendantes: les e_i sont prélevés indépendamment les uns des autres). F_n est donc un estimateur sans biais et convergent de p.

La valeur f_n observée sur un échantillon est donc prise pour estimer la valeur de la proportion p inconnue (estimation ponctuelle).

Exemple:

Un sondage effectué auprès de 150 personnes choisies de façon aléatoire dans une circonscription donne 45 suffrages au candidat A. Quelle est la proportion p des partisans de A dans la population ?

On choisit f_n comme estimation ponctuelle de la proportion p d'électeurs favorables au candidat *A*. La valeur observée de la fréquence de *A* dans l'échantillon est ici: $f_n = 0{,}3$.

Mais l'écart $|f_n - p|$ dépend aussi de l'aléa du prélèvement. On ne peut donc espérer obtenir qu'un contrôle probabiliste a priori de $|F_n - p|$, majorant l'erreur que l'on fera en prenant pour p la valeur $f_n$de la fréquence de l'événement E dans l échantillon prélevé.

L'inégalité de Bienaymé-Tchebycheff donne une réponse théorique à cette évaluation. Appliquée à la fréquence F_n, elle s'écrit: pour tout $\varepsilon > 0$: $P(|F_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2}$, probabilité qui tend donc vers 0 quand n tend vers l'infini. On dit que "la fréquence F_n tend vers p en probabilité". C'est le théorème de Bernoulli, forme la plus simple de la loi (faible) des grands nombres. L'inégalité s'écrit aussi:

$$P(F_n - \varepsilon < p < F_n + \varepsilon) \geq 1 - \frac{p(1-p)}{n\varepsilon^2}$$

Posant $\alpha = \frac{p(1-p)}{n\varepsilon^2}$, l'intervalle $]\, F_n - \varepsilon, F_n + \varepsilon \,[$ est donc un intervalle de confiance pour p de niveau $1 - \alpha$.

Mais avec $\alpha = 0{,}05$, $p = 1/2$ et $\varepsilon = 0{,}01$ (estimation de p à 1 % près) par exemple, il faut un échantillon de taille $n = 50\,000$!

Pour avoir un meilleur résultat, il nous faut utiliser la loi de F_n, moyenne des X_i. Or la loi binomiale de n F_n se prête mal aux calculs.

Le théorème de Moivre-Laplace (forme particulière d'un théorème puissant des probabilités, le théorème limite central) permet une bien meilleure estimation. Ce théorème dit que pour $n > 50$ (ce qui n'est pas trop demander), on fait une erreur négligeable sur la valeur de la probabilité $P(|F_n - p| < \varepsilon)$ en considérant que F_n suit une loi normale $N\left(p\,;\, \frac{p(1-p)}{n}\right)$. La condition de confiance $P(|F_n - p| < \varepsilon) = 1 - \alpha$ s'écrit alors:

$$P(|U| < \varepsilon) = 1 - \alpha,$$

où $U = \frac{(F_n - p)\sqrt{n}}{\sqrt{p(1-p)}}$ est une variable normale centrée réduite. Si u_α désigne le fractile de cette loi tel que $P(|U| < u_\alpha) = 1 - \alpha$, on a $P(|F_n - p| < u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}) = 1 - \alpha$,

Dans la pratique, pour $n > 50$, la valeur inconnue p peut être remplacée par son estimation ponctuelle, et on obtient l'intervalle de confiance pour p de niveau $1 - \alpha$:

$$]F_n - u_\alpha \frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} \; ; \; F_n + u_\alpha \frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}} \; [$$

Avec l'observation f de F_n, la demi-longueur de l'intervalle de confiance est alors $\varepsilon = u_\alpha \sqrt{\frac{f(1-f)}{n}}$ et la fourchette obtenue pour estimer la proportion p au niveau de confiance $1-\alpha$ est donc:

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

Dans l'exemple, on a la valeur observée $f = 0{,}3$ pour F_n et une taille $n = 150$ pour l'échantillon. Avec $\alpha = 0{,}05$, on obtient $u_\alpha = 1{,}96$, ce qui donne la fourchette]0,226 ; 0,374[pour estimer p au niveau de confiance 0,95.

Remarquons qu'en pourcentage, p est donnée à 7,4 % près avec 5 chances sur 100 de se tromper, ce qui n'est pas fameux pour un sondage. L'échantillon est trop petit pour estimer assez précisément cette proportion. En prenant $n = 1000$, on obtient la fourchette]0,271 ; 0,329[susceptible d'encadrer p à environ 3 % près, ce qui est la performance habituelle des sondages médiatisés.

Remarquons aussi que la précision de cette estimation ne dépend pas de la taille de la population *P*, et qu'elle est inversement proportionnelle à la racine carrée de la taille n de l'échantillon sondé.

Pour simplifier un peu grossièrement, on peut aussi majorer $\sqrt{f(1-f)}$ par 0,5, ce qui, en augmentant la valeur calculée pour ε, garantit un niveau de confiance supérieur à $1-\alpha$. Cette majoration n'est pas trop brutale: pour $f = 0{,}3$, on a $\sqrt{f(1-f)} = 0{,}46$ et pour $f = 0{,}1$, $= \sqrt{f(1-f)}$ 0,3.

Avec cette simplification, la condition $P(|U| < u) \geq 1 - \alpha$ donne $\varepsilon < \frac{u}{2\sqrt{n}}$, et puisque pour $\alpha = 0{,}05$ on a $u = 1{,}96$, ε est majoré par $\frac{1}{\sqrt{n}}$. On peut donc donner pour fourchette de sondage, au niveau de confiance $1 - \alpha = 0{,}95$, l'intervalle classique:

$$\left]f - \frac{1}{\sqrt{n}} \; ; \; f + \frac{1}{\sqrt{n}}\right[.$$

Il suffit donc d'un échantillon de taille $n > 10\,000$ (c'est encore cher) pour estimer p à 1 % près, avec une probabilité 0,95 de ne pas se tromper. Si on accepte une estimation de p à 3 % près, il suffit que $n > 1\,112$, taille approximative des sondages les plus courants.

Le tableau suivant donne les demi-fourchettes de sondage ε pour estimer une proportion p en pourcentages, en fonction de différentes valeurs de n et de α,

n α	500	800	1 000	2 000	10 000
0,1	3,7	2,9	2,6	1,8	0,8
0,05	4,4	3,5	3	2,2	1
0,01	5,7	4,5	4	2,9	1,3

Annexe

Lois des grands nombres et théorèmes de convergences

L`inégalité de Bienaymé – Tchebychev

Soit Y une variable aléatoire dont l'espérance et la variance existent. Pour tout $\varepsilon > 0$, on a

$$IP(|Y - E(Y)| \geq \varepsilon) \leq \frac{Var(Y)}{\varepsilon^2}$$

ce qui s'écrit aussi:

$$IP(Y - \varepsilon < E(Y) < Y + \varepsilon) \geq 1 - \frac{Var(Y)}{\varepsilon^2}$$

Le théorème de Bernoulli

Avec $Y = F_n$, on a $E(F_n) = p$ et $Var(F_n) = \frac{p(1-p)}{n}$, d'où pour tout $\varepsilon > 0$,

$$IP(|F_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \longrightarrow 0 \text{ quand } n \longrightarrow \infty$$

On a la condition de confiance:

$$IP(F_n - \varepsilon < p < F_n + \varepsilon) \geq 1 - \frac{p(1-p)}{n\varepsilon^2}$$

Pour un risque $\alpha = \frac{p(1-p)}{n\varepsilon^2}$ donné, la précision $\varepsilon = \frac{\sqrt{p(1-p)}}{\alpha\sqrt{n}}$ varie comme $1/\sqrt{n}$

Loi faible des grands nombres (Bernoulli 1713, Poisson 1837)

Enoncé moderne:

Si $(X_n)_{n\in IN}$ est une suite de v.a. définies sur Ω, indépendantes, de mêmes espérances $E(X_n) = m$ et de variances finies telles que $\frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \xrightarrow[n\to\infty]{} 0$, alors la suite des moyennes arithmétiques $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converge en probabilité vers m.

(On dit que Y_n converge en probabilité vers Y si pour tout $t > 0$, $IP(Y_n - Y > t) \xrightarrow[n\to\infty]{} 0$).

Démonstration: on a

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = m,$$

et du fait de l'indépendance des X_i,

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)\,.$$

D'après l'inégalité de B.T., on a pour tout $\varepsilon > 0$:

$$\mathrm{IP}\,(|\,\overline{X} - m\,| \geq \varepsilon) \leq \frac{\sum \mathrm{Var}(X_i)}{n^2\varepsilon^2} \longrightarrow 0 \text{ quand } n \longrightarrow \infty$$

Rq: il suffit pour cela que les $\mathrm{Var}(X_i)$ soient égales à une même valeur σ^2 (cas des échantillons), ou seulement bornées.

Pour un échantillon de Bernoulli, les X_i sont des v. a. de Bernoulli et $\sum X_i$ est le nombre de succès dans l'échantillon, d'où $Fn = \overline{X}$ et la loi des grands nombres se réduit au théorème de Bernoulli.

Loi forte des grands nombres (Borel 1909)

Si $(X_n)_{n\in IN}$ *est une suite de v.a. définies sur* Ω*, indépendantes, d'espérances* m *et de mêmes variances finies, alors la suite des moyennes arithmétiques* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ *converge presque sûrement vers* m. (i.e. $\mathrm{IP}(\{\omega \in \Omega \mid \overline{X}_n(\omega) \longrightarrow m\}) = 1$).

Autrement dit: sous ces hypothèses, on a strictement aucune chance de tomber sur une série d'observations dont la suite des moyennes arithmétiques ne converge pas.

Ceci s'applique à la suite des fréquences des succès observés dans un schéma de Bernoulli.

Théorème limite central ou central limite (Laplace 1821, Markov 1898)

Sous les hypothèses des lois des grands nombres, on sait que la suite des moyennes $\overline{X}_n$ converge (en probabilité et presque sûrement) vers m.

Mais il convient de contrôler l'écart $|\,\overline{X}_n - m\,|$, du moins la probabilité qu'il ne dépasse pas un ε. Pour avoir assez de précision, il faudrait connaître la loi de cet écart.

On a le théorème clé de la statistique:

Si $(X_n)_{n\in IN}$ *est une suite de v.a. définies sur* Ω*, indépendantes, de mêmes lois, d'espérance* m *et d'écart type fini* σ*, alors la suite des moyennes arithmétiques*

centrées $Y_n = \frac{\overline{X}_n - m}{\sigma}\sqrt{n}$ *converge en loi vers une v.a.* U*, normale centrée réduite,* $N(0, 1)$.

(i.e. la suite des fonctions de répartition des lois des Y_n converge simplement vers la fonction de répartition de la loi normale centrée réduite: pour tout y réel,

$$IP(Y_n \leq y) \longrightarrow IP(U \leq y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}}\, e^{\frac{1}{2}x^2}\, dx)$$

Ainsi pour un échantillon de Bernoulli, $Y_n = \frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n}$ et pour $n > 50$, la loi de Y_n est assez proche d'une loi normale centrée réduite. Autrement dit, la loi de F_n est proche de la loi normale

$N\left(p, \frac{p(1-p)}{n}\right)$, on a: $IP\left(|F_n - p| < \frac{u_\alpha\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1 - \alpha$.

Pour $1 - \alpha = 0{,}95$, on a $u_\alpha = 1{,}96$, et comme $p(1-p) < 1/4$, la précision ε est meilleure que $1/\sqrt{n}$.

Théorème de Moivre – Laplace (Moivre 1718, Laplace 1812)

En cherchant à améliorer les résultats de Bernoulli, Moivre dans sa *Doctrine of Chances*, publié en 1718, avait entrevu une forme particulière du T L C. Il a précisé ensuite ce résultat en 1756 qui fut entièrement démontré par Laplace dans sa *Théorie Analytique des Probabilités* en 1814.

Théorème:

Soit X_n *une variable binomiale* $B(n, p)$ et k_n *une suite d'entiers telle que* $n(k_{n/n} - p)^3 \to 0$ quand $n \to \infty$, *(ce qui suppose que* $k_{n/n} \to p$*), alors*

$$IP(X_n = k_n) \sim \frac{e^{\frac{1}{2}\frac{(k_n - np)^2}{np(1-p)}}}{\sqrt{2\pi}\sqrt{np(1-p)}}$$

et $$IP(a \leq X_n \leq b) \sim \frac{1}{\sqrt{2\pi}} \int_{\frac{a-np-1/2}{\sqrt{np(1-p)}}}^{\frac{b-np+1/2}{\sqrt{np(1-p)}}} e^{\frac{x^2}{2}}\, dx$$

Fréquences cumulées et médianes

Fonction de répartition empirique (ou fréquence cumulée)

C'est par definition la v.a.:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{]-\infty,\, x]}(X_i) = \frac{\text{nombre des } X_i \le x}{n} = \text{fréquence des } X_i \le x$$

Les $1_{]-\infty, x]}(X_i)$ (fonctions indicatrices) sont des variables de Bernoulli indépendantes de loi B (1, IP($X_0 \le x$)), d'où le *théorème de Glivenko-Cantelli* (aussi appelé le théorème fondamental de la statistique):

$F_n(x)$ *est une suite de v.a. qui tend en probabilité et presque sûrement vers la fonction de répartition* $F_{X_0}(x) = IP(X_0 \le x)$ *de la loi de* X_0, *et de plus, on a la convergence presque uniforme:*

$$\sup_{x \in IR} \left| F_n(x) - F_{X_0}(x) \right| \xrightarrow{p.s.} 0 .$$

Médianes empiriques

La médiane η de la loi de X_0 est définie par $IP(X \le \eta) = 1/2 = F_{X_0}(\eta)$.

La médiane empirique η_n est un nombre qui partage l'échantillon ordonné

$X_{(1)} \le ... \le X_{(k)} \le ... \le X_{(n)}$ en deux parties de même cardinal. On a la propriété:

Si F_{X_0} *est à densité strictement positive, alors* η_n *tend presque sûrement vers* η.

Bibliographie

CHAPUT, B. (2003). *Probabilités au lycée*. Comission Inter-Irem Statistique e Probabilités. Paris, APMEP (n. 143).

DRESS, F. (1997). *Probabilités Statistique*, Paris, Dunod.

HARTHONG, J. (1996). *Probabilités & statistiques*, Paris, Diderot.

HENRY, M. (org.) (1992/1994). *CII Statistique et Probabilités*. Actes de l´Universités d'Été de Statistique Inférentielle. La Rochelle, 1-5 Septembre, Rouen 29 Août-2 Septembre, Irem de Rouen Éditeur.

HENRY, M. (org.) (2001). *Autour de la modélisation en probabilities*. Besançon, Presses Universitaires Franc-Comtoises.

KAPADIA, R. et BOROVCNIK, M. (1991). *Chance Encounters: Probability in Education*. Dordrecht, Dordrecht: Kluwer Academic Publishers.

KAUFFMANN, P. (1994). *Information, Estimation, Tests*. Paris, Dunod.

SAPORTA, G. (1990). *Probabilités, Analyse des données et Statistique*. Paris, Technip.

Recebido em abr./2005; aprovado em jun./2005.