ECOLE DES MINES DE NANCY

Décision et Prévision Statistiques

Thierry Verdel (2007)

Chapitre 4

L'estimation statistique

Le problème traité dans ce chapitre est le suivant : on se trouve en présence d'un échantillon et l'on cherche à déterminer explicitement la loi de probabilité définissant la population de référence dont ces observations peuvent être considérées comme issues. Nous admettrons spécifiée la forme analytique de la loi de probabilité que suivent les observations. Dans ces conditions, on se trouve conduit à estimer les paramètres θ_1, θ_2, ... de la loi de probabilité $p(x\,;\,\theta_1,\,\theta_2,\,...)$ à partir de l'échantillon observé x_1, x_2, ... , x_n, c'est à dire à tirer de cet échantillon une information concernant la valeur des paramètres inconnus. Il s'agit de plus de pouvoir émettre un jugement sur la qualité de cette information

---

1. Estimateur et intervalle de confiance

1.1. La loi des grands nombres

Considérons une suite de variables aléatoires X_1, ..., X_i, ..., X_n indépendantes, et ayant toutes la même loi de probabilité qu'une variable aléatoire X. La loi de probabilité de X peut être *quelconque* de moyenne $E(X) = \mu$ et de variance σ^2.

Soit $M_n = \frac{1}{n}(X_1 + ... + X_i + ... + X_n)$ la moyenne arithmétique des variables X_1, ... , X_i, ... , X_n.

Nous avons calculé au chapitre précédent sa moyenne $E(M_n) = \mu$ et sa variance $\sigma^2(M_n) = \frac{\sigma^2}{n}$.

Soit ε un nombre choisi arbitrairement aussi petit que l'on veut. En utilisant l'inégalité de Bienaymé-Tchebichef, on peut écrire que :

$$\text{Prob}\{\,|M_n - \mu| > \varepsilon\} < \frac{\sigma^2}{n\varepsilon^2}$$

et, en posant $\frac{\sigma^2}{n\varepsilon^2} = \eta$, on voit qu'étant donnés ε et η aussi petits qu'on le veut, il est possible de trouver un nombre $N = \frac{\sigma^2}{\eta\varepsilon^2}$ tel que $n \geq N$ entraine :

$$\text{Prob}\{\,|M_n - \mu| > \varepsilon\} < \eta.$$

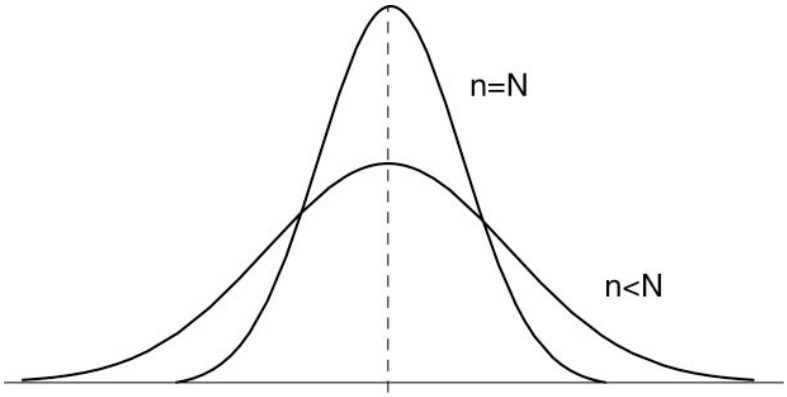
C'est la loi des grands nombres que l'on peut encore énoncer ainsi : quand n augmente, la moyenne M_n *converge en probabilité* vers l'espérance mathématique de X. Il est important de bien prendre conscience de la différence entre la notion de convergence classique et celle, toute nouvelle, de convergence en probabilité.

Dire que M_n tend au sens classique vers μ, ce serait dire qu'on peut déterminer n tel qu'étant donné ε aussi petit qu'on veut : $|M_n - \mu| \leq \varepsilon$.

Dire qu'il y a convergence en probabilité, c'est dire que l'évenement $\{\,|M_n - \mu| \leq \varepsilon\}$ n'est pas certain, mais que sa probabilité peut être rendue aussi voisine de 1 qu'on le veut, à condition que n soit suffisamment grand :

$$\text{Prob}\{\,|M_n - \mu| \leq \varepsilon\} \geq 1 - \eta.$$

On peut l'illustrer par la figure (4.1) : quand n augmente, la distibution de M_n se resserre autour de μ .



(4.1)

1.2. Estimation et estimateur

Pour simplifier la suite de l'exposé, nous supposerons que la loi de référence à déterminer dépend d'un seul paramètre θ. Le problème se réduit donc à déterminer une fonction des observations : $\theta^*(x_1, x_2, ..., x_n)$, aussi voisine que possible de la valeur vraie θ qui est inconnue. On dit alors que θ^* est une *estimation* de θ.

On peut utiliser, pour résoudre ce problème, la notion d'estimateur. Etant donné une variable aléatoire $T_n(X_1, X_2, ..., X_n)$ fonction des variables aléatoires $X_1, X_2, ..., X_n$, on dit qu'elle constitue un *estimateur* de θ si :

- son espérance mathématique tend vers θ quand n augmente indéfiniment : $E(T_n) \underset{n\to\infty}{\longrightarrow} \theta$,

- sa variance tend vers 0 quand n augmente indéfiniment : $E[T_n - E(T_n)]^2 \underset{n\to\infty}{\longrightarrow} 0$.

Dans le cas particulier où $E(T_n) = \theta$ quel que soit n, l'estimateur T_n est dit *sans biais*.

1.3. Estimateur et convergence en probabilité

Pour éclairer la compréhension de la définition d'un estimateur, on peut la rapprocher de celle de la convergence en probabilité.

Un estimateur T_n est dit convergent, si T_n converge en probabilité vers θ, quand n augmente indéfiniment, c'est-à-dire si, étant donnés deux nombres ε et η aussi petits qu'on le veut, il est possible de déterminer un nombre N tel que $n > N$ entraîne :

$$\text{Prob}\{\,|\,T_n - \theta\,|\, > \varepsilon\} < \eta.$$

On a alors l'important résultat suivant : *un estimateur* $\boldsymbol{T_n}$ *dont l'espérance mathématique tend vers* $\boldsymbol{\theta}$ *et dont la variance tend vers* $\boldsymbol{0}$*, quand n augmente indéfiniment, est convergent***.** La démonstration qui suit peut être omise.

D'après l'inégalité de Bienaymé-Tchebichef, on a la relation :

$$\text{Prob}\{\,|\,T_n - E(T_n)\,|\, < \tfrac{\varepsilon}{2}\} > 1 - 4\,\frac{\sigma^2(T_n)}{\varepsilon^2}$$

et a fortiori :

$$\text{Prob}\{\,|\,T_n - \theta\,|\, < \tfrac{\varepsilon}{2} + |\,\theta - E(T_n)\,|\} > 1 - 4\,\frac{\sigma^2(T_n)}{\varepsilon^2}.$$

Puisque $E(T_n)$ *converge vers* θ *au sens ordinaire de l'analyse, on peut trouver un nombre* N_1 *tel que* :

$$| E(T_n) - \theta | < \frac{\varepsilon}{2} \text{ pour } n > N_1,$$

soit, pour une telle valeur de n :

$$\text{Prob}\{ | T_n - \theta | < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \} > 1 - 4 \frac{\sigma^2(T_n)}{\varepsilon^2}.$$

Or, puisque $\sigma^2(T_n)$ *tend vers* 0 *quand n augmente indéfiniment, on peut trouver un nombre* N_2 *tel que* :

$$\sigma^2(T_n) < \frac{\eta \varepsilon^2}{4} \text{ pour } n > N_2.$$

Par suite, pour n supérieur au plus grand des deux nombres N_1 *et* N_2, *on aura* :

$$\text{Prob}\{ | T_n - \theta | > \varepsilon \} < \eta.$$

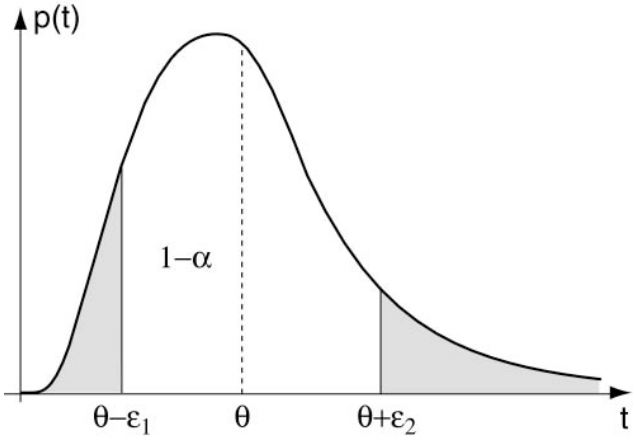
Ce critère permet de définir l'*efficacité* d'un estimateur : *un estimateur est d'autant plus efficace que sa variance est plus petite*.

1.4. Intervalle de confiance d'une estimation

Il reste maintenant à définir la précision de l'estimation. Considérons, pour cela, la distribution de la variable aléatoire T_n. Si nous convenons de considérer comme négligeable un certain seuil de probabilité α, nous pouvons déterminer un intervalle :

$$[\theta - \varepsilon_1, \ \theta + \varepsilon_2]$$

tel qu'il lui corresponde la probabilité $(1 - \alpha)$.



(4.2)

Il résulte de la définition même de cet intervalle que l'on a la probabilité $(1 - \alpha)$ d'observer l'évènement $\{\theta - \varepsilon_1 \leq T_n \leq \theta + \varepsilon_2\}$.

Cela étant, chaque fois que cette double inégalité est vérifiée, c'est-à-dire dans la proportion $(1 - \alpha)$ des cas, la double inégalité :

$$T_n - \varepsilon_2 \leq \theta \leq T_n + \varepsilon_1$$

est, elle aussi, vérifiée. L'intervalle :

$$[\, T_n - \varepsilon_2 \,, \ T_n + \varepsilon_1 \,]$$

est ainsi un intervalle aléatoire auquel peut être associée la probabilité $(1 - \alpha)$ de recouvrir la vraie valeur inconnue de θ :

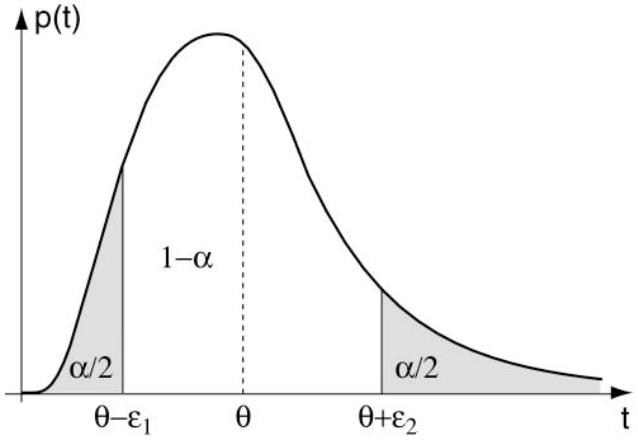
$$\text{Prob}\{T_n - \varepsilon_2 \leq \theta \leq T_n + \varepsilon_1\} = 1 - \alpha.$$

Si maintenant, nous observons les résultats de l'échantillon effectivement prélevé et calculons la valeur θ^* de T_n pour cet échantillon, l'intervalle :

$$[\theta^* - \varepsilon_2,\ \theta^* + \varepsilon_1]$$

est appelé *intervalle de confiance* de l'estimation de θ, au seuil de probabilité $(1 - \alpha)$.

Remarquons qu'il y a une infinité de façons de répartir la probabilité α, dont l'une correspond à un intervalle minimal, mais qui n'est pas toujours facile à déterminer en pratique. C'est pourquoi on convient généralement de répartir α par moitié de part et d'autre de l'intervalle. Cette répartition donne lieu à l'intervalle minimum dans le cas particulier où la densité de probabilité est symétrique et décroît pour des valeurs qui s'éloignent de θ.



(4.3)

1.5. Estimation d'une proportion

Considérons une population qui contient une proportion ϖ inconnue de pièces défectueuses. La variable aléatoire K_n, nombre de pièces défectueuses dans un échantillon de taille n, suit une loi binomiale de moyenne $n\,\varpi$ et de variance $n\,\varpi\,(1 - \varpi)$. Si nous considérons maintenant la variable fréquence $\frac{K_n}{n}$, elle a pour moyenne ϖ et pour variance $\frac{\varpi\,(1-\varpi)}{n}$. $\frac{K_n}{n}$ a donc les propriétés d'un estimateur sans biais de ϖ $\left(E(\frac{K_n}{n}) = \varpi\right)$ et convergent $\left(\sigma^2(\frac{K_n}{n}) \underset{n\to\infty}{\longrightarrow} 0\right)$.

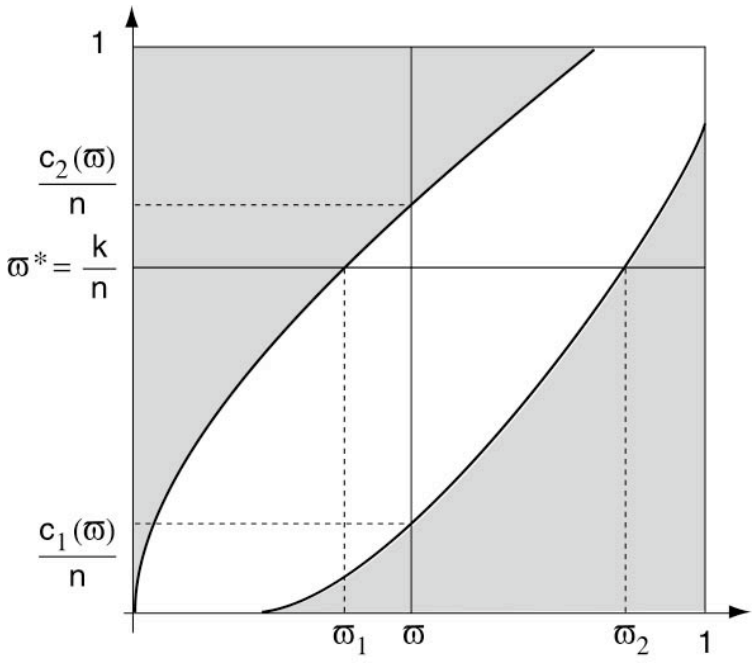
Pour estimer ϖ, il suffit donc, ayant prélevé un échantillon de taille n dans la population, de calculer le nombre k_n de pièces défectueuses puis :

$$\varpi^* = \frac{k_n}{n}.$$

Pour déterminer l'intervalle de confiance, au risque α, de cette estimation, soient $c_1(\varpi)$ et $c_2(\varpi)$ les nombres tels que pour chaque valeur possible de ϖ :

$$\frac{\alpha}{2} \leq \sum_{k=0}^{c_1} C_n^k\, \varpi^k (1-\varpi)^{n-k} \quad \text{et} \quad \frac{\alpha}{2} \leq \sum_{k=c_2}^{\infty} C_n^k\, \varpi^k (1-\varpi)^{n-k}$$

Il est alors possible de construire le graphe ci-dessous en portant ϖ en abcisse et $\frac{c_1(\varpi)}{n}$ ou $\frac{c_2(\varpi)}{n}$ en ordonnée. La surface comprise entre les deux courbes est ainsi le lieu des valeurs possibles de $\frac{k}{n}$ pour l'ensemble des valeurs possibles de ϖ, lorsqu'on néglige le seuil de probabilité α.



(4.4)

Lisant maintenant le graphique suivant l'horizontale d'ordonnée $\varpi^* = \frac{k}{n}$, on peut immédiatement obtenir l'intervalle de confiance $[\varpi_1, \varpi_2]$ correspondant à cette estimation.

1.6. Estimation d'une moyenne

Etant donnée une population de moyenne μ inconnue et de variance σ^2 connue, soit M_n la variable aléatoire moyenne d'un échantillon de taille n. On a montré que $E(M_n) = \mu$ et $\sigma^2(M_n) = \frac{\sigma^2}{n}$. M_n constitue donc un estimateur sans biais et convergent de μ. Par conséquent, ayant prélevé un échantillon, sa moyenne est une estimation de μ :

$$m = \mu^*.$$

Si ce résultat est absolument général et qu'il est, en particulier, indépendant de la forme analytique de la loi de probabilité suivie par les observations, la détermination d'un intervalle de confiance nécessite la connaissance de cette forme. Admettons qu'il s'agisse d'une *loi normale*, de variance σ^2 *connue*.

Il en résulte que M_n suit aussi une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Etant donné un seuil de probabilité α, on peut écrire :

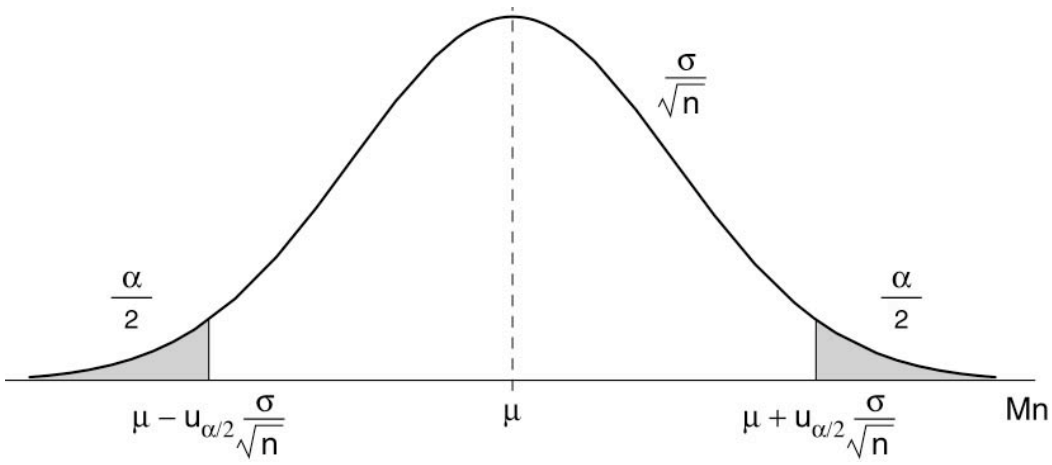
$$\text{Prob}\left\{\mu - u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < M_n < \mu + u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

où $u_{\alpha/2}$ est lu dans la table de la loi normale réduite de telle façon que :

$$\text{Prob}\{\,|\,U\,|\, > u_{\alpha/2}\} = \alpha\,.$$

L'intervalle de confiance de μ est donc :

$$\mu^* - u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \mu^* + u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$$



(4.5)

Bien souvent la variance σ^2 n'est pas connue. Nous verrons, dans la suite du chapitre, comment procéder dans ce cas.

Notons que le théorème *central limite* permet de généraliser ces résultats à une loi de probabilité *quelconque*, mais à condition que *n* soit suffisamment grand (quelques dizaines en pratique).

1.7. Estimation d'une variance

Soit une population quelconque de variance inconnue. Soit $S_n{}^2$ la variable aléatoire : variance d'un échantillon de taille *n*, qui s'écrit :

$$S_n{}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - M_n)^2$$

où M_n désigne la variable aléatoire moyenne. On peut encore écrire :

$$S_n{}^2 = \frac{1}{n} \sum_{i=1}^{n} [(X_i - \mu) - (M_n - \mu)]^2,$$

et, en développant le carré (moyenne des carrés moins carré de la moyenne) :

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

Calculons, sous cette dernière forme, $E(S_n{}^2)$. Il vient, compte tenu de la linéarité de l'opérateur Espérance :

$$E(S_n{}^2) = \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \mu)^2] - E[(M_n - \mu)^2],$$

et, en notant que $E[(X_i - \mu)^2]$ et $E[(M_n - \mu)^2]$ sont respectivement les variances de X_i et de M_n :

$$E(S_n{}^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2.$$

On pourrait montrer, d'autre part, mais les calculs sont assez laborieux, que :

$$\sigma^2(S_n{}^2) = \left(\frac{n-1}{n}\right)^2 \times \frac{\mu_4 - \sigma^4}{n} + 2 \frac{n-1}{n^3} \sigma^4,$$

où μ_4 désigne le moment centré d'ordre 4, $E[(\mathrm{Xi} - \mu)^4]$.

Il en résulte que $S_n{}^2$ est un estimateur convergent de σ^2, mais qu'il n'est pas sans biais. Le passage à un estimateur sans biais ne présente toutefois aucune difficulté. Il suffit de considérer la quantité : $\frac{n}{n-1} S_n{}^2$ dont l'espérance mathématique est :

$$E\left(\frac{n}{n-1} S_n{}^2\right) = \frac{n}{n-1} E(S_n{}^2) = \sigma^2.$$

Nous noterons, par la suite σ^{*2} cette estimation sans biais de σ^2 :

$$\sigma^{*2} = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n-1}$$

qui est calculée directement par presque toutes les calculettes effectuant des calculs statistiques.

Le calcul de l'intervalle de confiance, dans le cas d'une population normale, nécessite la connaissance préalable d'une nouvelle loi d'échantillonnage que nous allons établir maintenant.

2. Intervalle de confiance de la variance inconnue d'une population normale

2.1. Loi du χ^2

Soient U_1, U_2, ..., U_ν, ν variables aléatoires *indépendantes* qui suivent des lois *normales réduites*. Posons :

$$\chi_\nu^2 = U_1^2 + U_2^2 + ... + U_\nu^2.$$

La variable χ_ν^2 suit une loi appelée *loi du χ^2 (khi deux) à ν degrés de liberté*. Nous allons établir la forme analytique de cette loi qui joue un rôle essentiel en statistique, et dont il faut retenir la définition, mais le calcul qui suit n'est, quant à lui, pas essentiel.

Nous nous proposons de déterminer la loi de probabilité de la variable χ_ν^2 ; mais nous allons d'abord déterminer la loi de probabilité de la variable χ_ν, c'est-à-dire que nous allons calculer la probabilité pour que χ_ν se trouve compris dans l'intervalle $[c, c + dc]$.

La probabilité pour qu'on ait à la fois :

$$u_1 \leq U_1 < u_1 + du_1, \quad u_2 \leq U_2 < u_2 + du_2, \quad ... \text{ et } u_\nu \leq U_\nu < u_\nu + du_\nu$$

s'écrit, U_1, U_2, ..., U_ν étant indépendantes :

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{u_1^2}{2}} du_1 \times ... \times \frac{1}{\sqrt{2\pi}} e^{-\frac{u_\nu^2}{2}} du_\nu = \frac{1}{(2\pi)^{\nu/2}} e^{-\frac{1}{2}(u_1^2 + ... + u_\nu^2)} du_1 \, du_2 \, ... du_\nu$$

Ce résultat peut s'interpréter géométriquement de manière très simple. Considérons, dans l'espace à ν dimensions, le point P de coordonnées $(U_1, U_2, ..., U_\nu)$. *La probabilité pour que P tombe à l'intérieur d'un certain volume* dv *défini autour du point P est* :

$$\frac{1}{(2\pi)^{\nu/2}} e^{-\frac{1}{2}\overline{OP}^2} dv.$$

Dans ces conditions, la probabilité pour que la variable χ_ν soit comprise entre les deux valeurs c et $c + dc$ *est égale à la probabilité pour que P tombe dans la région comprise entre les deux sphères de centre O et de rayons c* et $c + dc$. *Or la densité de probabilité est constante dans cette région et elle est égale à* :

$$\frac{1}{(2\pi)^{\nu/2}} e^{-\frac{c^2}{2}}.$$

D'autre part, le volume de la sphère de centre O et de rayon c est de la forme $k\,c^{\nu}$ *où k est une certaine constante, si bien que le volume compris entre les deux sphères de rayons c et* $c + dc$ *est égal à :* $dv = k\,\nu\,c^{\nu-1}\,dc$. *On obtient alors finalement :*

$$\text{Prob}\{c \le \chi_\nu < c + dc\} = \frac{k\,\nu}{(2\pi)^{\nu/2}}\,c^{\nu-1}\,e^{-\frac{c^2}{2}}\,dc.$$

Pour obtenir maintenant la loi de probabilité de la variable χ^2_ν, *faisons le changement de variable* $x = c^2$. *On obtient :*

$$\text{Prob}\{x \le \chi^2_\nu < x + dx\} = \frac{k\,\nu}{2\,(2\pi)^{\nu/2}}\,x^{(\frac{\nu}{2}-1)}\,e^{-\frac{x}{2}}\,dx$$

qui définit la loi de probabilité cherchée, à la constante k près.

Pour calculer cette constante, on peut écrire que : $\text{Prob}\{\chi^2_\nu \ge 0\} = 1$, *soit :*

$$\frac{k\,\nu}{2\,(2\pi)^{\nu/2}}\int_0^\infty x^{(\frac{\nu}{2}-1)}\,e^{-\frac{x}{2}}\,dx = 1,$$

d'où finalement l'expression de la loi du χ^2 :

$$\text{Prob}\{x \le \chi^2_\nu < x + dx\} = \frac{x^{(\nu/2-1)}\,e^{-x/2}}{\int_0^\infty x^{(\nu/2-1)}\,e^{-x/2}\,dx}\,dx.$$

Notons qu'on définit en mathématiques les fonctions Γ *(gamma) dont les équations sont de la forme :* $\Gamma(a) = \int_0^\infty x^{a-1}\,e^{-x}\,dx$, *où a est une constante positive. L'intégrale au dénominateur de l' expression de la loi de probabilité du* χ^2 *peut alors s'écrire :* $\int_0^\infty x^{(\frac{\nu}{2}-1)}\,e^{-\frac{x}{2}}\,dx = 2^{(\frac{\nu}{2})}\,\Gamma(\frac{\nu}{2})$, *et on peut la calculer à partir des valeurs des fonctions* Γ.

Il existe des tables de la loi du χ^2 qui donnent généralement, pour chaque valeur du nombre ν de degrés de liberté, la valeur x telle que : $\text{Prob}\{\chi^2_\nu > x\} = \alpha$, où α est une probabilité donnée : 1 %, 5 %, ... Il est par conséquent inutile de retenir la formule de la loi du χ^2.

Par contre, il est indispensable de connaître les propriétés suivantes de la loi du χ^2.

2.1.1. Sommes de variables suivant des lois du χ^2

Si χ^2_1 et χ^2_2 sont deux variables *indépendantes* qui suivent des lois du χ^2 respectivement à ν_1 et ν_2 degrés de liberté, leur somme $(\chi^2_1 + \chi^2_2)$ suit une loi du χ^2 à $(\nu_1 + \nu_2)$ degrés de liberté.

Cela résulte immédiatement de la définition de la loi du χ^2.

2.1.2. Moyenne et variance d'une variable qui suit une loi du χ^2

La moyenne d'une variable suivant une loi du χ^2 à ν degrés de liberté est égale à ν et sa variance est égale à $2\,\nu$.

En effet, U étant une variable centrée réduite, sa variance est $\sigma^2(U) = E(U^2) - [E(U)]^2 = 1$ ($E(U)$ étant égal à 0) et, d'autre part, l'espérance d'une somme est égale à la somme des espérances, d'où :

$$E(\chi^2_\nu) = E(U_1{}^2) + E(U_2{}^2) + \ldots + E(U_\nu{}^2) = \nu.$$

Pour calculer la variance, on sait que la variance d'une somme de variables indépendantes est égale à la somme des variances :

$$\sigma^2(\chi_\nu^2) = \sigma^2(U_1{}^2) + \sigma^2(U_2{}^2) + \ldots + \sigma^2(U_\nu{}^2) = \nu\,\sigma^2(U^2).$$

et que la variance est égale à l'espérance du carré moins le carré de l'espérance :

$$\sigma^2(U^2) = E(U^4) - [E(U^2)]^2.$$

Enfin, on montre facilement, en intégrant par parties, que :

$$E(U^4) = \int_{-\infty}^{\infty} u^4\, e^{-\frac{u^2}{2}}\, du = 3.$$

D'où :

$$\sigma^2(\chi_\nu^2) = \nu\,(3-1) = 2\,\nu.$$

2.2. Loi de la variance d'un échantillon extrait d'une population normale dont l'écart-type est connu

Nous allons, en fait, chercher la loi de la quantité : $\sum_{i=1}^{n}(X_i - M_n)^2$ qui figure au numérateur de la variance et établir un résultat dont on pourra déduire immédiatement l'intervalle de confiance d'une variance. La démonstration peut être omise, mais les résultats suivants sont importants.

Etant donné un échantillon de taille *n* qui est extrait d'une population *normale* de variance égale à σ^2, la variable aléatoire :

$$\frac{\sum_{i=1}^{n}(X_i - M_n)^2}{\sigma^2}$$

suit une *loi du* χ^2 à $(n-1)$ degrés de liberté.

Et, d'autre part, les variables M_n et $\sum_{i=1}^{n}(X_i - M_n)^2$ sont des variables *indépendantes*. Cette dernière propriété sera utilisée un peu plus loin .

Nous avons montré, au cours de ce chapitre, que l'on peut écrire :

$$\sum_{i=1}^{n}(X_i - M_n)^2 = \sum_{i=1}^{n} X_i{}^2 - n\,M_n{}^2$$

Soit alors P le point de coordonnées $(X_1,\ X_2,\ \ldots,\ X_n)$, *et soient* $(Y_1,\ Y_2,\ \ldots,\ Y_n)$ *ses nouvelles coordonnées après un changement de coordonnées orthonormales, que nous allons choisir tel que la coordonnée* Y_n *soit justement égale à* $\sqrt{n}\ M_n$. *Dans ces conditions,* $\sum_{i=1}^{n} X_i{}^2 - n\,M_n{}^2$ *sera égal à* $\sum_{j=1}^{n-1} Y_i{}^2$, *c'est-à-dire à la somme de* $(n-1)$ *variables, dont nous allons montrer qu'elles sont indépendantes et qu'elles ont même variance* σ^2.

Ces conditions sont réalisées si l'axe des Y_n *est choisi passant par le vecteur unitaire dont toutes les coordonnées sont égales à* $\frac{1}{\sqrt{n}}$ *dans l'ancien système d'axes. Les nouvelles coordonnées de P s'écrivent* :

$$Y_1 = a_1{}^1\,X_1 + \ldots + a_i{}^1\,X_i + \ldots. + a_n{}^1\,X_n$$
$$\ldots$$
$$Y_j = a_1{}^j\,X_1 + \ldots + a_i{}^j\,X_i + \ldots. + a_n{}^j\,X_n$$
$$\ldots$$
$$Y_{n-1} = a_1{}^{n-1}\,X_1 + \ldots + a_i{}^{n-1}\,X_i + \ldots. + a_n{}^{n-1}\,X_n$$
$$Y_n = \frac{1}{\sqrt{n}}\,X_1 + \ldots + \frac{1}{\sqrt{n}}\,X_i + \ldots. + \frac{1}{\sqrt{n}}\,X_n$$

Les $a_i{}^j$ sont déterminés d'une infinité de façons par les relations :

$$\frac{a_1{}^j}{\sqrt{n}} + ... + \frac{a_i{}^j}{\sqrt{n}} + ... + \frac{a_n{}^j}{\sqrt{n}} = 0$$
$$a_1{}^i\, a_1{}^j + ... + a_i{}^i\, a_i{}^j + ... + a_n{}^i\, a_n{}^j = 0$$
$$(a_1{}^j)^2 + ... + (a_i{}^j)^2 + ... + (a_n{}^j)^2 = 1$$

Dans ces conditions, on montre facilement que les nouvelles variables Y_j suivent des lois normales, indépendantes, centrées et d'écart-type σ puisque : $E(Y_j) = 0$, $E(Y_j\, Y_{j'}) = 0$ et $E[(Y_j)^2] = \sigma^2$. *On a, d'autre part* : $\sum_{j=1}^{n} Y_j{}^2 = \sum_{i=1}^{n} X_i{}^2$ *puisqu'il s'agit d'un changement de coordonnées orthonormales. Comme Y_n a été choisi de telle sorte que* : $Y_n{}^2 = n\, M_n{}^2$, *la variable* :

$$\sum_{i=1}^{n} \frac{(X_i - M_n)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} X_i{}^2 - n\, M_n{}^2}{\sigma^2}$$

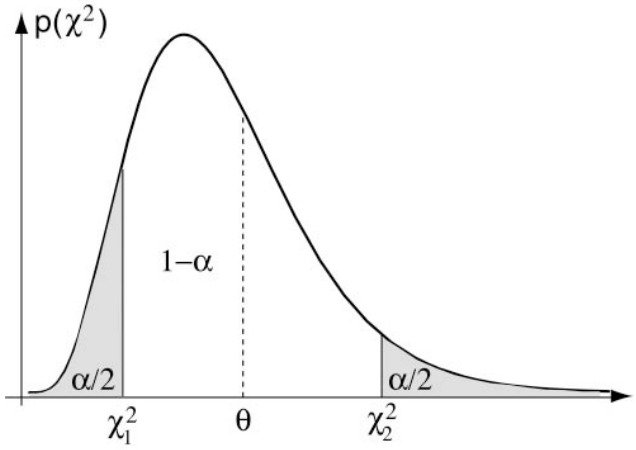
peut s'exprimer en fonction des carrés de $(n - 1)$ *variables normales, réduites, indépendantes* :

$$\sum_{i=1}^{n} \frac{(X_i - M_n)^2}{\sigma^2} = \sum_{j=1}^{n-1} \left(\frac{Y_j}{\sigma}\right)^2$$

Elle suit donc une loi du χ^2 à $(n - 1)$ degrés de liberté. Et la variable Y_n étant indépendante des variables Y_1, ..., Y_{n-1}, M_n et $\sum_{i=1}^{n} (X_i - M_n)^2$ *sont des variables indépendantes*.

2.3. Intervalle de confiance de la variance inconnue d'une population normale

Soit une population normale de variance σ^2 inconnue. La variable aléatoire $\frac{\sum_{i=1}^{n}(X_i - M_n)^2}{\sigma^2}$ suit une loi du χ^2 à $(n - 1)$ degrés de liberté.



(4.6)

Se fixant un seuil de probabilité α, il est possible de déterminer l'intervalle $[\chi_1{}^2,\ \chi_2{}^2]$ tel que :

$$\chi_1^2 < \frac{\sum_{i=1}^{n}(X_i - M_n)^2}{\sigma^2} < \chi_2^2 \quad \text{avec la probabilité } (1 - \alpha).$$

On en déduit l'intervalle de confiance pour σ^2, au risque α :

$$\frac{\sum_{i=1}^{n}(x_i - m)^2}{\chi_2^2} < \sigma^2 < \frac{\sum_{i=1}^{n}(x_i - m)^2}{\chi_1^2},$$

ou encore :

$$\frac{n\, s^2}{\chi_2^2} < \sigma^2 < \frac{n\, s^2}{\chi_1^2}.$$

3. Intervalle de confiance de la moyenne inconnue d'une population normale d'écart-type inconnu

3.1. Loi de Student

Considérons $(\nu + 1)$ variables aléatoires *normales, réduites, indépendantes* entre elles. Désignons les par U, U_1, ..., U_i, ..., U_ν. La variable :

$$T_\nu = \frac{U}{\sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu} U_i{}^2}}$$

suit, par définition, une loi de Student à ν degrés de liberté.

En remarquant que : $\sum_{i=1}^{\nu} U_i{}^2$ suit une loi du χ^2 à ν degrés de liberté, on peut encore écrire T_ν sous la forme :

$$T_\nu = \frac{U}{\sqrt{\chi_\nu^2/\nu}},$$

où U et χ_ν^2 sont des variables indépendantes qui suivent respectivement une loi normale réduite et une loi du χ^2 à ν degrés de liberté.

Pour $\nu = 1$, la loi de Student s'identifie à une loi appelée la loi de Cauchy connue pour n'avoir ni moyenne, ni variance finies. On montre d'autre part que, lorsque $\nu \to \infty$, la loi de Student tend vers une loi normale réduite. Mais, pour ν fini, elle est plus *étalée* que la loi normale, sa variance (pour $\nu > 2$) étant égale à $\frac{\nu}{\nu-2} > 1$.

Il existe des tables donnant, pour un nombre de degrés de liberté donné, et pour des seuils de probabilité α fixés les valeurs t telles que : Prob $\{ | T | > t\} = \alpha$.

3.2. Loi de la moyenne d'un échantillon extrait d'une population normale d'écart-type inconnu

En notant σ^{*2} l'estimateur sans biais de σ^2:

$$\sigma^{*2} = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1},$$

nous allons montrer que la quantité :

$$t = \frac{m-\mu}{\sigma^*/\sqrt{n}}$$

est une réalisation d'une variable de Student à $(n - 1)$ degrés de liberté.

En effet, la variable :

$$U = \frac{M_n - \mu}{\sigma/\sqrt{n}}$$

suit une loi normale réduite puisque, si les X_i suivent une loi normale de moyenne μ et d'écart-type σ, M_n suit une loi normale de moyenne μ et écart-type $\frac{\sigma}{\sqrt{n}}$. D'autre part, la variable :

$$\chi^2_{n-1} = \frac{\sum_{i=1}^{n}(X_i - M_n)^2}{\sigma^2}$$

suit une loi du χ^2 à $(n - 1)$ degrés de liberté. Et ces deux variables sont *indépendantes* .

Donc, la variable :

$$T = \frac{U}{\sqrt{\chi^2_{n-1}/(n-1)}} = \frac{M_n - \mu}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - M_n)^2}{n-1}} \Big/ \sqrt{n}}$$

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

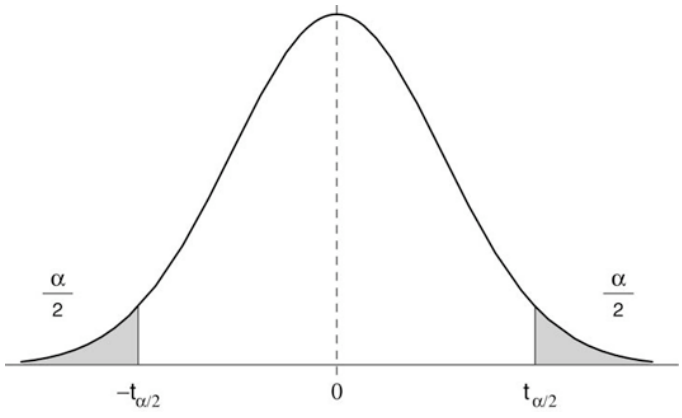
3.3. Intervalle de confiance de la moyenne inconnue d'une population normale d'écart-type inconnu

Soit $t_{\alpha/2}$ la valeur lue dans la table de Student à $(n - 1)$ degrés de liberté, correspondant au risque α réparti symétriquement. Il résulte immédiatement de ce qui précède que l'intervalle de confiance, au risque α, de la moyenne inconnue μ (avec $\mu^* = m$, son estimation) est le suivant :

$$\mu^* - t_{\alpha/2}\,\frac{\sigma^*}{\sqrt{n}} < \mu < \mu^* + t_{\alpha/2}\,\frac{\sigma^*}{\sqrt{n}}$$

expression dont la forme est la même que celle de l'intervalle de confiance établi dans le cas où l' écart-type σ est connu :

$$\mu^* - u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \mu^* + u_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}$$



(4.7)

Exercices du chapitre 4

Exercice 1

S'assurer que l'on connaît parfaitement bien les réponses aux questions suivantes.

a) Définition d'une variable normale réduite à partir d'une variable normale quelconque ?

b) Définition de la loi du χ^2 ?

c) Définition de la loi de Student ?

d) Quelle est la loi qui fait intervenir la moyenne d'un échantillon et les paramètres de la loi normale de référence ?

e) Comment estimer l'écart-type de la loi de référence

- si sa moyenne est connue,

- si elle n'est pas connue ?

f) Quelle est la loi qui fait intervenir la variance de la loi de référence et son estimation ?

g) Que devient la loi précisée en d) si l'écart-type de la loi de référence n'est pas connu et qu'il faille l'estimer ?

Exercice 2

On a mesuré la capacité (en microfarad) de 25 condensateurs et calculé la moyenne $m = 2.086$ et l'écart-type $s = 0.079$. Déterminer les intervalles de confiance de l'estimation de la moyenne μ de la population normale de référence, en choisissant un risque de 5 %, puis de 1 %. Est-il normal que le second soit plus grand que le premier ?

Exercice 3

L'airbag (ou coussin gonflable) est un système de sécurité de plus en plus souvent installé dans les automobiles. Son gonflement est assuré par un dispositif pyrotechnique dont les caractéristiques importantes sont la moyenne et l'écart-type du délai entre la mise à feu et l'explosion. Lors de l'étude d'un certain type de dispositif d'allumage, les résultats des mesures, effectuées sur 10 exemplaires, ont été (en millisecondes) : {28, 28, 31, 31, 33, 30, 31, 27, 32, 29}.

a) Calculer, au risque 5%, l'intervalle de confiance de la moyenne du délai si on connait l'écart-type de la population de référence et qu'il est égal à 2.

b) Calculer ce même intervalle si on ne connaît pas l'écart-type de la population de référence.

c) Calculer, au même risque, l'intervalle de confiance de la variance du délai, dont on déduira celui de l'écart-type dans le cas où on ne connait pas l'écart-type de la population de référence.

Exercice 4

Les poids de pièces usinées en cuivre sont distribués normalement. Ayant prélevé 9 pièces, on a obtenu les poids suivants en grammes : 18.457 ; 18.434 ; 18.444 ; 18.461 ; 18.453 ; 18.447 ; 18.452 ; 18.440 ; 18.443. Calculer m et s^2, puis les intervalles de confiance au risque 5 % de la moyenne μ et de la variance σ^2 inconnues.

Exercice 5

Pour estimer la porosité d'un certain matériau, on a fait 10 mesures et calculé $m = 94$ et $s^2 = 4$. Déterminer l'intervalle de confiance au risque 5 % de μ. Combien faudrait il faire de mesures pour que cet intervalle soit de ± 0.5 seulement ?

Exercice 6

Un fabricant de piles électriques indique sur ses produits que la durée de vie moyenne de ses piles est de 200 heures. Une association de consommateurs prélève au hazard un échantillon de 100 piles et observe une durée de vie de 190 heures en moyenne avec un écart type de 30 heures. S'agit-il de publicité mensongère ?

Exercice 7

On a mesuré les durées de vie en heures de fonctionnement de 10 tubes electroniques du même type. On a obtenu les résultats suivants : 26 ; 31 ; 34 ; 40 ; 49 ; 60 ; 72 ; 85 ; 123 ; 179. Trouver une estimation sans biais du taux de défaillance, en admettant le modèle du processus de Poisson.