Pierre Veuillez Statistiques inférentielle
Sources, et pour en savoir plus : **http://www.math-info.univ-paris5.fr/smel**

1 Problématique :

Exemple

une urne contient des boules rouges et blanches dont on ne connaît pas la composition.
En 100 tirages on a obtenu 30 Rouges et 70 Blanches.
A combien peut-on estimer la proportion de boules rouges dans l'urne ?

Formalisation X une variable aléatoire liée à une expérience aléatoire dont on ne connaît que partiellement la loi. (Ici, loi de Bernouilli valant 1 si l'on a R et 0 sinon)
Typiquement on connaît le type de la loi $\mathcal{L}$ mais pas son paramètre θ. On sait seulement que ce paramètre prend ses valeurs dans un ensemble $\Theta \subset \mathbb{R}$. (ici, le paramètre p qui est la proportion de boules Rouges)
La valeur x prise par X dans une expérience est appelée **réalisation de** X.
On cherche, via des réalisations de X à **estimer** (trouver une valeur approchée) la valeur du paramètre θ de la loi de X -estimation ponctuelle- ou un intervalle dans lequel il a une certaine probabilité de se trouver -estimation par intervalle de confiance-.
On pourra aussi faire ce travail pour d'autres grandeurs (espérance, variance ...) liées à X

Par exemple

Pour un lancer de pièce truquée, dans une suite de lancers Pile/Face on a obtenu 2 Pile 8 Face, on peut estimer que la probabilité de Pile est la **fréquence empirique** 2/10

Fréquence empirique

La fréquence empirique des succès est le nombre de succès sur le nombre d'expériences.
On peut la définir à partir de variables de Bernouilli X_i valant 1 pour succès et au $i^{\text{ème}}$ lancer et 0 sinon.
$F = \frac{\sum_{i=1}^{n} X_i}{10}$ est la fréquence empirique des succès lors des 10 premières expériences.

Modélisation

Pour modéliser la répétition de l'expérience, on se donne une liste $(X_1, \ldots, X_n)$ de variables aléatoires **indépendantes** et de même loi que X appelé n**-échantillon de variables aléatoires.**
Une liste de valeurs $(x_1, \ldots, x_n)$ prises par ces n variables est appelé n-échantillon **de données.**

2 Estimation ponctuelle

Un **estimateur** est une variable aléatoire T_n fonction du du n-échantillon de variables $T_n = f(X_1, \ldots, X_n)$ ou plus exactement une suite de telles variables$(T_n)_{n\in\mathbb{N}}$
La valeur $f(x_1, \ldots, x_n)$ souvent notée $\hat{\theta}$ prise par l'estimateur sur un n-échantillon de données est appelé **estimation** de θ. (ou d'autre grandeur)

2.1 Qualités

Biais

Le **biais** de T_n comme estimateur de θ est $b = E(T_n - \theta) = E(T_n) - \theta$. C'est l'écart moyen entre la valeur prise par T_n et la valeur à estimer θ.
Quand le biais est nul, on dit l'estimateur **sans biais;** il donne alors en moyenne la bonne valeur. Mais rien ne l'empêche de s'en éloigner car les écarts par excès et par défaut peuvent se compenser.

Exemple

Pour un lancer de pièce : $X = 1$ si Pile et $= 0$ si Face.X suit une loi de Bernouilli de paramètre $p = \mathrm{P}\,(\text{Pile})$
Et on se donne un n-échantillon de variables de même loi que X : $(X_1 \dots X_n)$
Soit $T_n = X_1$, on a $E(T_n) = E(X_1) = p$ l'estimateur est sans biais mais les valeurs prises par T_n (0 ou 1) ne s'approcheront jamais de la valeur à estimer p.

Risque quadratique.

Le **risque quadratique** de T_n comme estimateur de θ est $E\left((T_n - \theta)^2\right)$
Ici, les écarts en plus et en moins se cumulent. (le carré est positif)
De plus, l'écart de T_n avec θ étant élevé au carré, les grand écarts pèseront encore d'avantage que dans $E(|T_n - \theta|)$ par exemple.
C'est lui que l'on utilisera pour comparer deux estimateur. Plus le risque quadratique est petit, meilleur sera l'estimateur.

Théorème

Le risque quadratique est : $E\left((T_n - \theta)^2\right) = V(T_n) + b^2$ avec b le biais de T_n comme estimateur de θ.
Donc pour améliorer un estimateur,on peut diminuer sont biais, ou sa variance.

Exemple

Dans la suite de lancers Pile/Face ,

- Soit $T_n = X_1$, a pour risque quadratique $V(T_n) + b^2 = pq$: quelque soit la taille de l'échantillon, le risque quadratique restera le même.
- Soit $T'_n = \frac{\sum_{i=1}^{n} X_i}{n}$ la fréquence empirique.

  Alors son biais est $b = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) - p = \frac{1}{n} n\,p - p = 0$ donc T' est sans biais également.

  Pour calculer son risque quadratique, on cherche la variance de T'_n :

$$V\left(T'_n\right) = \frac{1}{n^2} V\left(\sum X_i\right) = \frac{1}{n^2} n\,p\,q = \frac{p\,q}{n}$$

- Donc le risque quadratique de T'_n est n fois plus petit que celui de T_n. De plus, il diminue avec la taille de l'échantillon. Plus l'échantillon est important, plus petit sera le risque quadratique.

2.2 Estimation de l'espérance

Pour une variable X ayant une espérance m et $(X_1;\dots, X_n)$ un n-échantillon de variables, l'espérance de X peut être estimée par la **moyenne empirique** : $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$

Exercice :

1. Montrer que $\overline{X_n}$ est un estimateur sans biais de m.
2. On suppose de plus que X a une variance

   Montrer qu'alors le risque quadratique de $\overline{X_n}$ tend vers 0 quand n tend vers $+\infty$

Exemple

Pour estimer le paramètre d'une loi binomiale, d'une loi de Poisson ou d'une loi Normale $\mathcal{N}(m, \nu)$: le paramètre est la moyenne.
On peut donc estimer ce paramètre par la moyenne empirique avec un risque quadratique qui tend vers 0 quand n tend vers l'infini.

2.3 Règles de calculs

$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E\left(X_i\right)$ et $\quad E\left(\alpha X\right) = \alpha E\left(X\right)$ si α est une constante.
$E\left(X\,Y\right) = E\left(X\right)E\left(Y\right)$ **si** X et Y sont indépendantes.
$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V\left(X_i\right)$ **si** les $\left(X_i\right)$ ssont indépendantes.
$V\left(\alpha X + \beta\right) = \alpha^2 V\left(X\right)$ si α et β sont une constante

3 Intervalle de confiance.

3.1 Définition

Soit X une variable aléatoire de loi $\mathcal{L}\left(\theta\right)$ et $\left(X_1 \ldots X_n\right)$ un n-échantillon de variables.
Soient U_n et V_n fonctions de cet échantillon
$[U_n, V_n]$ est un intervalle de confiance de θ de au niveau de confiance $1-\alpha$ (ou de niveau de risque α) si $\mathrm{P}\left(U_n \leq \theta \leq V_n\right) \geq 1-\alpha$
Très souvent, on prendra un intervalle centré autour d'un estimateur de θ

3.2 Inégalité de Bienaymé-Tchebichev

$\mathrm{P}\left(\left|X-m\right| \geq \varepsilon\right) \leq \frac{V(X)}{\varepsilon^2}$ donc $\mathrm{P}\left(\left|X-m\right| < \varepsilon\right) \geq 1 - \frac{V(X)}{\varepsilon^2}$ et $\mathrm{P}\left(X-\varepsilon \leq m \leq X+\varepsilon\right) \geq 1 - \frac{V(X)}{\varepsilon^2}$

3.3 Convergence : théorème de la limite centrée.

Si $\left(X_1 \ldots X_n\right)$ est un n-échantillon de variables indépendantes indépendants et de même loi que X ayant une espérance et une variance alors la loi de la moyenne empirique centrée réduite, ou de la somme centrée réduite converge en loi vers $\mathcal{N}\left(0,1\right)$ (peut être approchée par cette loi)
Ce qui se ramène à dire que la loi de $\overline{X_n}$ peut être approchée par $\mathcal{N}\left(m, \frac{\nu}{n}\right)$ (cf exercice)
Ou qu'une loi $\mathcal{B}\left(n,p\right)$ peut être approchée par $\mathcal{N}\left(n\,p, n\,p\,q\right)$ (condition : $n \geq 30$ et $n\,p \geq 15$ et $n\,p\,q \geq 5$ dans la littérature)

Exercice

Déterminer une valeur approchée de la loi de la moyenne empirique :
$E\left(\overline{X_n}\right) = E\left(X\right)$, $V\left(\overline{X_n}\right) = \frac{1}{n}V\left(X\right)$ donc $\overline{X_n} \underset{\sim}{\hookrightarrow} \mathcal{N}\left(E\left(X\right), \frac{1}{n}V\left(X\right)\right)$

3.4 Loi Normale

- **N.B.** Si $X \hookrightarrow \mathcal{N}\left(0,1\right)$ alors $\mathrm{P}\left(-t \leq X \leq t\right) = \Phi\left(t\right) - \Phi\left(-t\right) = 2\Phi\left(t\right) - 1$
- Si $X \hookrightarrow \mathcal{N}\left(m,\nu\right)$ alors $\mathrm{P}\left(X-t \leq m \leq X+t\right) = \mathrm{P}\left(\frac{-t}{\sigma} \leq \frac{X-m}{\sigma} \leq \frac{t}{\sigma}\right) = 2\Phi\left(\frac{t}{\sigma}\right) - 1$

Donc $\mathrm{P}\left(X-t \leq m \leq X+t\right) \geq 1-\alpha \Longleftrightarrow 2\Phi\left(\frac{t}{\sigma}\right) - 1 \geq 1-\alpha \Longleftrightarrow \Phi\left(\frac{t}{\sigma}\right) \geq 1-\alpha/2$

- Cas particulier : approximation de Biomiales centrée réduite : cf 4.4

Exemple :

pour $\alpha = 0,05$ (risque de 5%) on trouve $\Phi\left(1,96\right) = 0,975 = 1 - 0,05/2$ donc pour $\frac{t}{\sigma} = 1,96$ on a le risque voulu et $\mathrm{P}\left(X - 1,96\sigma \leq m \leq X + 1,96\sigma\right) \geq 0,95$ ... utilisable si on a la valeur de l'écart type (sinon, pratiquement, on en prend une estimation).

4 Exercices

4.1 Variance

Soit X ayant une espérance m et une variance v, sa **variance empirique** est $W_n = \frac{1}{n}\sum X_i^2 - \overline{X_n}^2$ avec $\overline{X_n}$ la moyenne empirique de X et $\frac{1}{n}\sum X_i^2$ la moyenne empirique de X^2.

1. Soit Y ayant une espérance et une variance. Calculer $E\left(Y^2\right)$ en fonction $E(Y)$ et $V(Y)$
2. Calculer $E\left(\overline{X_n}\right)$ et $V\left(\overline{X_n}\right)$ et en déduire $E\left(\overline{X_n}^2\right)$
3. Montrer enfin que $E(W_n) = \frac{n-1}{n}V(X)$ et en déduire un estimateur sans biais de la variance.

4.2 Question confidentielle.

Certains sujets abordés dans les enquêtes d'opinion sont parfois assez intimes, et on court le risque que les personnes interrogées se refusent à répondre franchement à l'enquêteur, faussant ainsi le résultat.
On peut alors avoir recours à une astuce consistant à inverser aléatoirement les réponses .
Considérons une question confidentielle pour laquelle on veut estimer la probabilité p de réponses positives.
L'enquêteur demande à chaque personne interrogée de lancer un dé.

- Si le dé tombe sur 6 , la personne doit donner sa réponse sans mentir,
- sinon elle doit donner l'opinion contraire à la sienne.

Si l'enquêteur ignore le résultat du dé, il ne pourra pas savoir si la réponse est franche ou non, et on peut espérer que la personne sondée acceptera de jouer le jeu.
Généralisons légèrement la situation en tirant pour chaque personne une variable de Bernoulli de paramètre α. ($\alpha = \frac{1}{6}$ dans l'exemple introductif)

- Si le résultat de cette variable est 1, la réponse est franche,
- sinon, elle est inversée.

Soit n le nombre de personnes interrogées.
L'enquêteur ne recueille que la fréquence empirique F_n des "oui".

1. Montrer que la probabilité de "oui" à l'issue de la procédure est $q = \alpha\, p + (1-\alpha)\,(1-p)$
2. Montrer que F_n, la fréquence observée par l'enquêteur, est un estimateur sans biais de q et de risque quadratique tendant vers 0 quand n tend vers $+\infty$
3. Pour $\alpha \neq 1/2$ exprimer p en en fonction de q.
4. En déduire que $T_n = \frac{F_n - 1 + \alpha}{2\alpha - 1}$ est un estimateur sans biais de p dont le risque quadratique tend vers 0 quand n tend vers $+\infty$.
5. Pour n fixé, quelle valeur attribuer à α pour que le risque quadratique soit minimum ? Est-ce acceptable ?

   Pour quelle valeur de α ce risque est-il maximum ?

   Quel sera le risque quadratique avec le dé ($\alpha = 1/6$)

4.3 Loi uniforme

Soit X de loi $\mathcal{U}[0,a]$ et $(X_1,\dots X_n)$ une n-échantillon de variables. Estimation de a :
X a une espérance de $a/2$. Soit $\overline{X_n}$ la moyenne empirique.

1. Soit $T_n = 2\overline{X_n}$. Montrer que T_n est sans biais et déterminer son risque quadratique
2. Soit $T'_n = \max(X_1,\dots,X_n)$

   Déterminer la fonction de répartition de X puis celle de T'_n

   En déduire sa densité puis son biais et son risque quadratique.
3. Soit $T''_n = \frac{n+1}{n}T'_n$ déterminer son biais et son risque quadratique.
4. Quel est le meilleur estimateur de a pour de grandes valeurs de n ?

4.4 Intervalle de confiance pour le paramètre d'une variable de Bernoulli.

Lors d'un sondage sur 100 personnes interrogée, 60 pensent voter pour A
On modélise le choix par un échantillon $(X_1,\dots,X_{100})$ de variable indépendantes de même loi de Bernouilli de paramètre p.
On cherche à déterminer un intervalle de confiance pour p au niveau de confiance 99% (1% de risque)

1. Déterminer l'espérance et la variance de la fréquence empirique $F = \frac{1}{100}\sum_{i=1}^{100} X_i$?
2. On note F^* la fréquence empirique centrée réduite.

   Par quelle loi peut on approcher celle de F^*? On suppose désormais que F^* suit $\mathcal{N}(0,1)$
3. Déterminer t tel que $\mathrm{P}(-t \leq F^* \leq t) \geq 0,99$ et en déduire que $\mathrm{P}\left(F - t\frac{\sqrt{p(1-p)}}{10} \leq p \leq F + t\frac{\sqrt{p(1-p)}}{10}\right) \geq 0,99$
4. Montrer que pour tout $p \in [0,1]$, $p(1-p) \leq \frac{1}{4}$ et en déduire que $[F - t/20\,;\, F + t/20]$ est un intervalle de confiance de p au niveau de confiance 99%

4.5 Intervalle de confiance par Bienaymé-Tchebichev

Soit $a \in \left[0; 2\sqrt{3}\right]$, $X \hookrightarrow \mathcal{U}_{[0,a]}$ et $(X_1 \dots X_n)$ un n-échantillon de variables de même loi que X et indépendantes.
On cherche un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99% (niveau de risque 1%).
On note $\overline{X_n}$ la moyenne empirique

1. Rappeler la moyenne m de X et montrer que $V(X) = \frac{a^2}{12}$. En déduire la moyenne et l'espérance de $\overline{X_n}$.
2. En déduire que $\mathrm{P}\left(\left|\overline{X_n} - \frac{a}{2}\right| > 0,1\right) \leq \frac{100}{n}$
3. Déterminer enfin n pour que $\left[\overline{X_n} - 0,1\,;\, \overline{X_n} + 0,1\right]$ soit un intervalle de confiance de $\frac{a}{2}$ au niveau de confiance 99%
4. Ecrire un programme PASCAL qui

   - choisit un nombre a au hasard dans $\left[0; 2\sqrt{3}\right]$
   - effectue 10000 tirages dans $[0,a]$
   - calcule et affiche la moyenne des résultats obtenus.

   Le programme a affiché 0,534.

   - Pensez vous que $\frac{a}{2} = 0,534$?
   - Pensez vous que $\frac{a}{2} > 0,7$?

- Pensez vous que $\frac{a}{2} \in [0,43\,;\,0,64]$?

5. dans la sutie, $n = 10000$. Par quelle loi peut-on approcher celle de $\overline{X_n}^*$ (centrée réduite) ?

6. Déterminer t pour que $\mathrm{P}\left(-t \leq \frac{\sqrt{12}}{a} 100\left(\overline{X_n} - \frac{a}{2}\right) < t\right) \geq 0,99$ et en déduire un autre intervalle de confiance de $\frac{a}{2}$ au niveau α

4.6 Comptage par capture et recapture

On cherche à évaluer le nombre N de poissons dans un étang.
Pour cela, on prélève dans l'étang m poissons que l'on bague avant les remettre dans l'étang.
On propose deux méthodes différentes d'estimation de N.

Méthode 1
Soit $n \in \mathbb{N}^*$, $n \geq m$.
On prélève des poissons dans l'étang, au hasard et avec remise.
On note X_n la variable aléatoire égale au nombre de poissons qu'il a été nécessaire de pêcher pour obtenir n poissons marqués.
Pour tout $i \in [2, n]$, on pose $D_i = X_i - X_{i-1}$. On pose $D_1 = X_1$ et on suppose que les D_i sont des variables indépendantes.

1. a) Pour tout $i \in [2, n]$, quelle est la signification de D_i ?
   b) Déterminer, pour $i \in [2, n]$, la loi de D_i, son espérance et sa variance.
   En déduire l'espérance et la variance de X_n.
   c) On pose $A_n = \frac{m}{n} X_n$. Montrer que A_n est un estimateur sans biais de N et déterminer son risque quadratique.

2. a) Pour n assez grand, par quelle loi peut-on approcher la loi de la variable aléatoire X_n^* (centrée réduite) ?

   b) On a marqué 200 poissons puis effectué 450 prélèvements pour obtenir 50 poissons marqués.
   On pose $\sigma = \sigma(A_n)$. On a pu prouver par ailleurs que $\sigma \leq 100$.
   Déterminer en fonction de σ, un intervalle de confiance pour N au seuil 0.9
   (On donne $\Phi(1,64) \simeq 0,95$).

Méthode 2
On prélève successivement et avec remise n poissons. Soit Y_n le nombre de poissons marqués parmi eux.

1. Montrer que $\frac{1}{nm} Y_n$ est un estimateur sans biais de $\frac{1}{N}$.

2. Pour quelle raison évidente ne peut-on pas prendre $\frac{nm}{Y_n}$ comme estimateur de N ?
   On pose alors $B_n = \frac{m(n+1)}{Y_n+1}$

   a) Calculer l'espérance de B_n (on montrera que $\frac{1}{k+1}\binom{n}{k} = \frac{1}{n+1}\binom{n+1}{k+1}$
   b) Est-il un estimateur sans biais de N ?