Licence de Biologie, 3e semestre
S. Vinatier Compléments de Mathématiques

7. Estimation des paramètres

Dans une population trop grande pour être étudiée entièrement, on constitue un échantillon de n individus, sur lesquels on mesure la valeur d'un caractère quantitatif x. Soit X la variable aléatoire (théorique) définie par les valeurs du caractère x sur l'**ensemble** de la population ; on dispose des *estimations ponctuelles* suivantes pour la moyenne, la variance et l'écart-type de X :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \quad Var_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \ , \quad s_x = \sqrt{Var_x} \ .$$

Noter que Var_x est égal à la *variance observée* multipliée par $\frac{n}{n-1}$, de façon à obtenir un estimateur *sans biais*, c'est-à-dire dont l'espérance (lorsqu'on considère un grand nombre d'échantillons) est égale à la variance théorique. La plupart des calculatrices à fonctions statistiques donnent à la fois l'écart-type *observé* (noté σ_n) et l'écart-type *débiaisé* s_x (noté σ_{n-1}).

Le principe de l'*estimation par intervalles* d'un paramètre t (moyenne, variance,...) est de choisir une *probabilité d'erreur* α et de déterminer un *intervalle de confiance* $]a, b[$ tel que

$$P(a < t < b) = 1 - \alpha \ .$$

Autrement dit, $P(t \notin \,]a, b[) = \alpha$. On appelle aussi α le *risque d'erreur* ou le *cœfficient de risque* et $1 - \alpha$ le *seuil de confiance* ou le *cœfficient de sécurité*. On choisit en général $\alpha = 1\%$ ou $\alpha = 5\%$.

En pratique, pour estimer la moyenne et la variance par intervalles de confiance, on distingue deux cas :

- pour les **grands** échantillons ($n \geq 30$), on lit u_α tel que $P(|\mathcal{N}| \geq u_\alpha) = \alpha$ ($\mathcal{N}$ est la loi normale centrée réduite) dans la *table de l'écart réduit* ; l'intervalle de confiance de la moyenne du caractère x au risque α est donné par :

  $$\left]\overline{x} - u_\alpha \frac{s_x}{\sqrt{n}} \ ; \ \overline{x} + u_\alpha \frac{s_x}{\sqrt{n}}\right[\ .$$

  En supposant de plus que X suit une **loi normale**, l'intervalle de confiance de la variance du caractère x au risque α est donné par :

  $$\left]\frac{2(n-1)s_x^2}{(\sqrt{2n-3} + u_\alpha)^2} \ ; \ \frac{2(n-1)s_x^2}{(\sqrt{2n-3} - u_\alpha)^2}\right[\ .$$

- Pour les **petits** échantillons ($n \leq 30$), on a besoin de l'hypothèse que la variable aléatoire X suit une **loi normale**. On lit alors t_α tel que $P(|T| \geq t_\alpha) = \alpha$ (T est la loi de Student à $n - 1$ degrés de liberté) dans la *table de la loi de Student* ; l'intervalle de confiance de la moyenne du caractère x au risque α est donné par :

  $$\left]\overline{x} - t_\alpha \frac{s_x}{\sqrt{n}} \ ; \ \overline{x} + t_\alpha \frac{s_x}{\sqrt{n}}\right[\ .$$

Pour la variance, on lit a et b tels que $P(Y^2 \geq b) = \frac{\alpha}{2}$ et $P(Y^2 \geq a) = 1 - \frac{\alpha}{2}$, où Y^2 est la loi du χ^2 à $n-1$ degrés de liberté, dans la *table de la loi du chi-deux*; l'intervalle de confiance de la variance du caractère x au risque α est donné par :

$$\left]\frac{(n-1)s_x^2}{b}\ ;\ \frac{(n-1)s_x^2}{a}\right[\ .$$

On rappelle qu'il est aussi possible d'estimer la *fréquence* d'apparition d'un caractère par intervalles de confiance.

Exercice 1

On a mesuré le poids en grammes du cerveau d'hommes et de femmes de 20 à 49 ans issus d'une région d'Europe. On a obtenu les résultats suivants.

Centre des classes	1170	1220	1270	1320	1370	1420	1470	total
Effectifs (hommes)	5	36	45	50	61	49	19	265

Centre des classes	1070	1120	1170	1220	1270	1320	1370	total
Effectifs (femmes)	12	22	45	54	52	20	10	215

Donner un intervalle de confiance au risque de 1% :

(a) pour la moyenne de la population des hommes;

(b) pour la moyenne de la population des femmes.

Exercice 2

On a mesuré la masse en kg de raisin par souche sur 10 souches prises au hasard dans une vigne et on a obtenu :

$$2{,}4\ ;\ 3{,}2\ ;\ 3{,}6\ ;\ 4{,}1\ ;\ 4{,}3\ ;\ 4{,}7\ ;\ 5{,}4\ ;\ 5{,}9\ ;\ 6{,}5\ ;\ 6{,}9\ .$$

(a) Déterminer une estimation ponctuelle non biaisée de la moyenne et de la variance de la vigne dont ces souches sont extraites.

(b) Donner un intervalle de confiance de la moyenne au risque 5%, en supposant que la masse de raisin par souche suit une loi normale dans cette vigne.

(c) Sous la même hypothèse, donner un intervalle de confiance de la variance au risque 5%. En déduire un intervalle de confiance au même risque pour l'écart-type.

Exercice 3

Dans une parcelle de soja, on a mesuré la hauteur en cm de 100 plantes à l'âge de 6 semaines :

Hauteurs	36	37	38	39	40	41
Effectifs	6	11	26	32	14	11

Déterminer sous les hypothèses adéquates des intervalles de confiance au cœfficient de sécurité 0,95 pour la moyenne et la variance de la hauteur des plantes de la parcelle.