Estimateurs

Leçon 17
EN2D2- Lycée Gustave Eiffel - Bordeaux
Thierry Sageaux.

Résumé

Cette leçon constitue une introduction aux statistiques inférentielles.

Table des matières

I Introduction **1**

II Estimer la proportion p **2**

III Estimer μ **5**
III.1 Quand σ est connu 5
III.2 Quand σ n'est pas connu 5

IV Estimer σ^2 **5**
IV.1 Une battle variance empirique - variance corrigée 5
IV.2 Si μ est connue 6
IV.3 Si μ est inconnu 7

V Application classique **8**

I Introduction

Le problème est le suivant : on a une population supposée très grande et l'on veut étudier un caractère qui suit la même loi pour chacun des n individus de la population. On note $(X_i)_{i=1}^n$ les variables aléatoires de chacun des individus de l'échantillon. Ne connaissant pas les paramètres de la loi en question suivie par les X_i, on cherche à les estimer. C'est ce que l'on appelle les **statistiques inférentielles**.

La plupart du temps, on cherche :

- la proportion p,
- l'espérance mathématique μ,
- la variance σ^2.

On utilise donc une série statistique obtenues sur un échantillon de la population. On a naturellement des estimateurs des trois indicateurs précédents avec la fréquence f (pour la proportion), la moyenne de la série $\bar{x}$ (pour l'espérance) ou l'écart type empirique s (pour σ). Cela vient évidemment de la loi des grands nombres [1].

Plus généralement, si θ est le paramètre à estimer, on appelle T un estimateur, c'est-à-dire une fonction des X_i qui est **convergente** (i.e. $T \underset{n\to+\infty}{\longrightarrow} \theta$ ou encore, pour être plus précis : $p(|T-\theta| > \varepsilon) \underset{n\to+\infty}{\longrightarrow} 0$). Un estimateur T d'un paramètre θ est donc une variable aléatoire dépendant des X_i à valeurs dans l'espace des paramètres.

Nous verrons que l'on peut trouver plusieurs estimateurs pour un même paramètre ; il conviendra donc de déterminer le meilleur via la notion de **précision**. Pour cela, on peut regarder $T-\theta$ qui est ce que l'on appelle **l'erreur d'estimation** :

1. qui dit *grosso modo* que les caractéristiques d'un échantillon aléatoire se rapprochent d'autant plus des caractéristiques statistiques de la population que la taille de l'échantillon augmente.

$$T-\theta = \underbrace{T-E(T)}_{\text{fluctuations aléatoires}} + \underbrace{E(T)-\theta}_{\text{biais}}.$$

La variable aléatoire $T-E(T)$ représente les fluctuations aléatoires de T autour de sa moyenne. Et l'erreur systématique $E(T)-\theta$ est assimilable à l'erreur due au fait que T varie autour de sa valeur centrale $E(T)$ et non autour de θ. C'est ce que l'on appelle le **biais**. L'idéal est d'avoir un estimateur sans biais.

Pour la notion de précision d'un estimateur T, on passe par **l'erreur quadratique moyenne** :

$$\begin{aligned} E[(T-\theta)^2] &= E[(T-E(T)+E(T)-\theta)^2] \\ &= \underbrace{E[(T-E(T))^2]}_{V(T)} + \underbrace{2E[(T-E(T))(E(T)-\theta)]}_{=(E(T)-\theta)E[T-E(T)]=0} + E[(E(T)-\theta)^2] \\ &= V(T)+[E(T)-\theta]^2. \end{aligned}$$

Ainsi, de deux estimateurs sans biais, le plus précis est celui de variance minimale, l'erreur systématique étant incompressible.

Exercice 1.

On considère une population distribuée selon une loi de Poisson de paramètre $\lambda > 0$. On extrait de cette population une observation X_1. Si notre but est d'estimer $e^{-3\lambda}$, que pouvez-vous dire de la statistique $(-2)^{X_1}$?

Nota Bene : Afin d'expliquer au mieux les statistiques inférentielles, on peut se baser sur ce que l'on sait au niveau des variables aléatoires. Le but est de remonter des informations sur la population à partir d'un échantillon.

Considérons un échantillon de taille n. On mesure le caractère : $x_1, \ldots, x_n$. Il s'agit d'une distribution d'échantillonage de moyenne $\bar{x}$ et d'écart type σ. Cette moyenne $\bar{x}$ est en fait une variable aléatoire

- d'espérance m et d'écart type $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ si on considère que la tirage est avec remise (car il s'agit d'une loi binomiale).
- d'espérance m et d'écart type $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{N-n}{N-1}}$ si on considère que la tirage est sans remise (car il s'agit d'une loi hypergéométrique). On retrouve au passage le coefficient d'exhaustivité.

II Estimer la proportion p

On se place, comme dit au début, dans le cadre d'une population très grande (le tirage d'un échantillon est alors sensé ne pas affecter la proportion). A défaut, on peut aussi avoir une situation avec remise, ce qui revient au même. On cherche à estimer la part de la population présentant un caractère donné (savoir si la personne vote pour un candidat particulier par exemple), donc une proportion.

La plus naturel est évidemment de prendre la **fréquence observée** f. On note F la variable aléatoire correspondant à f (nombre de X_i favorables sur n personnes interrogées). On a X_i qui suit une loi de Bernoulli et nF suit une loi binomiale $\mathcal{B}(n,p)$, dont l'espérance est np et l'écart type $\sqrt{np(1-p)}$.

Ainsi, F a pour espérance p et pour écart type $\sqrt{\dfrac{p(1-p)}{n}}$.

Dans le cas très particulier où n est petit, on peut utiliser les tables de la loi binomiale ou la calculatrice (quand elle est autorisée).

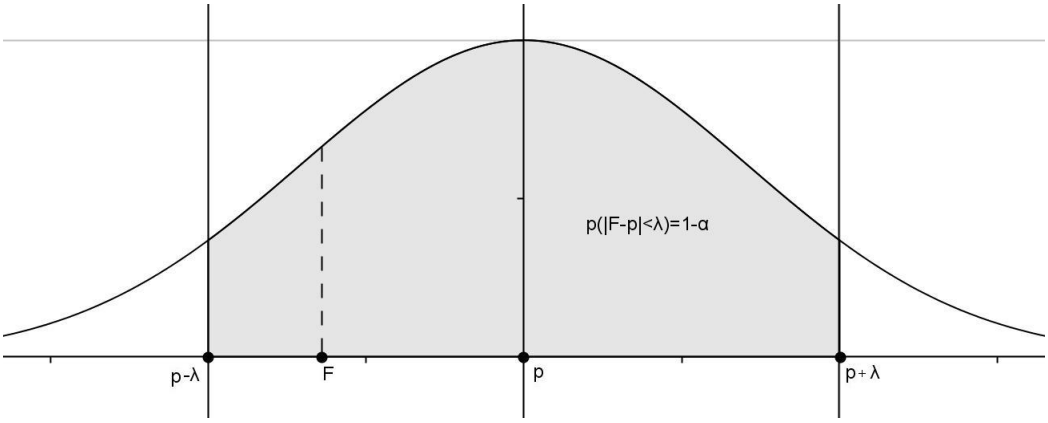
Sinon, on approxime la loi binomiale par une loi normale et on sait que

$$nF \text{ suit } \mathcal{N}(np; np(1-p)) \quad \Leftrightarrow \quad \boxed{F \text{ suit } \mathcal{N}\left(p; \frac{p(1-p)}{n}\right)}.$$

Ainsi, on cherche à donner une "fourchette" dans laquelle se trouve p avec une certaine probabilité d'erreur. C'est ce que l'on appelle un **intervalle de confiance**.

En général, on pose α le risque que l'on prend [2]. On choisit souvent $\alpha = 0,05$ car dans le cas d'un risque à 95%, on trouve une valeur simple pour le $u_{\frac{\alpha}{2}} \simeq 1,96$ de la suite.

La probabilité symétrique $p(|F-p| < \lambda) = 1-\alpha$ est représentée sur le schéma suivant :



Ici, compte tenu de la loi qui n'est pas centrée réduite, on a $\lambda = u_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$. On a donc α chance de se tromper (i.e. 5% dans le cas classique) en affirmant que :

"la fréquence observée est dans l'intervalle $[p-\lambda, p+\lambda]$".

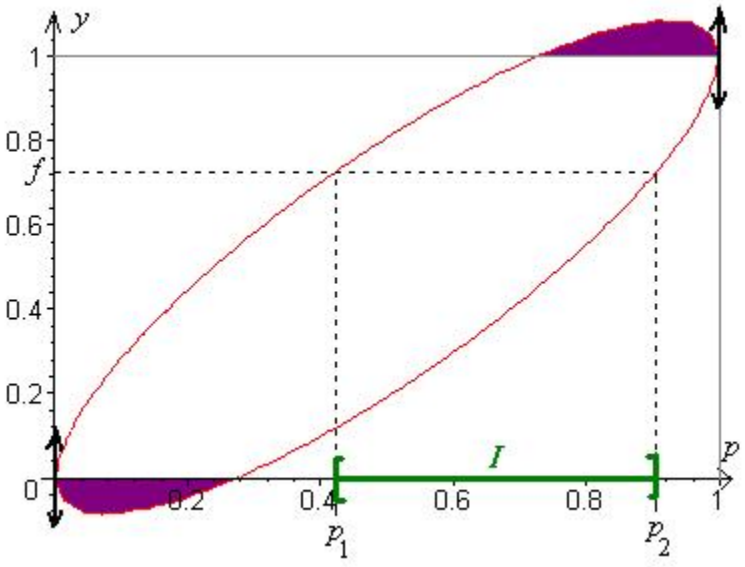
N.B. : Il vaut mieux le voir comme cela (i.e. avec un risque) qu'avec une réussite : "On a 95% de chance que la fréquence observée soit dans l'intervalle $[p-\lambda, p+\lambda]$". Effectivement, la confusion avec les intervalles de fluctuations est facile et impropre et la remarque qui suit confirme cette façon d'écrire. On prendra donc l'habitude de commencer systématiquement sa phrase par : "on a 5% de chance de se tromper en affirmant que..."

Intéressons-nous à une simplification des bornes de cet intervalle, que l'on note y. On a alors :

$$y = p \pm u_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \quad \Leftrightarrow \quad (y-p)^2 = u_{\frac{\alpha}{2}}^2 \frac{p(1-p)}{n} \quad \Leftrightarrow \quad \boxed{y^2 + p^2\left(1 + \frac{u_{\frac{\alpha}{2}}^2}{n}\right) - 2py - \frac{u_{\frac{\alpha}{2}}^2 p}{n} = 0}.$$

Il s'agit de l'équation d'une ellipse en les variables p et y. Sans aller plus loin dans l'étude des ellipses, on obtient le graphe suivant qui passe par les points de coordonnées $(0,0)$ et $(1,1)$ avec des tangentes verticales (faites-moi confiance).

2. On trouve aussi l'appellation "coefficient de sécurité" $1-\alpha$



Les zones en violet correspondent à des valeurs de paramètres pour lesquelles l'approximation n'est pas acceptable. Pour une fréquence f fixée, on trouve **l'intervalle de confiance** I correspondant noté en vert sur le dessin.

En approximant (voir addendum pour le calcul complet), on trouve :

$$\boxed{f - u_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < p < f + u_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}}\,.$$

Remarque: IMPORTANT ! Si l'on part à nouveau du cas où $\alpha = 0,05$, on a :

$$p\left(F - 1,96\sqrt{\frac{F(1-F)}{n}} \leq p \leq F + 1,96\sqrt{\frac{F(1-F)}{n}}\right) = 0,95 \quad (*)$$

On remarque que Si $F \in [0,3;0,7]$ (cas le plus intéressant car sinon, la question est tranchée et le sondage n'a pas d'intérêt), alors $F(1-F) \in [0,21;0;25]$ et $1,96\sqrt{F(1-F)} \sim 1$ (tout en restant plus grand que 1.

donc l'égalité $(*)$ devient :

$$p\left(F - \frac{1}{\sqrt{n}} \leq p \leq F + \frac{1}{\sqrt{n}}\right) \geq 0,95$$

i.e. il y a (au plus) 5% de chance que l'intervalle $\left[F - \frac{1}{\sqrt{n}}; F + \frac{1}{\sqrt{n}}\right]$ ne contienne pas p.

Si f est la fréquence observée, alors $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est la **fourchette de sondage de l'échantillon au seuil de confiance** $0,95$.

Exercice 2. *Un exemple d'application : Le test d'hypothèse*

On fait l'hypothèse qu'une pièce est bien équilibrée et on la jette 860 fois. On trouve une proportion de $0,481$ de face.

1) L'hypothèse est-elle recevable au risque de 5% ?

2) A supposé que l'on continue à obtenir une fréquence de $0,481$, combien aurait-il fallu faire de lancers pour rejeter l'hypothèse ?

III Estimer μ

III.1 Quand σ est connu

Pour estimer l'espérance d'une loi inconnue, le modèle le plus simple est de prendre la **moyenne empirique** de la série statistique :

$$\boxed{T = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i}.$$

Cette moyenne empirique converge naturellement vers μ. Elle est aussi un estimateur sans biais.

De plus, $\bar{X}$ suit une loi normale $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Ainsi, de la même façon que pour la proportion, la probabilité symétrique $p(|\bar{x} - \mu| < \lambda) = 1 - \alpha$ donne l'intervalle de confiance :

$$\boxed{\bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}.$$

III.2 Quand σ n'est pas connu

L'estimateur $\bar{X}$ peut être amélioré dans le cas d'une loi normale dont on ne connaît pas l'écart type. Dans ce cas, on lui préfère

$$\boxed{T = \frac{\bar{X} - \mu}{S}\sqrt{n-1}}.$$

Cet estimateur suit une loi de Student à $(n-1)$ degrés de liberté.

Avec la même idée que pour les intervalles trouvés précédemment pour la loi normale, on trouve l'intervalle de confiance :

$$\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}},$$

avec s l'écart type empirique, et $t_{\frac{\alpha}{2}}$ la valeur donnée par les tables de la loi de Student ; ou encore

$$\boxed{\bar{x} - t_{\frac{\alpha}{2}}\frac{s^*}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}}\frac{s^*}{\sqrt{n}}}.$$

IV Estimer σ^2

On se place dans la même situation d'une population dont on cherche cette fois-ci la variance.

IV.1 Une battle variance empirique - variance corrigée

Mais **la variance empirique** d'échantillon

$$\boxed{S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

est biaisé pour σ^2. En effet, $E(S^2) = \dfrac{n-1}{n}\sigma^2$ car

$$\begin{aligned} E(S^2) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i-\mu)^2 - E(\bar{X}-\mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} V(X_i) - V(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2. \end{aligned}$$

Le biais vaut donc $\dfrac{\sigma^2}{n}$ et tend vers 0. Ainsi, S^2 a tendance à sous-estimer σ^2. On lui préfère alors la **variance corrigée**

$$\boxed{S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

On a alors par la linéarité de l'espérance, $E(S^{*2}) = \sigma^2$. On peut donc dire que $E(S^{2*})$ est un estimateur sans biais de σ^2. En revanche, S^* n'est pas un estimateur sans biais de σ. En effet, $E(\sqrt{S^{*2}}) \neq \sqrt{E(S^{*2})}$ (mais il est asymptotiquement sans biais).

Mais il y a mieux :

IV.2 Si μ est connue

Si μ est connue, l'estimateur

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

est meilleur que S^{*2}. En effet :

$$V(T) = \frac{1}{n^2}V\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}V[(X-\mu)^2] = \frac{1}{n}[\mu_4-\sigma^4].$$

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

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

Les deux derniers tendent vers $\dfrac{\mu_4-\sigma^4}{n} = V(T)$. Mais on peut montrer facilement que $V(T) < V(S^{*2})$. Donc T est un meilleur estimateur de σ^2 que S^{*2} ; encore faut-il connaître μ...

De plus, si on suppose que la loi cherchée est une loi normale, alors $\dfrac{nT}{\sigma^2}$ suit une loi χ^2_n en tant que somme de n carrés de loi $\mathcal{N}(0;1)$ indépendantes.

De la même façon que précédemment, mais avec une loi χ^2_n, on trouve deux valeurs k_1 et k_2 telles que

$$p\left(k_1 < \frac{nT}{\sigma^2} < k_2\right) = 1-\alpha.$$

Et l'intervalle de confiance qui en résulte :

$$\boxed{\frac{nt}{k_2} < \sigma^2 < \frac{nt}{k_1}}.$$

IV.3 Si μ est inconnu

On n'a pas d'autre choix que de prendre la variance empirique $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ et comme $\frac{nS^2}{\sigma^2}$ suit une loi χ^2_{n-1}, on trouve deux valeurs l_1 et l_2 telles que

$$p\left(l_1 < \frac{nS^2}{\sigma^2} < l_2\right) = 1 - \alpha.$$

Et l'intervalle de confiance qui en résulte :

$$\boxed{\frac{ns^2}{l_2} < \sigma^2 < \frac{ns}{l_1}}.$$

Addendum : Calcul de l'intervalle de confiance d'une proportion

En reprenant le calcul plus haut, et en posant $k = u_{\frac{\alpha}{2}}$, on peut trouver l'intervalle de confiance en résolvant l'équation :

$$f^2 + p^2\left(1 + \frac{k^2}{n}\right) - 2pf - \frac{k^2}{n}p = 0.$$

Ce qui donne en p :

$$p^2\left(1 + \frac{k^2}{n}\right) - p\left(\frac{k^2}{n} + 2f\right) + f^2 = 0.$$

On trouve

$$\Delta = \left(\frac{k^2}{n} + 2f\right)^2 - 4\left(1 + \frac{k^2}{n}\right)f^2 = \frac{k^4}{n^2} + 4f\frac{k^2}{n} - 4f^2\frac{k^2}{n}.$$

D'où,

$$p = \frac{\left(\frac{k^2}{n} + 2f\right) \pm \sqrt{\frac{k^4}{n^2} + 4f\frac{k^2}{n} - 4f^2\frac{k^2}{n}}}{2\left(1 + \frac{k^2}{n}\right)} = \frac{\frac{k^2}{n} + 2f}{2\left(1 + \frac{k^2}{n}\right)} \pm \frac{\sqrt{\frac{k^4}{n^2} + 4f\frac{k^2}{n} - 4f^2\frac{k^2}{n}}}{2\left(1 + \frac{k^2}{n}\right)}.$$

- Le premier terme donne par un DL :

$$\frac{\frac{k^2}{n} + 2f}{2\left(1 + \frac{k^2}{n}\right)} = f + o(\tfrac{1}{n}).$$

- Quant au second, on trouve :

$$\frac{\sqrt{\frac{k^4}{n^2} + 4f\frac{k^2}{n} - 4f^2\frac{k^2}{n}}}{2\left(1 + \frac{k^2}{n}\right)} = \sqrt{\frac{k^4 + 4fnk^2 - 4f^2nk^2}{4(n+k^2)^2}} = \sqrt{\frac{k^4 + 4fnk^2 - 4f^2nk^2}{4n^2 + 8k^2n + 4k^4}} \sim \sqrt{\frac{fnk^2 - f^2nk^2}{n^2}} =$$

$$k\sqrt{\frac{f(1-f)}{n}}.$$

Donc en considérant n grand, on a l'approximation trouvée de l'intervalle de confiance.

V Application classique

Exercice 3. *(EDHEC 2013)*

Dans cet exercice, n désigne un entier naturel supérieur ou égal à 2.

On considère une variable aléatoire réelle X dont une densité de probabilité est donnée par :

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{si } x \geq \theta \\ 0 & \text{sinon.} \end{cases}$$

où θ est un réel strictement positif inconnu.

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de la loi X. On pose

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad Z_n = \inf(X_1, X_2, \ldots, X_n).$$

1) a) Déterminer la fonction de répartition F de X puis reconnaître la loi de $X - \theta$. En déduire l'espérance et la variance de X.

b) Montrer que $T_{1,n} = Y_n - 1$ est un estimateur sans biais de θ.

2) a) Déterminer la fonction de répartition de Z_n puis reconnaître la loi de $Z_n - \theta$.

b) Montrer que $T_{2,n} = Z_n - \dfrac{1}{n}$ est un estimateur sans biais de θ.

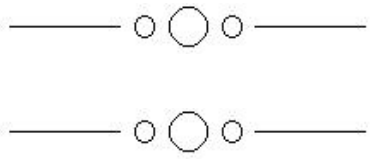
3) Calculer les variances de $T_{1,n}$ et $T_{2,n}$. Que peut-on en conclure ?

4) On veut construire un estimateur T_n sans biais et convergent de θ, comme combinaison linéaire de $T_{1,n}$ et $T_{2,n}$. On souhaite, de plus, que la variance de T_n soit la plus petite possible.

On note ρ le coefficient de corrélation linéaire de $T_{1,n}$ et $T_{2,n}$, supposé constant, et on pose :

$$T_n = a_{1,n}T_{1,n} + a_{2,n}T_{2,n}.$$

a) Donner la valeur de $a_{1,n} + a_{2,n}$.

b) Ecrire la variance de T_n comme une fonction g de $a_{1,n}$, puis déterminer $a_{1,n}$ et $a_{2,n}$ (en fonction de n et de ρ) afin que cette fonction soit minimale.

c) Vérifier que la variance de T_n est inférieure ou égale à celles de $T_{1,n}$ et $T_{2,n}$.