Inférences Statistiques

Le calcul des probabilités permet, à partir d'un modèle théorique, d'associer à un événement une probabilité qui mesure la fréquence de son apparition dans une suite d'épreuves identiques. La construction du modèle même est une autre investigation qui reste à faire. Les techniques statistiques vont nous aider à reconstruire le modèle ou nous autoriser à nous poser un certain nombre de questions sur le modèle à partir d'un bon nombre d'observations (un échantillon) de réalisations de l'événement étudié. Ces inférences sont fondées sur des critères tels que la vraisemblance des valeurs observées, l'absence de biais ou la convergence des prévisions suggérées par l'échantillon lorsque la taille de celui-ci augmente.

Estimation

La théorie de l'estimation est une problématique importante en statistique théorique. Disposant d'un échantillon $x_1, ..., x_n$ de réalisations d'une v.a. X d'une loi donnée, où un ou plusieurs paramètres sont inconnus, quelle(s) valeur(s) peut-on proposer pour le(s) paramètre(s) inconnu(s) ?

Nous appelons dans la suite la v.a. X *variable aléatoire parente*. Pour l'évaluation d'un paramètre (ou un vecteur paramètre) inconnu θ dont dépend la loi de X, considérons l'application :

$$\widehat{\theta}_n = \widehat{\theta}_n(X_1, ..., X_n),$$

qui est une v.a., prenant une *valeur possible de* θ. L'application $\widehat{\theta}_n$ doit être mesurable.

Elle sera appelée un *estimateur de* θ ; i.e. pour une réalisation $(x_1, ..., x_n)$, la valeur de θ est évaluée par :

$$\widehat{\theta}_n(x_1, ..., x_n).$$

Il faut noter que le nombre d'arguments d'un estimateur n'est pas fixé et varie avec n, qui est la taille de l'échantillon.

Qualités Souhaitées

Jusqu'ici un estimateur est défini de façon arbitraire. Il est évident qu'une telle introduction ne présente aucun intérêt que si on l'enrichit de certaines qualités ou de certaines mesures de performances. Nous énumérons dans la suite quatre propriétés souhaitées. Elles ne sont pas nécessairement compatibles. De plus, la recherche d'un estimateur, possédant une performance souhaitée peut ne pas être une tâche facile.

Maximum de Vraisemblance

Définition. On peut justifier le choix d'un estimateur par le fait qu'il maximise la vraisemblance de la réalisation de l'échantillon $(x_1, ..., x_n)$.

- Si la loi est discrète, la *fonction de vraisemblance* est la probabilité pour que : $(X_1, ..., X_n) = (x_1, ..., x_n)$.
- Si la loi admet une densité f, la *fonction de vraisemblance* est le produit : $f(x_1)...f(x_n)$.

Dans les deux cas la fonction de vraisemblance, notée $L_n(\theta)$ dépend de n, de $(x_1, ..., x_n)$ et de θ.

L'estimateur $\widehat{\theta}_n(x_1, ..., x_n)$ est appelé un *estimateur du maximum de vraisemblance*, s'il maximise la fonction de vraisemblance.

Exemple. On dispose d'un échantillon $(x_1, ..., x_n)$ d'une v.a. X suivant une loi de Poisson de paramètre λ inconnu. Trouver un estimateur du maximum de vraisemblance de λ.

Solution. La fonction de vraisemblance vaut :

$$L_n(\lambda) \;=\; e^{-n\lambda}\, \frac{\lambda^{x_1+\ldots+x_n}}{x_1!\ldots x_n!}.$$

Pour maximiser cette expression par rapport à λ, il suffit de maximiser $e^{-n\lambda}\ \lambda^{x_1+\ldots+x_n}$ ou de façon équivalente son logarithme.

Un calcul simple, passant par la dérivation du logarithme, aboutit à l'estimateur du maximum de vraisemblance :

$$\widehat{\lambda}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

On en déduit que l'estimation du paramètre d'une loi de poisson par la méthode du maximum de vraisemblance nous conduit à identifier le paramètre par la moyenne arithmétique de l'échantillon.

Biais d'un Estimateur

Définitions. Un estimateur $\widehat{\theta}_n$ de θ est dit *sans biais*, si son espérance vaut θ pour tout $n \in \mathbb{N}^*$; sinon on dit qu'il est *biaisé*. Si l'espérance de $\widehat{\theta}_n$ converge vers θ, il est alors appelé *asymptotiquement sans biais*.

L'estimateur $\widehat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ est un estimateur sans biais du paramètre de la loi de Poisson.

Convergence d'un Estimateur

Définition. Un estimateur $\widehat{\theta}_n$ de θ est dit *convergent*, s'il tend en probabilité vers θ, lorsque $n \to \infty$.

Efficacité Absolue d'un Estimateur

Définition. Un estimateur $\widehat{\theta}_n$ de θ est *absolument efficace* si, étant convergent et sans biais, il possède la plus petite variance parmi tous les estimateurs partageant les deux premières propriétés.

Quelques Estimateurs Standard

Estimateur standard de l'espérance. Soit $(x_1, ..., x_n)$ un échantillon d'une v.a. possédant une espérance inconnue μ. Son estimateur standard est la moyenne de l'échantillon :

$$\widehat{\mu}_n(x_1, ..., x_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Cet estimateur est sans biais. Il est de plus convergent, lorsque la v.a. parente admet une variance (la loi des grands nombres).

Dans la suite, pour alléger la notation, nous supprimons la référence de l'estimateur à l'indice n et aux arguments $x_1, ..., x_n$, lorsqu'il n'y a pas le danger de confusion.

Estimateur standard de la variance. Soit X une v.a. admettant une espérance μ et une variance V, toutes deux inconnues. Soit $(x_1, ..., x_n)$ un échantillon de X. L'estimateur standard joint $(\widehat{\mu}, \widehat{V})$ est donné par :

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \widehat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2.$$

Cet estimateur de la variance est convergent et sans biais. On notera qu'il vaut la variance empirique multipliée par le facteur $\frac{n}{n-1}$. On démontre en effet que l'espérance de cette dernière vaut $\frac{n-1}{n}\ V$ et il faut, donc, la multiplier par $\frac{n}{n-1}$ pour la "débiaiser".

Estimateurs standard de la covariance et du coefficient de corrélation. Étant donné un échantillon $((x_1, y_1), ..., (x_n, y_n))$ du couple (X, Y) de v.a. réelles admettant chacune une espérance et une variance inconnues. L'estimateur standard $\widehat{cov}$ de $cov(X, Y)$ est donné par :

$$\widehat{cov} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})(y_i - \widehat{\nu}),$$

où $\widehat{\mu}$ et $\widehat{\nu}$ sont les estimateurs standard de l'espérance de X et de celle de Y respectivement.

De même, l'estimateur standard $\widehat{corr}$ de $corr(X, Y)$ est donné par :

$$\widehat{corr} = \frac{\widehat{cov}}{\widehat{\sigma}.\widehat{\tau}},$$

où $\widehat{\sigma}$ et $\widehat{\tau}$ sont les estimateurs standard de l'écart-type de X et de celui de Y respectivement. Ces deux estimateurs sont sans biais.

Estimateur standard d'une distribution de probabilité finie. Soit $(x_1, ..., x_n)$ un échantillon d'une v.a. X prenant une des valeurs $\alpha_1, ..., \alpha_k$ avec les probabilités inconnues $p_1, ..., p_k$ respectivement. L'estimateur standard du vecteur $p = (p_1, ..., p_k)$ est défini par :

$$\widehat{p}_i = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{\alpha_i\}}(x_j), \quad i = 1, ..., k.$$

Autrement dit, l'estimateur standard de chaque p_i est la fréquence relative d'occurrences de α_i dans l'échantillon. C'est un estimateur sans biais. On peut aussi démontrer que cet estimateur maximise la fonction de vraisemblance.

Dans le cas où l'espérance μ est *connue*, l'estimateur standard de la variance devient :

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

et cet estimateur de la variance est convergent et sans biais.

Estimateur standard de l'écart-type. Il est donnée par :

$$\widehat{\sigma} = \sqrt{\widehat{V}},$$

où $\widehat{V}$ est l'estimateur standard de V, donné par l'une des deux formules de la section précédente.

Intervalle de Confiance (J. Istas)

Étant donné un estimateur $\widehat{\theta}_n$ d'un paramètre inconnu θ, quelle confiance peut-on lui accorder ? Prenons un intervalle centré de rayon δ autour de $\widehat{\theta}$, et cherchons à déterminer si θ appartient ou pas à cet intervalle. Évidemment, sauf dans des cas très particuliers, on ne peut répondre avec certitude à la question "$\theta \in [\widehat{\theta}_n \pm \delta]$?". On peut toutefois confirmer, du moins en principe, qu'avec une certaine probabilité α, la vraie valeur du paramètre se trouve dans cet intervalle. Nous dirons alors que l'intervalle $[\widehat{\theta}_n \pm \delta]$ est un *intervalle de confiance de niveau* α.

Exemple. Soit $(x_1, ..., x_n)$ un échantillon d'une v.a. normale réduite d'espérance μ inconnue (i.e. $\mathcal{N}(\mu, 1)$). Trouver un intervalle de confiance de niveau α pour l'estimateur standard de l'espérance.

Solution. L'estimateur standard de l'espérance étant défini par :

$$\widehat{\mu}_n(x_1, ..., x_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i,$$

on démontre facilement que c'est une v.a. normale d'espérance μ et de variance $\frac{1}{n}$. On en déduit que la v.a. $\sqrt{n}\,(\widehat{\mu}_n - \mu)$ est une v.a. normale centrée-réduite. Nous avons alors, pour tout $\delta > 0$:

$$Pr\left(|\widehat{\mu}_n - \mu| \le \frac{\delta}{\sqrt{n}}\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_{|x|\le\delta} e^{-x^2/2}dx.$$

Pour un niveau α donné, choisissons δ de sorte que :

$$\frac{1}{\sqrt{2\pi}}\int_{|x|\le\delta} e^{-x^2/2}dx \;=\; \alpha.$$

On peut donc affirmer que μ appartient à l'intervalle $[\widehat{\mu} - \delta/\sqrt{n}\ ,\ \widehat{\mu} + \delta/\sqrt{n}]$ avec la probabilité α.

Il s'agit donc d'un intervalle de confiance de niveau α pour l'estimateur.

Intervalle de Confiance Asymptotique

Il arrive souvent que le calcul exact d'un intervalle de confiance soit difficile, alors que le calcul d'un intervalle de confiance approché, lorsque la taille de l'échantillon tend vers l'infini, soit accessible. On parle alors d'*intervalle de confiance asymptotique*.

Exemple. Soit X une v.a. quelconque d'espérance μ inconnue et de variance 1. Soit $(x_1, ..., x_n)$ un échantillon de X. On considère l'estimateur standard $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_n$ de μ. Trouver un intervalle de confiance asymptotique de niveau α pour $\widehat{\mu}_n$.

Solution. Nous ne connaissons pas la loi de la v.a. $\widehat{\mu}_n$. Nous savons en revanche que, lorsque $n \to \infty$, $\sqrt{n}(\widehat{\mu}_n - \mu)$ *converge en loi vers une v.a. normale centrée et de variance 1*. Nous avons donc asymptotiquement :

$$Pr\left(|\widehat{\mu}_n - \mu| \le \frac{\delta}{\sqrt{n}}\right) \;\asymp\; \frac{1}{\sqrt{2\pi}}\int_{|x|\le\delta} e^{-x^2/2}dx.$$

Choisissons alors δ de sorte que :

$$\frac{1}{\sqrt{2\pi}}\int_{|x|\le\delta} e^{-x^2/2}dx \;=\; \alpha.$$

Nous pouvons alors confirmer que la probabilité pour que le paramètre inconnu μ se trouve dans l'intervalle $[\widehat{\mu} - \delta/\sqrt{n}\ ,\ \widehat{\mu} + \delta/\sqrt{n}]$ vaut *asymptotiquement* α. Par conséquent cet intervalle est un intervalle de confiance asymptotique de niveau α pour l'estimateur proposé.