UNIVERSITÉ DE RENNES 1

Introduction à la statistique - 1ère partie

Estimation - Intervalles de confiance

Préparation au Capes - Université Rennes 1

Une entreprise reçoit un lot important de pièces fabriquées en série. L'entreprise n'accepte la livraison que si la proportion de pièces défectueuses est inférieure à 5%. Il est impossible d'inspecter chaque pièce, on extrait alors du lot un échantillon de 200 pièces. Sur ces 200 pièces, on dénombre 15 défectueuses. L'entreprise doit-elle accepter ce lot ?

Ce cours a pour but de répondre à cette question.

Hélène Guérin, helene.guerin@univ-rennes1.fr

Définition Un **échantillon** est une suite de variables $X_1, X_2, \ldots$ où les X_i sont indépendantes et identiquement distribuées (i.i.d.).

La loi commune $\mathbb{P}_\theta$ de ces variables dépend d'un paramètre θ inconnu. Le but va être d'estimer la valeur de θ à partir de l'échantillon.

Concrètement Les valeurs X_i sont le résultat d'observations ou d'un sondage. Les valeurs sont donc connues.

On peut reprendre les deux exemples du cours sur les statistiques descriptives : Le premier concerne l'âge des étudiants préparant le capes de mathématiques à la fac de Rennes en 2005-2006 et l'autre concerne la durée de vie des ampoules d'une grande marque connue.

Dans le premier X_i représente l'âge de la $i^{\text{ème}}$ personne et dans le second X_i représente la durée de vie de la $i^{\text{ème}}$ ampoule.

Exemples de base

1. **Sondage avec remise** : On souhaite évaluer la proportion de mésanges bleues parmi la population de mésanges en Ille et Vilaine. Pour cela, on poste des observateurs à différents endroits dans le département (la mésange n'étant pas un oiseau migrateur). Chaque observateur compte le nombre de mésanges bleues ainsi que le nombre total de mésanges obervées à la jumelle. *On souhaite connaitre la proportion p de mésanges bleues*.
On associe à la variable X_i la valeur 1 si la $i^{\text{ème}}$ mésange observée est bleue et 0 si la mésange n'est pas bleue. On limite notre étude à n oiseaux (par exemple $n = 1000$ mésanges). On suppose que les X_i sont indépendants (en effet, les mésanges vivent en groupes formés de plusieurs espèces de mésanges), et de même loi de Bernoulli $\mathcal{B}(p)$. Par conséquent *le nombre de mésanges bleues $S_n = \sum_{i=1}^{n} X_i$ sur n mésanges observées suit la loi Binomiale $\mathcal{B}(n, p)$*.

Chercher la mésange bleue . . .

2. **Sondage sans remise** : On reprend l'exemple initial. Une entreprise reçoit un lot important de pièces fabriquées en série. On connait le nombre total de pièces, il vaut N. L'entreprise souhaite connaitre le nombre n_1 de pièces défectueuses dans ce lot.

On extrait du lot un échantillon de n pièces choisies au hasard. On associe à la variable X_i la valeur 1 si la $i^{\text{ème}}$ pièce est défectueuses, et la valeur 0 sinon. *Le nombre de pièces défectueuses $S_n = \sum_{i=1}^{n} X_i$ parmi les n pièces étudiées suit la loi Hypergéométrique $H(n_1, n, N)$:*

$$\mathbb{P}(S_n = k) = \frac{\begin{pmatrix} n_1 \\ k \end{pmatrix} \begin{pmatrix} N - n_1 \\ n - k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad \text{pour } k \in \{0, .., \min(n_1, n)\}.$$

Remarque

On sait que lorsque la taille de la population initiale N est très grande, on peut approcher la loi Hypergéométrique par la loi Binomiale $\mathcal{B}(n, p)$, où p est la proportion de pièces défectueuses dans tout le lot ($p = \frac{n_1}{N}$).

La plupart des études réalisées sur une population divisée en deux classes (il n'y a que deux réponses possible pour chaque individu) sont des sondages sans remise, on utilise pourtant la loi binomiale en se basant sur la remarque précédente.

I - Estimation ponctuelle

On considère un échantillon $(X_i)_{1\leq i\leq n}$ de même loi $\mathbb{P}_\theta$ qui dépend d'un paramètre inconnu θ. Les X_i sont connus, le but est d'estimer la valeur de θ à partir des X_i.

On va limiter ce cours au cas où le paramètre inconnu θ est la moyenne $m = E[X_1]$ et où $\sigma^2 = Var(X_1)$ existe.

Exemple Dans les exemples précédents, le paramètre inconnu est

1. l'âge moyen des étudiants de capes,
2. la durée moyenne d'une ampoule électrique,
3. la proportion de mésanges bleues parmi les mésanges en Ille et Vilaine,
4. la proportion de pièces défectueuses.

Définition Soit $m \in \mathbb{R}$ inconnu. Un **estimateur** $\hat{m}$ de m est une variable aléatoire de la forme $\hat{m} = f(X_1, X_2, ..., X_n)$.

Attention ! Un estimateur de m ne doit pas dépendre de m !!

Évidement, il faut choisir la fonction f de façon à "bien" estimer m.

Exemple initial On reprend l'exemple des pièces défectueuses. On note p la proportion de pièces défectueuses.

Les variables $\hat{p} = 10$, $\hat{p} = X_1$, $\hat{p} = \sqrt{X_1 X_2}$ sont des estimateurs de p. Cependant, un estimateur naturel de p est la fréquence de *succès* :

$$\hat{p} = \frac{X_1 + ... + X_n}{n} = \frac{\sharp\text{pièces défectueuses}}{\sharp\text{pièces étudiées}}.$$

L'entreprise ne considère qu'un échantillon de 200 pièces, par conséquent on estime p par $\frac{15}{200} = 0.075$.

De manière générale, comme les variables X_i sont indépendantes et de même loi, d'après la loi faible des grands nombres on a

$$\frac{X_1 + X_2 + ... + X_n}{n} \underset{n \to +\infty}{\longrightarrow} m \quad \text{en proba.}$$

Donc un estimateur naturel de m est $\dfrac{X_1 + X_2 + ... + X_n}{n}$.

Cet estimateur est appelé **moyenne empirique**, il est souvent noté $\overline{X}_n$.

Propriétés

- $E[\overline{X}_n] = m$,
- $Var(\overline{X}_n) = \dfrac{\sigma^2}{n}$ où $\sigma^2 = Var(X_1)$ est la variance de la loi de l'échantillon.

On a estimé l'inconnu m par une certaine quantité $\overline{X}_n$, avec n fixé. Il serait intéressant de connaître la *distance* entre m et son estimateur $\overline{X}_n$, savoir si l'approximation est bonne ou pas.

II - Estimation par intervalle de confiance

Définition

Un **intervalle de confiance** $I = [a, b]$ **au niveau** $1 - \alpha$ est un intervalle aléatoire qui dépend de l'échantillon $X_1, ..., X_n$, mais pas de m, tel que

$$\mathbb{P}_m(m \in I) = \mathbb{P}_m(a \leq m \leq b) = 1 - \alpha \quad \forall m \in \mathbb{R}.$$

Les bornes a, b de l'intervalle I dépendent de l'échantillon :
$a = a(X_1, \ldots, X_n)$, $b = b(X_1, \ldots, X_n)$.

On parle d'**intervalle de confiance asymptotique au niveau** $1 - \alpha$ si I dépend de la taille de l'échantillon n et si

$$\lim_{n \to +\infty} \mathbb{P}_m(m \in I_n) = 1 - \alpha \quad \forall m \in \mathbb{R}.$$

Le nombre α représente le taux d'erreur maximal que l'on accepte de prendre. On prend souvent $\alpha = 5\%$.

Comparaison d'intervalles de confiance :
Le meilleur est celui de longueur la plus petite.

Remarque Dans la suite je distinguerais rarement les termes *intervalle de confiance* et *intervalle de confiance asymptotique*.

Exemple

On considère un échantillon $(X_1, ..., X_n)$ de loi $\mathcal{N}(m, 1)$. On estime m par $\overline{X}_n = \frac{X_1+X_2+...+X_n}{n}$. Comme les variables sont indépendantes, $\overline{X}_n$ suit la loi $\mathcal{N}(m, 1/n)$.

Voici deux exemples d'intervalles de confiance au niveau de confiance 95% :

- $I_1 = [\overline{X}_n - 1.96/n, \overline{X}_n + 1.96/n]$,

- $I_2 = [\overline{X}_n - 1.645/n, +\infty[$.

Le problème est qu'en général, la loi de $\overline{X}_n$ est compliquée ou inconnue.

Lorsque les variables $X_i \sim \mathcal{B}(p)$, on a $X_1 + X_2 + ... + X_n \sim \mathcal{B}(n, p)$. On ne sait pas calculer t tel que $\mathbb{P}(|\overline{X}_n - p| \leq t) = 0,95$ ou tel que $P(\overline{X}_n - p \leq t) = 0,95$.

Utilisation du théorème central limite

1. Si σ est connu.

D'après le théorème de la limite centrale, on a

$$\forall t > 0 \quad \mathbb{P}_m \left(\frac{\sqrt{n}}{\sigma} |\overline{X}_n - m| \leq t \right) \underset{n \to +\infty}{\longrightarrow} \mathbb{P} \left(|Z| \leq t \right)$$

où Z est une v.a. de loi $\mathcal{N}(0,1)$.
On choisit t_α tel que $\mathbb{P} \left(|Z| \leq t_\alpha \right) = 1 - \alpha$. (Par exemple : pour $1 - \alpha = 0.95$, on a $t_\alpha = 1.96$).

Par conséquent

$$I_n = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} t_\alpha, \overline{X}_n + \frac{\sigma}{\sqrt{n}} t_\alpha \right]$$

est un intervalle de confiance asymptotique de m au niveau $1 - \alpha$ lorsque σ est connu.

Problème ! Que faire quand σ est inconnu ?

2. Si σ est inconnu.

• Si σ s'écrit comme une fonction de m : $\sigma = f(m)$, on peut estimer σ par $f(\overline{X}_n)$.

• De manière générale, on peut toujours estimer la variance σ^2 par la **variance empirique** :

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

D'après la loi faible des grands nombres, on a $\hat{\sigma}_n^2 \overset{\text{Proba}}{\underset{n\to+\infty}{\longrightarrow}} \sigma^2$.

Une variante du théorème central limite existe :

$$\forall t > 0 \quad \mathbb{P}\left(\frac{\sqrt{n}}{\hat{\sigma}_n}|\overline{X}_n - m| \leq t\right) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left(|Z| \leq t\right) \quad \text{avec } Z \sim \mathcal{N}(0,1).$$

On choisit t_α tel que $\mathbb{P}(|Z| \leq t_\alpha) = 1 - \alpha$.

Par conséquent

$$I_n = \left[\overline{X}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} t_\alpha, \overline{X}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} t_\alpha\right]$$

est un intervalle de confiance asymptotique de m au niveau $1 - \alpha$ lorsque σ est connu.

Exemple initial On voudrait avoir un intervalle de confiance à 95% de la proportion de pièces défectueuses.

Dans le cas présent l'échantillon suit la loi de Bernoulli $\mathcal{B}(p)$, où p est la proportion de pièces défectueuses dans le lot entier :

$$m = p \quad \text{et} \quad \sigma^2 = p(1-p).$$

On estime alors l'espérance par $15/200 = 0.075$ et la variance par $0.075 \times (1 - 0.075) = 0.069$.

Finalement, la proportion de pièces défectueuses du lot est avec probabilité 0.95 dans l'intervalle

$$I = \left[0.075 - \sqrt{\frac{0.069}{200}}1.96, 0.075 + \sqrt{\frac{0.069}{200}}1.96\right] = [0.038, 0.111]$$

L'entreprise accepte le lot si la proportion de pièces défectueuses est inférieure à 5%. Que faire du lot ?