Points importants du cours 2

Vincent Guillemot

14 janvier 2009

1 Estimation ponctuelle

Soit X une variable aléatoire quelconque, on supposera dans tout le paragraphe que sa moyenne et sa variance son finies, c'est le cas pour la plupart des variables que nous avons vues dans le cours précédent. D'autres hypothèses viendront s'ajouter lorsque l'on voudra calculer des intervalles de confiance.

1.1 Remarques préliminaires (rappels rapides)

Dans toute la suite, une variable aléatoire sera notée par une des lettres de la deuxième moitié de l'alphabet latin en majuscule et en italique, le plus souvent N, T, X, Y, Z etc. Une réalisation d'une variable aléatoire X sera notée par la minuscule correspondante x, modulo des indices, exposants ou symboles supplémentaires. Ainsi un échantillon de taille n de la variable aléatoire X sera noté $(x_1, ..., x_n)$, la moyenne empirique de cet échantillon est notée $\overline{x}$ et vaut

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Mais il est possible que nous voulions étudier l'estimateur $\overline{X}$ de la moyenne $E(X)$, on se donnera alors un échantillon aléatoire $(X_1, ..., X_n)$ et on travaillera sur le comportement de la nouvelle variable aléatoire

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Soit θ l'estimateur (c'est une variable aléatoire) d'un paramètre θ^*. On parlera d'estimateur sans biais si $E(\theta) = \theta^*$, si de plus $V(\theta) \to 0$, c'est un estimateur consistant.

Remarque : Cette notation θ utilisée pour désigner un estimateur est courante : pour se familiariser avec son utilisation, nous l'adopterons dans quelques-uns des exemples suivants.

1.2 Proportion

Soit une population d'individus possédant ou non un caractère avec la probabilité p. La variable aléatoire modélisant une telle population est une variable aléatoire de Bernoulli

$X \sim \mathcal{B}(1,p)$ et un échantillon aléatoire de taille n de cette variable est un ensemble de variables $\mathcal{B}(1,p)$ noté $(X_1,...,X_n)$.

Définition 1 - *Indépendance de deux variables aléatoires X et Y*

X et Y sont indépendantes si et seulement $\mathrm{P}(X_1 \in I_1, X_2 \in I_2) = \mathrm{P}(X_1 \in I_1)P(X_2 \in I_2)$, où I_1 et I_2 sont des intervalles de $\mathbb{R}$.

Proposition 1

X et Y indépendantes implique que $E(XY) = E(X)E(Y)$.

Exercice : Quelle est la loi de la variable aléatoire $Y = \sum_{i=1}^{n} X_i$?

Exercice : Calculez l'espérance de la variable aléatoire $\pi = \frac{Y}{n}$. Montrez que

$$V(\pi) = \frac{p(1-p)}{n}.$$

1.3 Moyenne

Le paramètre θ à estimer est l'espérance de la variable X. L'estimateur classique est la moyenne empirique

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

L'espérance de cet estimateur vaut

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

c'est donc un estimateur sans biais ! Notons maintenant $\mu = E(X)$. La variance de l'estimateur vaut

$$E\left((\overline{X}-\mu)^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left((X_i-\mu)(X_j-\mu)\right) = \frac{1}{n}V(X).$$

L'estimateur est donc de plus convergent : $V(\overline{X}) \to 0$.

Remarque : Lorsque l'on suppose que X est une variable aléatoire de paramètres μ et σ^2, $\overline{X}$ est également une variable aléatoire gaussienne (car c'est une combinaison linéaire de variables gaussiennes, résultat important et que nous admettrons) de paramètres μ et σ^2/n.

1.4 Variance

Deux cas de figure se présentent : la moyenne de la variable aléatoire est connue ou non. Cela change assez radicalement l'estimation de la variance.

(Cas 1) *La moyenne μ de X est connue :* On note S_μ^2 l'estimateur de la variance de X

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

On trouve facilement que l'estimateur est sans biais :

$$E(S_\mu^2) = \frac{1}{n}\sum_{i=1}^{n} E\left((X_i-\mu)^2\right) = V(X).$$

Cependant, lorsqu'il faut calculer la variance d'un échantillon, il est fort peu probable de connaître à l'avance la moyenne, on doit donc l'estimer !

(Cas 2) *La moyenne μ de X est inconnue :* elle est estimée par la moyenne empirique. Nous étudierons deux estimateurs différents θ_1 et θ_2, et nous ne garderons que l'estimateur sans biais lorsque l'estimation de la variance interviendra dans les tests d'hypothèses. Ces estimateurs valent

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \text{ et } \theta_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Calculons l'espérance du premier estimateur :

$$E(\theta_1) = \frac{1}{n}\sum_{i=1}^{n} E\left((X_i-\overline{X})^2\right).$$

L' "astuce" qui permet de faire apparaître $V(X)$ consiste à intercaler μ dans le carré. Soit $i \in \{1,...,n\}$,

$$\begin{aligned}(X_i-\overline{X})^2 &= (X_i-\mu+\mu-\overline{X})^2 \\ &= (X_i-\mu)^2+(\mu-\overline{X})^2-2(X_i-\mu)(\overline{X}-\mu),\end{aligned}$$

ainsi, le terme général de la somme vaut

$$\begin{aligned}& E\left((X_i-\mu)^2\right)+E\left((\overline{X}-\mu)^2\right)-2E\left((X_i-\mu)(\overline{X}-\mu)\right) \\ = \;& V(X)+V(\overline{X})-2E\left((X_i-\mu)(\overline{X}-\mu)\right).\end{aligned}$$

Il ne reste qu'un seul terme qui dépend de i, de plus, comme nous l'avons vu au paragraphe précédent, $V(\overline{X}) = V(X)/n$. Nécessairement,

$$\begin{aligned}E(\theta_1) &= V(X)+\frac{1}{n}V(X)-2E\left((\overline{X}-\mu)\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right) \\ &= V(X)+\frac{1}{n}V(X)-\frac{2}{n}V(X) \\ &= \frac{n-1}{n}V(X).\end{aligned}$$

$E(\theta_1)$ ne vaut pas exactement $V(X)$, θ_1 est un estimateur biaisé ! Or, après avoir remarqué que $\theta_2 = \frac{n}{n-1}\theta_1$ et par linéarité de l'espérance,

$$E(\theta_2) = V(X),$$

θ_2 n'est pas biaisé.

Lorsqu'il s'agira d'introduire dans des statistiques de tests un ou des estimateurs de variance, nous choisirons donc l'estimateur sans biais que nous avons noté θ_2 dans ce paragraphe par commodité, il sera noté S^2 dans la suite :

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

L'estimation pour n réalisations $(x_1, ..., x_n)$ de la variable aléatoire de référence X sera notée avec une lettre minuscule :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Dans toute cette partie, nous ne nous sommes intéressés qu'aux moments d'ordre 1 et 2 d'estimateurs donnés. On peut également s'intéresser à la réalisation d'un estimateur, sans se préoccuper de sa loi de probabilité. Quand on ne propose qu'une seule valeur pour la paramètre à estimer, on parle d'estimation ponctuelle.

2 Estimation par intervalle de confiance

On peut vouloir enrichir cette information sur le paramètre à estimer en se servant de la loi que suit l'échantillon considéré dont on observe une réalisation : on peut alors proposer une fourchette de valeurs pour le paramètre. On parle d'intervalle de confiance.

Nous avons jusqu'à présent formulé des hypothèses assez simples sur la nature de notre échantillon. Cependant, pour déterminer un intervalle de confiance, il faut pouvoir formuler des hypothèses quant à la loi de probabilité de la variable aléatoire étudiée. Soit un paramètre θ^* que l'on veut estimer par un estimateur θ. On suppose que cet estimateur a une densité de probabilité quelconque (si possible pour laquelle on dispose de tables) notée $x \mapsto f(x)$ dont le quantile d'ordre q est noté x_q. Déterminer un intervalle de confiance $[a; b]$ pour la valeur calculée $\widehat{\theta}$ au niveau α, se traduit en termes de probabilité de la façon suivante

$$\mathrm{P}\left(\theta^* \in [a; b]\right) = 1 - \alpha.$$

Nous nous intéresserons dans ce chapitre à des intervalles "bilatéraux" :

$$[a; b] = [x_{1-\alpha/2}; x_{\alpha/2}],$$

ou unilatéraux, par exemple

$$]a; b] =]-\infty; x_\alpha] \text{ ou } [a; b[= [x_{1-\alpha}; +\infty[.$$

On parle le plus souvent du risque α plutôt que de la probabilité $1-\alpha$ (c'est équivalent).

Définition 2 - *Intervalle de confiance au risque α du paramètre θ*

On appelle intervalle de confiance au niveau $1-\alpha$, ou au risque α, l'intervalle I tel que $\mathrm{P}(\theta \in I) = 1-\alpha$.

2.1 Proportion

Soit une population d'individus possédant ou non un caractère avec la probabilité p. La variable aléatoire modélisant une telle population est une variable aléatoire de Bernoulli $X \sim \mathcal{B}(1,p)$. Soit un échantillon aléatoire de taille n de cette variable $(X_1, ..., X_n)$. Nous supposons de plus que la population est suffisamment grande pour pouvoir utiliser l'approximation de la variable $Y = X_1 + ... + X_n$ binomiale par une variable de loi gaussienne $\mathcal{N}(np, np(1-p))$. Nous avons considéré au chapitre précédent l'estimateur sans biais $\pi = \frac{Y}{n}$ pour p. Ainsi,

$$\pi \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Pour une estimation $\widehat{\pi} = \frac{\sum_i x_i}{n}$ correspondant à une réalisation de notre échantillon aléatoire, nous avons un intervalle de confiance bilatéral de niveau α pour la valeur du paramètre p :

$$p^* \in \left[\widehat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}\,(1-\widehat{\pi})}{n}}; \widehat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\widehat{\pi}\,(1-\widehat{\pi})}{n}}\right],$$

avec z_q le quantile d'ordre q d'une variable aléatoire gaussienne centrée réduite $Z \sim \mathcal{N}(0,1)$.

2.2 Moyenne

Supposons que l'on dispose d'un n-échantillon d'une variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$. L'estimateur "moyenne empirique" du paramètre μ de la variable X présenté dans les paragraphes précédents est sans biais :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Nous avons déjà calculé sa moyenne et sa variance, et, par théorème (non vu), $\overline{X}$ est une variable aléatoire gaussienne : $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. L'intervalle de confiance de μ au risque α se calcule sous deux configurations différentes :

- la variance de X, σ^2 est connue,
- elle est inconnue.

Dans cette partie, $\theta^* = \mu$ est la valeur à estimer, et l'estimateur que nous utilisons est $\theta = \overline{X}$.

σ^2 connue La nouvelle variable

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

L'intervalle de confiance bilatéral au risque α pour le paramètre μ est donc

$$\mu \in \left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

l'intervalle est également noté

$$\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

σ^2 inconnue La variance de X est estimée par S^2. Ainsi, en procédant de la même manière que lorsque σ^2 est connue, la nouvelle variable

$$\frac{\overline{X} - \mu}{S/\sqrt{n-1}} \sim \mathcal{T}(\nu = n-1),$$

un moyen "mnémotechnique" (qui ne constitue en aucun cas une preuve) permettant de se rappeler le nombre de degrés de libertés consiste à considérer qu'un degré de liberté est "mangé" par l'estimation de la moyenne dans la somme du dénominateur.

L'intervalle de confiance bilatéral de niveau α pour μ est donc :

$$\mu \in \left[\overline{X} - t_{1-\alpha/2}^{\nu=n-1}\frac{S}{\sqrt{n-1}}; \overline{X} + t_{1-\alpha/2}^{\nu=n-1}\frac{S}{\sqrt{n-1}}\right],$$

l'intervalle est également noté

$$\overline{X} \pm t_{1-\alpha/2}^{\nu=n-1}\frac{S}{\sqrt{n-1}}.$$

2.3 Variance

On suppose que la variable étudiée est une variable gaussienne de moyenne μ et de variance σ^2. Le paramètre étudié est ici la variance de X, σ^2. Deux cas de figure se présentent :

- μ est connue,
- μ est inconnue.

μ connue Nous allons utiliser l'estimateur sans biais :

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

Grâce à l'hypothèse de normalité de la variable X étudiée,

$$\frac{S_\mu^2}{\sigma^2} \sim \chi^2(\nu = n),$$

l'intervalle de confiance au niveau α est donc (attention, la loi du χ^2 n'est pas symétrique !)

$$\sigma^2 \in \left[\frac{nS_\mu^2}{\chi^2_{1-\alpha/2,\nu=n}}; \frac{nS_\mu^2}{\chi^2_{\alpha/2,\nu=n}} \right].$$

μ inconnue L'estimateur sans biais sera utilisé :

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

L'intervalle de confiance au niveau α est très similaire au précédent, attention donc à ne pas les confondre, pour de petits échantillons, cela peut faire une grosse différence :

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\nu=n-1}}; \frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu=n-1}} \right].$$

3 Tests d'hypothèses

3.1 Exemple préliminaire

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. μ est inconnu, et σ est connu. On cherche à décider si μ est égal à une valeur μ_0 fixe. On appelle $\mathcal{H}_0$ l'hypothèse nulle $\mu = \mu_0$. On appelle $\mathcal{H}_1$ l'hypothèse alternative $\mu \neq \mu_0$ (il existe d'autres possibilités).

Le résultat d'un test est une prise décision :

- soit on la rejette,
- soit on ne peut pas la rejeter (et dans ce cas on l'accepte).

Cette décision est prise sur la base d'observations $(x_1, ..., x_n)$ qui forment un échantillon de taille n. Dans la suite nous considérerons un échantillon aléatoire et l'estimateur classique de μ : $\overline{X} = \frac{1}{n} \sum_i X_i$.

$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, nous aimerions savoir si $\overline{X}$ est significativement différent de μ_0 !

Supposons que $\mathcal{H}_0$ est vraie (on dit également "sous $\mathcal{H}_0$"). Alors $\overline{X} \sim \mathcal{N}(\mu_0, \frac{\sigma^2}{n})$. Quand peut-on considérer que $\overline{X}$ est trop éloigné de μ_0 pour lui être égal ? On considérera une variable centrée $Z \sim \mathcal{N}(0,1)$ (pour laquelle on dispose de table). Sous $\mathcal{H}_0$, $P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2})$, on décide $\mathcal{H}0$ si $\overline{X}$ appartient à l'intervalle $\mu_0 \pm z\frac{\sigma}{\sqrt{n}}$, on décide $\mathcal{H}_1$ sinon.

Définition 3 - *Risque de première et deuxième espèce*

α = probabilité de rejeter $\mathcal{H}_0$ alors qu'elle est vraie.
β = probabilité d'accepter $\mathcal{H}_0$ alors qu'elle est fausse.

Application numérique : $\mu_0 = 15$, $\sigma^2/n = 2.5^2/100$, $z_{1-\alpha/2} = 1.96$, $\alpha = 0.05$ et $\overline{x} = 14.2$.

Calcul de β : Supposons qu'en fait $\mu = \mu_1 = 14.4$. Sous $\mathcal{H}_1$, $Z = \sqrt{n}\frac{\overline{X}-\mu_1}{\sigma}$.

$$\begin{aligned}
\beta &= \mathrm{P}\left(x_1 \leq \overline{X} \leq x_2 | \overline{X} \sim \mathcal{N}(\mu_1, \sigma^2/n)\right) \\
&= \mathrm{P}\left(x_1 \leq \frac{\sqrt{n}}{\sigma} Z + \mu_1 \leq x_2\right) \\
&= \mathrm{P}\left(\frac{\sqrt{n}}{\sigma}(x_1 - \mu_1) \leq Z \leq \frac{\sqrt{n}}{\sigma}(x_2 - \mu_1)\right) \\
&= \ldots
\end{aligned}$$