Mathématiques : statistiques et simulation

Stéphane Ducay

Université de Picardie - LAMFA CNRS UMR 6140

PAF Amiens - Formation Enseignement des Mathématiques - 28 janvier 2011

La répétition de n expériences identiques et indépendantes à deux ou trois issues peut être représentée par un arbre pondéré. La probabilité d'une liste de résultats (un chemin sur l'arbre) est égale au produit des probabilités de chaque résultat.

La répétition de n expériences identiques et indépendantes à deux ou trois issues peut être représentée par un arbre pondéré. La probabilité d'une liste de résultats (un chemin sur l'arbre) est égale au produit des probabilités de chaque résultat.

1) On peut commencer par l'exemple de 3 lancers d'une pièce équilibrée. Il y a $2^3 = 8$ chemins possibles, chaque chemin ayant pour probabilité $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$.

La répétition de n expériences identiques et indépendantes à deux ou trois issues peut être représentée par un arbre pondéré. La probabilité d'une liste de résultats (un chemin sur l'arbre) est égale au produit des probabilités de chaque résultat.

1) On peut commencer par l'exemple de 3 lancers d'une pièce équilibrée. Il y a $2^3 = 8$ chemins possibles, chaque chemin ayant pour probabilité $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$.

On peut faire remarquer la cohérence avec l'équiprobabilité "naturelle" P sur l'univers
$\Omega = \{\text{résultats de l'expérience aléatoire}\} = \{\text{chemins}\}$. L'ensemble Ω peut être décrit complètement, et on peut faire le lien entre chaque élément de l'ensemble et le chemin correspondant.

On peut alors s'intéresser à la probabilité de l'événement A_k : "obtenir k fois Pile en 3 lancers", pour $k = 0, 1, 2, 3$.
La probabilité de l'événement A_k étant égale à la somme des probabilités des événements élémentaires qui le constituent, on l'obtiendra en additionnant les probabilités de tous les chemins réalisant k Pile en 3 lancers. Comme tous les chemins ont la même probabilité $\frac{1}{8}$, on aura $P(A_k) = \frac{1}{8} \times$ nombre de chemins réalisant k "Pile" en 3 lancers. Désignant par $\binom{3}{k}$ ce nombre, on aura

$$P(A_k) = \binom{3}{k}\frac{1}{8} = \binom{3}{k}\left(\frac{1}{2}\right)^3.$$

On peut alors s'intéresser à la probabilité de l'événement A_k : "obtenir k fois Pile en 3 lancers", pour $k = 0, 1, 2, 3$.
La probabilité de l'événement A_k étant égale à la somme des probabilités des événements élémentaires qui le constituent, on l'obtiendra en additionnant les probabilités de tous les chemins réalisant k Pile en 3 lancers. Comme tous les chemins ont la même probabilité $\frac{1}{8}$, on aura $P(A_k) = \frac{1}{8} \times$ nombre de chemins réalisant k "Pile" en 3 lancers. Désignant par $\binom{3}{k}$ ce nombre, on aura

$$P(A_k) = \binom{3}{k}\frac{1}{8} = \binom{3}{k}\left(\frac{1}{2}\right)^3.$$

Sur cet exemple, on peut compter les chemins et obtenir $\binom{3}{0} = 1$, $\binom{3}{1} = 3$, $\binom{3}{2} = 3$ et $\binom{3}{3} = 1$.

2) On peut ensuite modifier cette situation en prenant une pièce truquée donnant Pile avec une probabilité $\frac{1}{4}$, et donc Face avec une probabilité $\frac{3}{4} = 1 - \frac{1}{4}$. Tous les chemins n'ont plus la même probabilité. On remarque cependant que, suivant le principe de multiplication, la probabilités de tous les chemins réalisant k "Pile" en 3 lancers ont la même probabilité $\left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{3-k}$, ce qui donne $P(A_k) = \binom{3}{k} \left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{3-k}$.

2) On peut ensuite modifier cette situation en prenant une pièce truquée donnant Pile avec une probabilité $\frac{1}{4}$, et donc Face avec une probabilité $\frac{3}{4} = 1 - \frac{1}{4}$. Tous les chemins n'ont plus la même probabilité. On remarque cependant que, suivant le principe de multiplication, la probabilités de tous les chemins réalisant k "Pile" en 3 lancers ont la même probabilité $\left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{3-k}$, ce qui donne $P(A_k) = \binom{3}{k} \left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{3-k}$.

3) On peut ensuite généraliser en remplaçant $\frac{1}{4}$ par $p \in]0;1[$, puis 3 par n, ce qui donne $P(A_k) = \binom{n}{k} p^k (1-p)^{n-k}$.

4) Le passage à une variable aléatoire X suivant la loi binomiale se fait "naturellement" : on considère la variable aléatoire X qui à chaque résultat de l'expérience (les 3 lancers) associe le nombre k de Pile obtenus. On observe alors que k peut prendre les valeurs $0, 1, 2, 3$, et que X "prend la valeur k" correspond à l'événement A_k, ce que l'on note $(X = k) = A_k$. On a alors

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

4) Le passage à une variable aléatoire X suivant la loi binomiale se fait "naturellement" : on considère la variable aléatoire X qui à chaque résultat de l'expérience (les 3 lancers) associe le nombre k de Pile obtenus. On observe alors que k peut prendre les valeurs $0, 1, 2, 3$, et que X "prend la valeur k" correspond à l'événement A_k, ce que l'on note $(X = k) = A_k$. On a alors $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

5) Reste le problème de la détermination générale des coefficients binomiaux $\binom{n}{k}$.

Considérons un entier naturel non nul n, un entier naturel k tel que $0 \leq k \leq n$. Considérons la répétition de n expériences identiques et indépendantes à deux issues (succès ou échec) représentée par un arbre pondéré.

Le coefficient binomial $\binom{n}{k}$ correpond au nombre de chemins réalisant k succès en n expériences.

Considérons un entier naturel non nul n, un entier naturel k tel que $0 \leq k \leq n$. Considérons la répétition de n expériences identiques et indépendantes à deux issues (succès ou échec) représentée par un arbre pondéré.

Le coefficient binomial $\binom{n}{k}$ correpond au nombre de chemins réalisant k succès en n expériences.

1) Propriété d'initialisation : $\binom{n}{0} = \binom{n}{n} = 1$.

En effet, il est évident qu'il n'y a qu'un seul chemin avec 0 succès (chemin avec uniquement des échecs), et un seul chemin avec n succès (chemin avec uniquement des succès).

2) Propriété de symétrie : $\binom{n}{k} = \binom{n}{n-k}$.

En effet, par symétrie succès/échec, $\binom{n}{k}$ correpond aussi au nombre de chemins réalisant k échecs en n expériences. De plus, "k échecs en n expériences" est équivalent à "$n-k$ succès en n expériences", ce qui donne le résultat.

2) Propriété de symétrie : $\binom{n}{k} = \binom{n}{n-k}$.

En effet, par symétrie succès/échec, $\binom{n}{k}$ correpond aussi au nombre de chemins réalisant k échecs en n expériences. De plus, "k échecs en n expériences" est équivalent à "$n-k$ succès en n expériences", ce qui donne le résultat.

3) Propriété de récurrence : $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$.

On peut commencer par observer ce phénomène pour $n = 2$ et $k = 0, 1$ sur l'exemple précédent de 3 lancers d'une pièce.

On peut commencer par observer ce phénomène pour $n = 2$ et $k = 0, 1$ sur l'exemple précédent de 3 lancers d'une pièce.

Plus généralement, $\binom{n+1}{k+1}$ correspond au nombre de chemins réalisant $k+1$ succès en $n+1$ expériences. Or ces chemins sont de deux types :

On peut commencer par observer ce phénomène pour $n = 2$ et $k = 0, 1$ sur l'exemple précédent de 3 lancers d'une pièce.

Plus généralement, $\binom{n+1}{k+1}$ correspond au nombre de chemins réalisant $k+1$ succès en $n+1$ expériences. Or ces chemins sont de deux types :

- il y a ceux pour lesquels il y a succès lors de la $(n+1)^{ème}$ expérience, et ils réalisent donc k succès lors des n premières expériences, soit $\binom{n}{k}$ chemins possibles ;
- il y a ceux pour lesquels il y a échec lors de la $(n+1)^{ème}$ expérience, et ils réalisent donc $k+1$ succès lors des n premières expériences, soit $\binom{n}{k+1}$ chemins possibles.

4) Triangle de Pascal.

Les propriétés 1 et 3 permettent de construire ce triangle.

Utilisation du tableur Excel : construire le triangle de Pascal pour $n = 20$.

Algorithmique : écrire un algorithme permettant de calculer $\binom{n}{k}$

4) Triangle de Pascal.

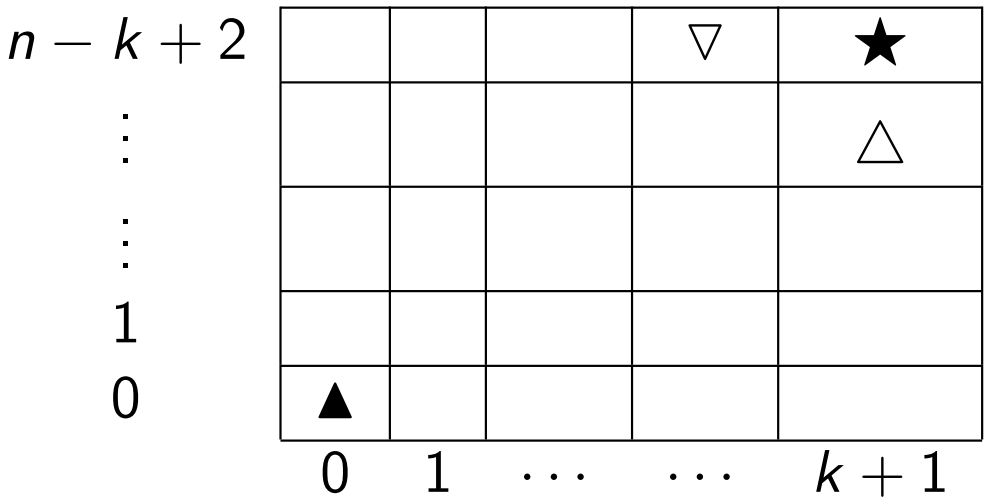
Les propriétés 1 et 3 permettent de construire ce triangle.

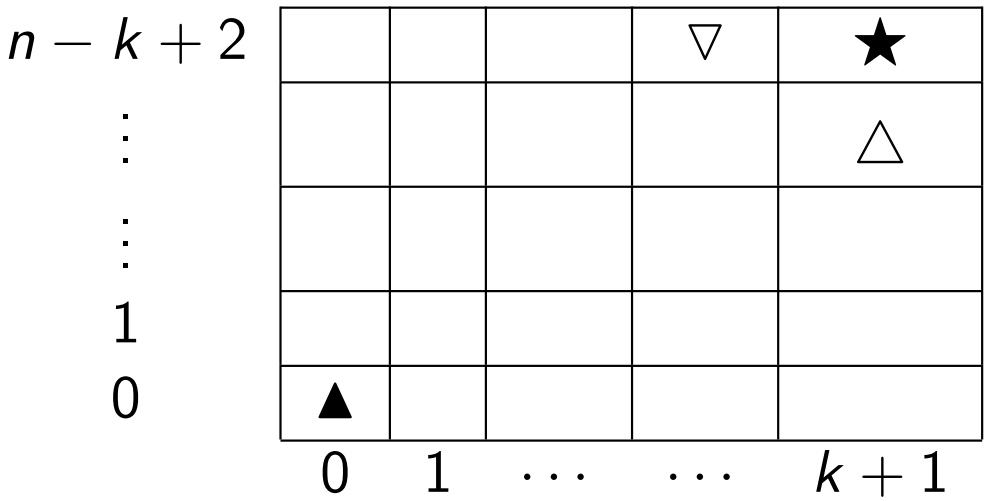
Utilisation du tableur Excel : construire le triangle de Pascal pour $n = 20$.

Algorithmique : écrire un algorithme permettant de calculer $\binom{n}{k}$

5) Une variante.

Soient n et k deux entiers supérieurs ou égaux à 1 et tels que $n - k + 2 \geqslant 1$. On considère la grille ci-dessous, constituée de $k + 2$ colonnes et $n - k + 3$ lignes, sur laquelle on se déplace de case en case. On cherche à déterminer le nombre de chemins croissants permettant de relier la case de départ ▲ à la case d'arrivée ★, les seuls déplacements autorisés étant donc vers la droite et vers le haut. A chaque étape du chemin, on choisit donc de se déplacer vers la droite ou vers le haut, en ne sortant évidemment pas de la grille.


$n-k+2$
1
0
0
1
$\cdots$
$\cdots$
$k+1$



a) Pour un chemin croissant reliant ▲ à ★, combien doit-on faire de déplacements vers la droite ? de déplacements vers le haut ? de déplacements au total ?

b) On convient d'appeler "mot" toute suite finie de lettres, même si elle ne figure pas dans le dictionnaire.

i) Combien de mots de $n+1$ lettres peut-on écrire avec les lettres D et H ? Expliquer.

ii) Combien parmi eux contiennent exactement $k+1$ fois la lettre D ?

iii) En déduire, en justifiant, le nombre de chemins croissants reliant ▲ à ★.

b) On convient d'appeler "mot" toute suite finie de lettres, même si elle ne figure pas dans le dictionnaire.

i) Combien de mots de $n+1$ lettres peut-on écrire avec les lettres D et H ? Expliquer.

ii) Combien parmi eux contiennent exactement $k+1$ fois la lettre D ?

iii) En déduire, en justifiant, le nombre de chemins croissants reliant ▲ à ★.

c) On considère les deux cases $\triangledown$ et $\triangle$ situées immédiatement à gauche et au dessous de ★.

i) En utilisant le résultat du b) iii), donner (sans calcul) le nombre de chemins croissants reliant ▲ à $\triangle$, et le nombre de chemins croissants reliant ▲ à $\triangledown$.

ii) Retrouver alors la formule de Pascal : $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$.

Soit (Ω, A, P) un espace probabilisé.
Une variable aléatoire X suit la loi de Bernoulli de paramètre $p \in]0,1[$, que l'on note $\mathcal{B}(p)$, si et seulement si X est à valeurs dans $\{0;1\}$, et $P(X=1)=p$ et $P(X=0)=p$.

Soit (Ω, A, P) un espace probabilisé.
Une variable aléatoire X suit la loi de Bernoulli de paramètre $p \in]0,1[$, que l'on note $\mathcal{B}(p)$, si et seulement si X est à valeurs dans $\{0;1\}$, et $P(X=1)=p$ et $P(X=0)=p$.

Une telle variable aléatoire permet d'indiquer si un événement A est réalisé $(X=1)$ ou pas $(X=0)$. Comme exemples d'application on peut citer :

- lancer d'une pièce menant à Pile ou Face, $A =$ "obtenir Pile" ;
- tirage d'une boule dans une urne contenant des boules blanches et noires, $A =$ "obtenir une boule blanche" ;
- vote d'un électeur, $A =$ "voter pour le candidat Untel".

Soit (Ω, A, P) un espace probabilisé.
Une variable aléatoire X suit la loi de Bernoulli de paramètre $p \in]0,1[$, que l'on note $\mathcal{B}(p)$, si et seulement si X est à valeurs dans $\{0;1\}$, et $P(X=1)=p$ et $P(X=0)=p$.

Une telle variable aléatoire permet d'indiquer si un événement A est réalisé $(X=1)$ ou pas $(X=0)$. Comme exemples d'application on peut citer :

- lancer d'une pièce menant à Pile ou Face, $A =$ "obtenir Pile" ;

- tirage d'une boule dans une urne contenant des boules blanches et noires, $A =$ "obtenir une boule blanche" ;

- vote d'un électeur, $A =$ "voter pour le candidat Untel".

Ainsi , une telle variable permet de représenter un caractère à deux modalités.

p étant donné dans $]0,1[$, on considère une urne contenant une proportion p de boules blanches. Plus précisément, on considère l'entier N plus petit multiple de 10 tel que Np soit entier, et ainsi une urne contenant N boules, dont Np boules blanches et $N(1-p)$ boules noires. Par exemple, pour $p = 0,42$, on a $N = 100$, $Np = 42$ et $N(1-p) = 58$.
On suppose que les N boules sont numérotées de 1 à N, de 1 à Np pour les boules blanches, de $Np+1$ à n pour les noires.

p étant donné dans $]0,1[$, on considère une urne contenant une proportion p de boules blanches. Plus précisément, on considère l'entier N plus petit multiple de 10 tel que Np soit entier, et ainsi une urne contenant N boules, dont Np boules blanches et $N(1-p)$ boules noires. Par exemple, pour $p = 0,42$, on a $N = 100$, $Np = 42$ et $N(1-p) = 58$.
On suppose que les N boules sont numérotées de 1 à N, de 1 à Np pour les boules blanches, de $Np+1$ à n pour les noires.

A l'expérience aléatoire " tirer une boule au hasard dans l'urne ", on peut associer l'univers $\Omega = \{1, ..., N\}$ et le munir de l'équiprobabilité P.
Dans ce contexte, l'événement A "obtenir une boule blanche" est $A = \{1, ..., Np\}$, sa probabilité étant alors

$$P(A) = \frac{cardA}{card\Omega} = \frac{Np}{N} = p.$$

Considérant la variable aléatoire X qui à chaque tirage d'une boule associe 1 si elle est blanche et 0 sinon, on a $(X = 1) = A$ et $(X = 0) = \overline{A}$, et donc $P(X = 1) = P(A) = p$ et $P(X = 0) = P\left(\overline{A}\right) = 1 - P(A) = 1 - p$.

Considérant la variable aléatoire X qui à chaque tirage d'une boule associe 1 si elle est blanche et 0 sinon, on a $(X = 1) = A$ et $(X = 0) = \overline{A}$, et donc $P(X = 1) = P(A) = p$ et $P(X = 0) = P\left(\overline{A}\right) = 1 - P(A) = 1 - p$.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 1)

Le tirage d'une boule de l'urne est simulé par l'instruction =ALEA.ENTRE.BORNES(1 ;N) à entrer dans la cellule B8 (par exemple).

La valeur correspondante de X est alors obtenue par l'instruction =SI(B8<= Np ;1 ;0).

Simulation 2

A l'expérience aléatoire "choisir un nombre au hasard dans l'intervalle $[0;1]$" on peut associer une variable aléatoire Y suit la loi Uniforme sur l'intervalle $[0;1]$ (loi à densité) ; Y indique le nombre obtenu. On sait que pour tout $y \in [0;1]$, $P(Y \leq y) = y$.

Simulation 2

A l'expérience aléatoire "choisir un nombre au hasard dans l'intervalle $[0;1]$" on peut associer une variable aléatoire Y suit la loi Uniforme sur l'intervalle $[0;1]$ (loi à densité) ; Y indique le nombre obtenu. On sait que pour tout $y \in [0;1]$, $P(Y \leq y) = y$.

p étant donné dans $]0,1[$, on a alors $P(Y \leq p) = p$. Considérant la variable aléatoire X définie par $(X = 1) = (Y \leq p)$ et $(X = 0) = \overline{(Y \leq p)} = (Y > p)$, X suit la loi de Bernoulli $\mathcal{B}(p)$.

Simulation 2

A l'expérience aléatoire "choisir un nombre au hasard dans l'intervalle $[0;1]$" on peut associer une variable aléatoire Y suit la loi Uniforme sur l'intervalle $[0;1]$ (loi à densité) ; Y indique le nombre obtenu. On sait que pour tout $y \in [0;1]$, $P(Y \leq y) = y$.

p étant donné dans $]0,1[$, on a alors $P(Y \leq p) = p$. Considérant la variable aléatoire X définie par $(X = 1) = (Y \leq p)$ et $(X = 0) = \overline{(Y \leq p)} = (Y > p)$, X suit la loi de Bernoulli $\mathcal{B}(p)$.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 2)

Une valeur de Y est simulée par l'instruction =ALEA() à entrer dans la cellule B7 (par exemple).

La valeur correspondante de X est alors obtenue par l'instruction =SI(B7<= p ;1 ;0).

Simulation 2 bis

p étant donné dans $]0,1[$ et considérant une variable aléatoire Y suit la loi Uniforme sur l'intervalle $[0;1]$, on a clairement $0 \leq Y \leq 1$ et donc $0 < p \leq Y + p \leq 1 + p < 2$.
Désignant par $[x]$ la partie entière d'un réel x, la variable aléatoire $X = [Y + p]$ ne peut prendre que les valeurs 1 ou 0. De plus,
$P(X = 1) = P([Y + p] = 1) = P(Y + p \geqslant 1) =$
$1 - P\left(\overline{Y + p \geqslant 1}\right) = 1 - P(Y + p < 1) = 1 - P(Y < 1 - p) =$
$1 - (1 - p) = p$.
Ainsi, X suit la loi de Bernoulli $\mathcal{B}(p)$.

Simulation 2 bis

p étant donné dans $]0,1[$ et considérant une variable aléatoire Y suit la loi Uniforme sur l'intervalle $[0;1]$, on a clairement $0 \leq Y \leq 1$ et donc $0 < p \leq Y + p \leq 1 + p < 2$.
Désignant par $[x]$ la partie entière d'un réel x, la variable aléatoire $X = [Y + p]$ ne peut prendre que les valeurs 1 ou 0. De plus, $P(X = 1) = P([Y + p] = 1) = P(Y + p \geqslant 1) =$
$1 - P\left(\overline{Y + p \geqslant 1}\right) = 1 - P(Y + p < 1) = 1 - P(Y < 1 - p) =$
$1 - (1 - p) = p$.
Ainsi, X suit la loi de Bernoulli $\mathcal{B}(p)$.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 2)
Une valeur de Y est simulée par l'instruction =ALEA() à entrer dans la cellule D8 (par exemple). La valeur correspondante de X est alors obtenue par l'instruction =ENT(D8+p).

Reprenons l'exemple d'une urne contenant une proportion $p = 0,42$ de boules blanches.
On tire une boule au hasard dans l'urne : le nombre de "boule blanche" obtenu en un tirage est une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$: $P(X = 1) = p = 0.42$ et $P(X = 0) = 1 - p = 0,58$. On a $E(X) = p = 0,42$ et $Var(X) = p(1 - p) = 0,2436$.

Reprenons l'exemple d'une urne contenant une proportion $p = 0,42$ de boules blanches.
On tire une boule au hasard dans l'urne : le nombre de "boule blanche" obtenu en un tirage est une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$: $P(X = 1) = p = 0.42$ et $P(X = 0) = 1 - p = 0,58$. On a $E(X) = p = 0,42$ et $Var(X) = p(1 - p) = 0,2436$.

Si on effectue $n = 50$ tirages avec remise d'une boule, on observe la réalisation de X_1 , X_2 , ... , X_{50} , variables aléatoires indépendantes de même loi que X. On dit que l'on a un échantillon aléatoire simple de taille $n = 50$ de loi de Bernoulli de paramètre $p = 0,42$.

La proportion de "boules blanches" obtenue est une variable aléatoire : $F_n = \dfrac{X_1 + X_2 + \cdots + X_{50}}{50} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre de "boules blanches" obtenues en $n = 50$ tirages.

La proportion de "boules blanches" obtenue est une variable aléatoire : $F_n = \dfrac{X_1 + X_2 + \cdots + X_{50}}{50} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre de "boules blanches" obtenues en $n = 50$ tirages.

Ayant procédé par répétitions d'expériences indépendantes, $nF_n = \sum_{i=1}^{n} X_i$ est une variable aléatoire de la loi Binomiale $\mathcal{B}(50; 0,42) = \mathcal{B}(n, p)$.

La proportion de "boules blanches" obtenue est une variable aléatoire : $F_n = \dfrac{X_1 + X_2 + \cdots + X_{50}}{50} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre de "boules blanches" obtenues en $n = 50$ tirages.

Ayant procédé par répétitions d'expériences indépendantes, $nF_n = \sum_{i=1}^{n} X_i$ est une variable aléatoire de la loi Binomiale $\mathcal{B}(50; 0,42) = \mathcal{B}(n, p)$.

On a donc $nE(F_n) = E(nF_n) = np$ et $n^2 Var(F_n) = Var(nF_n) = np(1-p)$, d'où $E(F_n) = p = 0,42$ et $Var(F_n) = \dfrac{p(1-p)}{n} = \dfrac{0,2436}{n}$.

On constate donc que lorsqu'on augmente la taille n de l'échantillon, l'espérance de F_n reste constante, égale à $0,42$, alors que la variance diminue.

On constate donc que lorsqu'on augmente la taille n de l'échantillon, l'espérance de F_n reste constante, égale à $0,42$, alors que la variance diminue.

Utilisation du tableur Excel (voir fichier excel - feuille Bernoulli simulation 1 et 2)
On reprend les simulations 1 et 2 en répétant 50 les instructions précédentes sur 50 lignes. Il suffit ensuite de "sommer" les valeurs de X obtenues pour avoir le nombre de boules blanches obtenues, puis de diviser par 50 pour avoir la fréquence.

Le cadre mathématique peut être décrit comme suit ([SN]).

Statistique et probabilités :

Description des observations et modèle théorique.

La Statistique consiste à étudier un ensemble d'objets (on parle de population, composée d'individus ou unités statistiques) sur lesquels on observe des caractéristiques, appelées variables statistiques.

Le calcul des Probabilités permet de proposer un modèle théorique d'une situation concrète afin de quantifier la fiabilité des affirmations.

Population et échantillon :

Dans certains cas on peut obtenir les valeurs de ces variables sur l'ensemble de la population ; en appliquant les méthodes de la statistique descriptive il est possible, au moyen de tableaux, graphiques, paramètres, d'analyser ces résultats. Exemples : Recensement de la population française, notes obtenues par tous les candidats à un examen, salaires de tous les employés d'une entreprise, etc...

Mais la population peut être trop vaste pour être étudiée dans sa totalité, par manque de moyens, ou de temps. (C'est le cas si on s'intéresse aux intentions de vote des Français pour une élection). Elle peut même être considérée comme infinie. C'est le cas si l'on note la qualité (défectueuse ou non) des pièces produites par un certain procédé : le nombre de ces pièces est a priori illimité, et on ne peut toutes les tester.

Population et échantillon (suite) :

De même, si l'on s'intéresse aux fréquences d'obtentions de "pile" et "face" avec une pièce de monnaie, le nombre de lancers de pièce à étudier est a priori infini : on a ici une population latente infinie.

Il arrive aussi que la mesure d'une variable soit destructrice pour l'individu : si on étudie la durée de vie de certains appareils, il serait absurde de les faire tous fonctionner jusqu'à la panne, les rendant inutilisables.

Dans tous ces cas, on est amené à n'étudier qu'une partie de la population, un échantillon, obtenu par sondage, dans le but d'extrapoler à la population entière des observations faites sur l'échantillon.

Fluctuation d'échantillonnage

Lorsqu'on étudie un caractère sur plusieurs échantillons d'une même population, on peut observer que les résultats ne sont pas identiques selon les échantillons. Plus la taille de l'échantillon étudié est grande, plus les résultats obtenus seront fiables. Cela s'explique par la diminution de la variance, et aussi par la loi des grands nombres.

La fluctuation d'échantillonnage représente la fluctuation entre les différents résultats obtenus d'une même enquête sur différents échantillons d'une même population.

Ces différents résultats présentent une certaine régularité, ce qui se traduit par la notion d'intervalle de confiance.

Considérons une population Ω sur laquelle on définit un caractère qualitatif à deux modalités A et B. On convient de représenter la modalité A par 1 et la modalité B par 0.

Considérons une population Ω sur laquelle on définit un caractère qualitatif à deux modalités A et B. On convient de représenter la modalité A par 1 et la modalité B par 0.

Le caractère est ainsi représenté par une application X de Ω dans $\mathbb{R}$ qui, à tout individu ω, associe un réel
$x = X(\omega) \in X(\Omega) = \Omega_X = \{0, 1\}$ ensemble des "valeurs" du caractère.

Considérons une population Ω sur laquelle on définit un caractère qualitatif à deux modalités A et B. On convient de représenter la modalité A par 1 et la modalité B par 0.

Le caractère est ainsi représenté par une application X de Ω dans $\mathbb{R}$ qui, à tout individu ω, associe un réel $x = X(\omega) \in X(\Omega) = \Omega_X = \{0, 1\}$ ensemble des "valeurs" du caractère.

Cette application modélise le caractère de façon déterministe : si on connaît l'individu ω, on connaît aussitôt la valeur x. Son étude relève de la statistique descriptive qui conduit, par exemple, au tableau des couples (x_i, f_i) où x_i est une valeur observée et f_i sa fréquence.

Supposons maintenant que l'on tire au hasard un individu ω dans cette population Ω pour consigner la valeur x du caractère. Ne pouvant pas prévoir quel individu précis sera tiré, on ne peut pas prévoir non plus la valeur précise de x qui sera consigner. On aimerait donc disposer d'un moyen d'attribuer une probabilité aux éléments de Ω_X.

Supposons maintenant que l'on tire au hasard un individu ω dans cette population Ω pour consigner la valeur x du caractère. Ne pouvant pas prévoir quel individu précis sera tiré, on ne peut pas prévoir non plus la valeur précise de x qui sera consigner. On aimerait donc disposer d'un moyen d'attribuer une probabilité aux éléments de Ω_X.

L'idée est de transporter sur Ω_X la probabilité P sur Ω construite pour modéliser la situation aléatoire correspondant au tirage aléatoire d'un individu.

Supposons maintenant que l'on tire au hasard un individu ω dans cette population Ω pour consigner la valeur x du caractère. Ne pouvant pas prévoir quel individu précis sera tiré, on ne peut pas prévoir non plus la valeur précise de x qui sera consigner. On aimerait donc disposer d'un moyen d'attribuer une probabilité aux éléments de Ω_X.

L'idée est de transporter sur Ω_X la probabilité P sur Ω construite pour modéliser la situation aléatoire correspondant au tirage aléatoire d'un individu.

Ici, X est une variable aléatoire de loi de Bernoulli $\mathcal{B}(p)$ où p est la proportion d'individus ayant la modalité A dans la population : $P(X = 1) = p$ et $P(X = 0) = 1 - p$.

Lorsqu'on n'a pas accès à l'ensemble de la population, la proportion p est inconnue. On procède à un **échantillonnage**, i.e. au choix de n individus dans la population, sur lesquels on observe la valeur x du caractère X. Lorsque les tirages ont lieu avec (respectivement sans) remise, l'échantillonnage est dit non-exhaustif (resp. exhaustif). Lorsque la taille n de l'échantillon est faible par rapport à celle N de la population ($N \geqslant 10n$), alors tout échantillonnage est assimilable au cas non-exhaustif.

Lorsqu'on n'a pas accès à l'ensemble de la population, la proportion p est inconnue. On procède à un **échantillonnage**, i.e. au choix de n individus dans la population, sur lesquels on observe la valeur x du caractère X. Lorsque les tirages ont lieu avec (respectivement sans) remise, l'échantillonnage est dit non-exhaustif (resp. exhaustif). Lorsque la taille n de l'échantillon est faible par rapport à celle N de la population ($N \geqslant 10n$), alors tout échantillonnage est assimilable au cas non-exhaustif.

Pour un premier échantillonnage, on observera des valeurs x_1, x_2, ..., x_n du caractère. Pour un deuxième échantillonnage de même taille, on observera des valeurs x'_1, x'_2, ..., x'_n du caractère. Et ainsi de suite. On peut alors considérer la suite x_1, x'_1, ... comme les valeurs observées d'une même variable aléatoire X_1, la suite x_2, x'_2, ... comme les valeurs observées d'une même variable aléatoire X_2, ...

Ainsi, pour tout $i = 1, ..., n$, la variable aléatoire X_i correspond aux valeurs du caractère du i-ème individu obtenu par échantillonage, et aura donc la **même loi de probabilité que** X. De plus, l'échantillonnage étant non-exhaustif (tirages avec remise), les variables aléatoires X_i sont indépendantes.

Ainsi, pour tout $i = 1, ..., n$, la variable aléatoire X_i correspond aux valeurs du caractère du i-ème individu obtenu par échantillonage, et aura donc la **même loi de probabilité que** X. De plus, l'échantillonnage étant non-exhaustif (tirages avec remise), les variables aléatoires X_i sont indépendantes.

Plus précisément, les variables aléatoires X_i sont des applications de Ω^n dans $\mathbb{R}$, qui à tout échantillonnage $(\omega_1, \omega_2, ..., \omega_n)$ associe $x_i = X_i(\omega_1, \omega_2, ..., \omega_n) = X(\omega_i)$

Ainsi, pour tout $i = 1, ..., n$, la variable aléatoire X_i correspond aux valeurs du caractère du i-ème individu obtenu par échantillonage, et aura donc la **même loi de probabilité que** X. De plus, l'échantillonnage étant non-exhaustif (tirages avec remise), les variables aléatoires X_i sont indépendantes.

Plus précisément, les variables aléatoires X_i sont des applications de Ω^n dans $\mathbb{R}$, qui à tout échantillonnage $(\omega_1, \omega_2, ..., \omega_n)$ associe $x_i = X_i(\omega_1, \omega_2, ..., \omega_n) = X(\omega_i)$

On dira que $(X_1, X_2, ..., X_n)$ est un **échantillon** (aléatoire simple) **de taille** n **de** X, et que $(x_1, x_2, ..., x_n)$ est une observation de l'échantillon.
Le terme d'échantillon désigne à la fois les n individus choisis et le n-uple de variables aléatoires $(X_1, X_2, ..., X_n)$.

Objectif : déterminer p à l'aide d'informations obtenues à partir d'un échantillonnage de taille n extrait de la population. Impossible tant que $n < N$, mais la théorie de l'échantillonnage conduit à des **estimations** $\widehat{p}$ de p, d'autant meilleures que n est grand.

Objectif : déterminer p à l'aide d'informations obtenues à partir d'un échantillonnage de taille n extrait de la population. Impossible tant que $n < N$, mais la théorie de l'échantillonnage conduit à des **estimations** $\widehat{p}$ de p, d'autant meilleures que n est grand.

Statistique : toute variable aléatoire Y fonction des variables aléatoires $X_1, X_2, ..., X_n$: $Y = \varphi(X_1, X_2, ..., X_n)$.

Objectif : déterminer p à l'aide d'informations obtenues à partir d'un échantillonnage de taille n extrait de la population. Impossible tant que $n < N$, mais la théorie de l'échantillonnage conduit à des **estimations** $\widehat{p}$ de p, d'autant meilleures que n est grand.

Statistique : toute variable aléatoire Y fonction des variables aléatoires $X_1, X_2, ..., X_n$: $Y = \varphi(X_1, X_2, ..., X_n)$.

Estimateur du paramètre p : suite $T = (T_n)$ de statistiques telles que $T_n = \varphi(X_1, X_2, ..., X_n)$ et $\lim T_n = \Theta$, où Θ est la fonction constante de valeur p.
Les variables aléatoires T_n, définies sur l'ensemble des échantillonnages de taille n, sont appelées **estimateur de taille** n de θ.

Estimation (ponctuelle) de p : toute valeur $\widehat{p}$ prise par T_n sur un échantillonnage quelconque de taille n.

Estimation (ponctuelle) de p : toute valeur $\widehat{p}$ prise par T_n sur un échantillonnage quelconque de taille n.

Un estimateur T est **sans biais** si $E(T_n) = p$ pour tout n de $\mathbb{N}^*$.

Estimation (ponctuelle) de p : toute valeur $\widehat{p}$ prise par T_n sur un échantillonnage quelconque de taille n.

Un estimateur T est **sans biais** si $E(T_n) = p$ pour tout n de $\mathbb{N}^*$.

Proportion (ou fréquence) **d'échantillon** $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre d'individus de l'échantillonnage ayant la modalité A.

Estimation (ponctuelle) de p : toute valeur $\widehat{p}$ prise par T_n sur un échantillonnage quelconque de taille n.

Un estimateur T est **sans biais** si $E(T_n) = p$ pour tout n de $\mathbb{N}^*$.

Proportion (ou fréquence) **d'échantillon** $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre d'individus de l'échantillonnage ayant la modalité A.

Pour une observation $(x_1, x_2, ..., x_n)$ de l'échantillon (en pratique on observe souvent directement $\sum_{i=1}^{n} x_i$), une **estimation ponctuelle** de p est $f_n = \dfrac{\sum_{i=1}^{n} x_i}{n} = \widehat{p}$.

Un exemple sur la proportion ([CHA])

Un groupe de 4 enfants, Alexis, Benjamin, Cyril et David, d'âges respectifs 12, 13, 14 et 15 ans.

On choisit un enfant au hasard dans le groupe, on peut considérer :

- X, indicatrice du fait que l'enfant plus 14,5 ans,

variable aléatoire de loi de Bernoulli $\mathcal{B}\left(\frac{1}{4}\right)$:

$$P(X=1)=\frac{1}{4}=p \text{ et } P(X=0)=\frac{3}{4}=1-p.$$

Cherchons à retrouver ou à approcher ces résultats à partir d'échantillons non-exhaustifs (**avec remise**) de taille $n = 3$.
Il y en a $4^3 = 64$, ils forment un univers Ω', ensemble des résultats possibles de l'expérience aléatoire "choisir un échantillon".
On peut munir Ω' de la tribu des événements $\mathcal{A}' = \mathcal{P}(\Omega')$ et de l'équiprobabilité P' sur $(\Omega', \mathcal{A}')$.
A chacun des résultats (échantillons) ω, on peut associer la proportion $F_n(\omega) = f_n$ d'enfants ayant plus de 14,5 ans.
On obtient les résultats suivants :

ω	f_n	ω	f_n
(A,A,A)	0	(B,A,A)	0
(A,A,B)	0	(B,A,B)	0
(A,A,C)	0	(B,A,C)	0
(A,A,D)	$1/3$	(B,A,D)	$1/3$
(A,B,A)	0	(B,B,A)	0
(A,B,B)	0	(B,B,B)	0
(A,B,C)	0	(B,B,C)	0
(A,B,D)	$1/3$	(B,B,D)	$1/3$
(A,C,A)	0	(B,C,A)	0
(A,C,B)	0	(B,C,B)	0
(A,C,C)	0	(B,C,C)	0
(A,C,D)	$1/3$	(B,C,D)	$1/3$
(A,D,A)	$1/3$	(B,D,A)	$1/3$
(A,D,B)	$1/3$	(B,D,B)	$1/3$
(A,D,C)	$1/3$	(B,D,C)	$1/3$

ω	f_n
(C,A,A)	0
(C,A,B)	0
(C,A,C)	0
(C,A,D)	$1/3$
(C,B,A)	0
(C,B,B)	0
(C,B,C)	0
(C,B,D)	$1/3$
(C,C,A)	0
(C,C,B)	0
(C,C,C)	0
(C,C,D)	$1/3$
(C,D,A)	$1/3$
(C,D,B)	$1/3$
(C,D,C)	$1/3$

ω	f_n
(D,A,A)	$1/3$
(D,A,B)	$1/3$
(D,A,C)	$1/3$
(D,A,D)	$2/3$
(D,B,A)	$1/3$
(D,B,B)	$1/3$
(D,B,C)	$1/3$
(D,B,D)	$2/3$
(D,C,A)	$1/3$
(D,C,B)	$1/3$
(D,C,C)	$1/3$
(D,C,D)	$2/3$
(D,D,A)	$2/3$
(D,D,B)	$2/3$
(D,D,C)	$2/3$

On définit ainsi une variable aléatoire F_n, dont on peut obtenir la loi de probabilité :

x_i	0	1/3	2/3	1
$P(F_n = x_i)$	27/64	27/64	9/64	1/64

On peut alors calculer :

- $E(F_n) = \frac{1}{4}$: on remarque que $E(F_n) = p = E(X)$.
- $Var(F_n) = \frac{1}{16}$:

  on remarque que $Var(F_n) = \frac{p(1-p)}{n} = \frac{Var(X)}{n}$.

Propriétés générales de $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$

Propriétés générales de $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$

$nF_n = \sum_{i=1}^{n} X_i$ suit la loi Binomiale $\mathcal{B}(n, p)$

$nE(F_n) = E(nF_n) = np$ et $n^2 Var(F_n) = Var(nF_n) = np(1-p)$

d'où $E(F_n) = p$ et $Var(F_n) = \dfrac{p(1-p)}{n}$.

Propriétés générales de $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$

$nF_n = \sum_{i=1}^{n} X_i$ suit la loi Binomiale $\mathcal{B}(n, p)$

$nE(F_n) = E(nF_n) = np$ et $n^2 Var(F_n) = Var(nF_n) = np(1-p)$

d'où $E(F_n) = p$ et $Var(F_n) = \dfrac{p(1-p)}{n}$.

On a ainsi $E(F_n) = p$ et on dit que F_n est un **estimateur sans biais** de p.

On a de plus $\lim_{n \to +\infty} Var(F_n) = 0$ et on dit que F_n est un **estimateur convergent** de p.

Théorème

Loi faible des grands nombres (voir [LES] par exemple)
Si les X_i sont indépendantes et admettent la même espérance p et la même variance σ^2,
alors pour tout $\varepsilon > 0$, $\lim_{n\to+\infty} P\left(|F_n - p| > \varepsilon\right) = 0$;
cette convergence étant uniforme en p.
Cela signifie que (F_n) converge en probabilité vers p.

Théorème

Loi faible des grands nombres (voir [LES] par exemple)
Si les X_i sont indépendantes et admettent la même espérance p et la même variance σ^2,
alors pour tout $\varepsilon > 0$, $\lim_{n\to+\infty} P(|F_n - p| > \varepsilon) = 0$;
cette convergence étant uniforme en p.
Cela signifie que (F_n) converge en probabilité vers p.

Preuve. Ce résultat découle de l'inégalité de Bienaymé-Tchebychev qui donne $P(|F_n - E(F_n)| > \varepsilon) \leq \frac{1}{\varepsilon^2} Var(F_n)$, et donc

$$P(|F_n - p| > \varepsilon) \leq \frac{1}{\varepsilon^2} \frac{p(1-p)}{n} \leq \frac{1}{4n\varepsilon^2}.$$

L'inégalité de B-T découle de l'inégalité de Markov

$P(X \geqslant a) \leq \frac{E(X)}{a}$, où X est une variablé aléatoire positive et a un réel > 0, appliqué avec $X = (F_n - E(F_n))^2$ et $a = \varepsilon^2$. □

Entonnoir déterministe ([SUQ])

Proposition

Si les X_i sont indépendantes, de même loi, et s'il existe des constantes a et b telles que $P(a \leq X_1 \leq b) = 1$,

alors pour tout $\varepsilon > 0$, $P(|F_n - p| > \varepsilon) \leq 2\exp\left(-n\dfrac{2\varepsilon^2}{(b-a)^2}\right)$

On peut utiliser ce type d'inégalité pour étudier quantitativement les fluctuations asymptotiques de F_n autour de $E(X_1) = p$.
Pour simplifier, on suppose désormais $a = 0$ et $b = 1$.

Entonnoir déterministe (suite)

On en déduit que pour tout entier $K \geqslant 2$ et tout réel $\alpha > 1/2$,

$$P\left(\forall n \geqslant K\ ,\ |F_n - p| \leq \sqrt{\frac{\alpha \ln n}{n}}\right) \geqslant 1 - \frac{2}{2\alpha - 1} K^{1-2\alpha}.$$

Pour $K = 200$ et $\alpha = 1$, on a donc

$$P\left(\forall n \geqslant 200\ ,\ |F_n - p| \leq \sqrt{\frac{\ln n}{n}}\right) \geqslant 1 - 2\,(200)^{-1} = 0.99.$$

On obtient ainsi un « entonnoir » déterministe qui avec une probabilité d'au moins $0,99$ encadre jusqu'à l'infini la ligne polygonale de sommets (n, F_n).

Théorème

Théorème central limite

Si les X_i sont indépendantes, de même espérance mathématique μ et de même écart-type σ,

si $Z_n = \dfrac{\overline{X_n} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ *et F_{Z_n} sa fonction de répartition,*

alors, pour tout réel x,

$$\lim_{n\to+\infty} F_{Z_n}(x) = \lim_{n\to+\infty} P(Z_n \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = \phi(x).$$

Ainsi, si X_1, X_2, ..., X_n sont n variables aléatoires indépendantes de même loi espérance mathématique μ et de même écart-type σ, on dira que

$\dfrac{\overline{X_n} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$,

autrement dit que

$\overline{X_n}$ suit approximativement la loi normale $\mathcal{N}\left(\mu; \dfrac{\sigma}{\sqrt{n}}\right)$.

Proportion d'échantillon (suite)

Si $np \geqslant 10$ et $n(1-p) \geqslant 10$, on peut approcher la loi Binomiale $\mathcal{B}(n,p)$ par la loi normale $\mathcal{N}\left(np\,;\sqrt{np(1-p)}\right)$.

On en déduit que nF_n suit approximativement la loi normale $\mathcal{N}\left(np\,;\sqrt{np(1-p)}\right)$,

et donc F_n suit approximativement la loi normale

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

Ainsi, $U = \dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$.

Commentaires de ces résultats ([SN])

F_n a toujours pour espérance p : la proportion dans l'échantillon est, "en moyenne", celle de la population.

La variance de F_n est d'autant plus faible que n est grand : la proportion dans l'échantillon varie d'autant moins d'un échantillon à l'autre que la taille de cet échantillon est grande.

A la limite, si n tend vers l'infini, $Var(F_n)$ tend vers 0 et donc F_n tend vers la constante p. La fréquence observée F tend, quand n tend vers l'infini, vers la proportion p dans la population-mère.

Ce dernier résultat est la première forme de la loi des grands nombres, formulée par Jacques Bernoulli (Ars conjectandi, 1713).

Commentaires de ces résultats (suite)
L'applet : pile ou face sur
http ://www-sop.inria.fr/mefisto/java/tutorial1/tutorial1.html
illustre la loi des grands nombres pour une proportion, number of throwes est le nombre de lancés.
On dira que **les fluctuations d'échantillonnage de f_n autour de p sont d'autant plus faibles que n est grand.**
Quand la taille de l'échantillon, n, tend vers l'infini, la fréquence observée f_n tend vers p.

Commentaires de ces résultats (suite)

De plus, si n est grand, F_n suit à peu près une loi normale.

Dans la pratique, l'approximation de la loi de F_n par une loi normale est correcte dès que $np \geqslant 10$ et $n(1-p) \geqslant 10$, ou dès que $np(1-p) > 18$, ou sous d'autres conditions proches, d'autant plus que n est grand et p proche de 0.5.

Lorsque p n'est pas connu, on vérifie ces conditions sur la fréquence f_n observée.

Dans le cas des échantillons obtenus par tirage sans remise (ou tirage exhaustif), on peut établir des formules analogues dans une population comprenant N individus :

$E(F_n) = p$ et $Var(F_n) = \dfrac{p(1-p)}{n} \times \dfrac{N-n}{N-1}$, où $\dfrac{N-n}{N-1}$ est appelé facteur d'exhaustivité ; il est < 1, mais tend vers 1 si N tend vers l'infini.

Commentaires de ces résultats (suite)

Il est possible en fait de démontrer que F_n, estimateur sans biais et convergent de p, est le plus efficace (il n'en existe pas de variance plus faible).

Ceci reste vrai, que la population soit finie ou infinie, et que le tirage soit ou non exhaustif (sans ou avec remise).

La meilleure estimation ponctuelle d'une proportion inconnue p dans une population est toujours la fréquence f_n obtenue dans l'échantillon (tirage aléatoire simple, avec ou sans remise).

Considérons une variable aléatoire X de loi de Bernoulli $\mathcal{B}(p)$, un échantillon $(X_1, X_2, ..., X_n)$ de taille n de X,

la proportion (ou fréquence) d'échantillon $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$.

On sait que si $np \geqslant 10$ et $n(1-p) \geqslant 10$, alors

$U = \dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathcal{N}(0;1)$.

On détermine le réel u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, i.e. $u_\alpha = \phi^{-1}\left(1 - \dfrac{\alpha}{2}\right)$, où ϕ est la fonction de répartition de la loi $\mathcal{N}(0;1)$. Pour $\alpha = 5\%$, on a $u_\alpha = 1.96$.

Remarque. Lorsque n est petit, on doit utiliser la loi exacte de nF_n, à savoir la loi Binomiale $\mathcal{B}(n,p)$.

Intervalle de fluctuation de la fréquence F_n

- On suppose que l'on connait p. On obtient :

$$P\left(p-\sqrt{\frac{p(1-p)}{n}}u_\alpha \leq F_n \leq p+\sqrt{\frac{p(1-p)}{n}}u_\alpha\right)=1-\alpha,$$

et donc $P(F_n \in IF_p)=1-\alpha$,

avec $IF_p=\left[p-\sqrt{\frac{p(1-p)}{n}}u_\alpha \ ; \ p+\sqrt{\frac{p(1-p)}{n}}u_\alpha\right]$

intervalle de fluctuation IF_p de F_n au niveau $1-\alpha=0.95$.

- Pour tout $p \in [0,1]$, on a $0 \leq p(1-p) \leq \frac{1}{4}$,

$0 \leq \sqrt{p(1-p)} \leq \frac{1}{2}$, $\sqrt{\frac{p(1-p)}{n}}u_\alpha \leq \frac{1.96}{2}\frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ et donc on a l'inclusion d'événements

$IF_p \subset IF'_p=\left[p-\frac{1}{\sqrt{n}} \ ; \ p+\frac{1}{\sqrt{n}}\right]$

et donc $P\left(F \in IF'_p\right) \geqslant P\left(F \in IF_p\right)=1-\alpha=0.95$;

Intervalle de fluctuation de la fréquence F_n

- On suppose que l'on connait p. On obtient :

  $$P\left(p-\sqrt{\frac{p(1-p)}{n}}u_\alpha \leq F_n \leq p+\sqrt{\frac{p(1-p)}{n}}u_\alpha\right)=1-\alpha,$$

  et donc $P(F_n \in IF_p)=1-\alpha$,

  avec $IF_p=\left[p-\sqrt{\frac{p(1-p)}{n}}u_\alpha \ ; \ p+\sqrt{\frac{p(1-p)}{n}}u_\alpha\right]$

  intervalle de fluctuation IF_p de F_n au niveau $1-\alpha=0.95$.

- Pour tout $p \in [0,1]$, on a $0 \leq p(1-p) \leq \frac{1}{4}$,

  $0 \leq \sqrt{p(1-p)} \leq \frac{1}{2}$, $\sqrt{\frac{p(1-p)}{n}}u_\alpha \leq \frac{1.96}{2}\frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ et

  donc on a l'inclusion d'événements

  $IF_p \subset IF'_p=\left[p-\frac{1}{\sqrt{n}} \ ; \ p+\frac{1}{\sqrt{n}}\right]$

  et donc $P\left(F \in IF'_p\right) \geqslant P(F \in IF_p)=1-\alpha=0.95$:

Intervalle de fluctuation de la fréquence F_n (suite)

$IF'_p = \left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$ est un intervalle de fluctuation de F_n de niveau (supérieur ou égal à) $1 - \alpha = 0.95$.

Pour tout $p \in [0.2, 0.8]$, on a $0.16 \leq p(1-p) \leq 0.25$, $0.4 \leq \sqrt{p(1-p)} \leq 0.5$ et donc, avec $u_\alpha = 1.96$,

$$0.784\frac{1}{\sqrt{n}} \leq \sqrt{\frac{p(1-p)}{n}}u_\alpha \leq 0.98\frac{1}{\sqrt{n}}.$$

Intervalle de confiance de la proportion p

On suppose que l'on ne connait pas p mais que l'on a une observation f_n de F_n à partir d'un échantillon. On a

$$P\left(F_n - \sqrt{\frac{p(1-p)}{n}}u_\alpha \leq p \leq F_n + \sqrt{\frac{p(1-p)}{n}}u_\alpha\right) = 1-\alpha, \text{ et}$$

donc $P(p \in IC_p) = 1-\alpha$, avec

$$IC_p = \left[F_n - \sqrt{\frac{p(1-p)}{n}}u_\alpha \ ; \ F_n + \sqrt{\frac{p(1-p)}{n}}u_\alpha\right]$$ **intervalle de confiance IC_p de p au niveau** $1-\alpha = 0.95$.

Bien remarquer que p est fixé et que ce sont les bornes de l'intervalle, et donc l'intervalle, qui sont aléatoires ; chaque échantillon donne a priori un intervalle différent. Cela s'interprète donc en disant que 95% des échantillons fournissent un intervalle contenant p.

Intervalle de confiance de la proportion p (suite)

Pour tout $p \in [0,1]$, on a $0 \leq p(1-p) \leq \frac{1}{4}$, $0 \leq \sqrt{p(1-p)} \leq \frac{1}{2}$,

$\sqrt{\frac{p(1-p)}{n}} u_\alpha \leq \frac{1.96}{2} \frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ et donc on a l'inclusion

d'événements $IC_p \subset IC'_p = \left[F_n - \frac{1}{\sqrt{n}} \ ; \ F_n + \frac{1}{\sqrt{n}} \right]$ et donc

$P\left(p \in IC'_p\right) \geqslant P\left(p \in IC_p\right) = 1 - \alpha = 0.95$:

$IC'_p = \left[F_n - \frac{1}{\sqrt{n}} \ ; \ F_n + \frac{1}{\sqrt{n}} \right]$ est un intervalle de confiance de la proportion p de niveau (supérieur ou égal à) $1 - \alpha = 0.95$.

Pour une observation f_n de F_n,

on obtient l'intervalle $ic'_p = \left[f_n - \frac{1}{\sqrt{n}} \ ; \ f_n + \frac{1}{\sqrt{n}} \right]$.

Intervalle de confiance de la proportion p (suite)

Remarque

Comme $\dfrac{F_n(1-F_n)}{n-1}$ est un estimateur sans biais de $\dfrac{p(1-p)}{n}$, on en déduit, si $nf_n \geqslant 10$ et $n(1-f_n) \geqslant 10$, un intervalle de confiance de la proportion p au niveau $1-\alpha$:

$$ic_p = \left[f_n - \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha \ ,\ f_n + \sqrt{\frac{f_n(1-f_n)}{n-1}}\, u_\alpha \right].$$

Exemple d'intervalle de confiance

Dans une certaine espèce de rongeur, on a compté 206 mâles sur 400 naissances.

On peut considérer la situation suivante.

Population : les rongeurs d'une certaine espèce.

Variable : le sexe, à deux modalités (mâle et femelle),

représenté par une variable aléatoire de loi de Bernoulli $\mathcal{B}(p)$,

où p est la proportion de mâles dans la population ;

on a ainsi $P(X = 1) = p$ et $P(X = 0) = 1 - p$.

Echantillon $(X_1, X_2, ..., X_n)$ de taille $n = 400$ de X.

Observation de l'échantillon : $(x_1, x_2, ..., x_n) = (1, 1, 0, 1, ..., 0)$.

Estimateur de la proportion p : $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$,

proportion (ou fréquence) de mâles dans l'échantillon,

où $\sum_{i=1}^{n} X_i$ représente le nombre de mâles de l'échantillon.

Estimation ponctuelle de la proportion p :

$$f_n = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{206}{400} = 0.515,$$

fréquence (ou proportion) de mâles dans l'observation de l'échantillon.

Intervalle de confiance de la proportion p au niveau 0,95 :

1er calcul :

$$ic'_p = \left[0.515 - \frac{1}{\sqrt{400}} \ ; \ 0.515 + \frac{1}{\sqrt{400}}\right]$$

$$ic'_p = [0.465 \ ; \ 0.565]$$

2ème calcul : $nf_n = 206 \geqslant 10$ et $n(1 - f_n) = 194 \geqslant 10$

Pour $\alpha = 0,05$ (i.e. 5%), on a $u_\alpha = 1,96$.

$$ic_p = \left[f_n - \sqrt{\frac{f_n(1-f_n)}{n-1}}u_\alpha \ ; \ f_n + \sqrt{\frac{f_n(1-f_n)}{n-1}}u_\alpha\right].$$

$$ic_p = [0,466 \ ; \ 0,564].$$

Exemple d'application de l'intervalle de fluctuation

Reprenons l'exemple précédent et supposons savoir qu'il y a équiprobabilité male/femelle à chaque naissance, autrement dit que $p = 0,5$.

Pour un échantillon de $n = 400$ naissances, l'intervalle de fluctuation de F_n est $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right] =$

$$\left[0.5 - \frac{1}{\sqrt{400}} \ ; 0.5 + \frac{1}{\sqrt{400}}\right] = [0.45 \ ; \ 0.55].$$

Ainsi, 95 % des échantillons de 400 naissances donneront une fréquence d'échantillon comprise entre 0.45 et 0.55.

On s'appuie ici sur le document d'accompagnement qui précise le contenu ≪ Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence ≫ et la capacité correspondante,
≪ Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion ≫, des programmes du lycée.

On s'appuie ici sur le document d'accompagnement qui précise le contenu ≪ Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence ≫ et la capacité correspondante, ≪ Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion ≫, des programmes du lycée.

On considère une population dans laquelle on suppose que la proportion d'un certain caractère est p. Pour juger de cette hypothèse, on y prélève, au hasard et avec remise, un échantillon de taille n sur lequel on observe une fréquence f du caractère.

On s'appuie ici sur le document d'accompagnement qui précise le contenu ≪ Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence ≫ et la capacité correspondante, ≪ Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion ≫, des programmes du lycée.

On considère une population dans laquelle on suppose que la proportion d'un certain caractère est p. Pour juger de cette hypothèse, on y prélève, au hasard et avec remise, un échantillon de taille n sur lequel on observe une fréquence f du caractère.

On rejette l'hypothèse selon laquelle la proportion dans la population est p lorsque la fréquence f observée est trop éloignée de p, dans un sens ou dans l'autre. On choisit de fixer le seuil de décision de sorte que la probabilité de rejeter l'hypothèse, alors qu'elle est vraie, soit inférieure à 5 %.

Lorsque la proportion dans la population vaut p, la variable aléatoire X correspondant au nombre de fois où le caractère est observé dans un échantillon aléatoire de taille n, suit la loi binomiale de paramètres n et p. On cherche à partager l'intervalle $[0, n]$, où X prend ses valeurs, en trois intervalles $[0, a-1]$, $[a, b]$ et $[b+1, n]$ de sorte que X prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur.

Lorsque la proportion dans la population vaut p, la variable aléatoire X correspondant au nombre de fois où le caractère est observé dans un échantillon aléatoire de taille n, suit la loi binomiale de paramètres n et p. On cherche à partager l'intervalle $[0, n]$, où X prend ses valeurs, en trois intervalles $[0, a-1]$, $[a, b]$ et $[b+1, n]$ de sorte que X prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur.

En tabulant les probabilités cumulées $P(X \leq k)$, pour k allant de 0 à n, il suffit de déterminer le plus petit entier a tel que $P(X \leq a) > 0,025$ et le plus petit entier b tel que $P(X \leq b) \geqslant 0,975$, c'est-à-dire $P(X > b) \leq 0,025$. Autrement dit, a est le plus grand entier tel que $P(X < a) \leq 0.25$. On observe aussi que $a < b$.

On a ainsi
$P((X < a) \cup (X > b)) = P(X < a) + P(X > b) \leq 0.05$
et donc $P(a \leq X \leq b) = P\left(\overline{(X < a) \cup (X > b)}\right) \geqslant 0.95$, en étant "assez proche" de 0.95.

On a ainsi
$P((X < a) \cup (X > b)) = P(X < a) + P(X > b) \leq 0.05$
et donc $P(a \leq X \leq b) = P\left(\overline{(X < a) \cup (X > b)}\right) \geqslant 0.95$, en étant "assez proche" de 0.95.

Comme $F_n = \dfrac{X}{n}$, on a ainsi $P\left(\dfrac{a}{n} \leq F_n \leq \dfrac{b}{n}\right) \geqslant 0.95$, en étant "assez proche" de 0.95.

On a ainsi
$P((X < a) \cup (X > b)) = P(X < a) + P(X > b) \leq 0.05$
et donc $P(a \leq X \leq b) = P\left(\overline{(X < a) \cup (X > b)}\right) \geqslant 0.95$, en étant "assez proche" de 0.95.

Comme $F_n = \dfrac{X}{n}$, on a ainsi $P\left(\dfrac{a}{n} \leq F_n \leq \dfrac{b}{n}\right) \geqslant 0.95$, en étant "assez proche" de 0.95.

La règle de décision est la suivante : si la fréquence observée f_n appartient à l'intervalle de fluctuation à 95 % $\left[\dfrac{a}{n}, \dfrac{b}{n}\right]$, on considère que l'hypothèse selon laquelle la proportion est p dans la population n'est pas remise en question et on l'accepte ; sinon, on rejette l'hypothèse selon laquelle cette proportion vaut p.

Pour $n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$, on observe que l'intervalle de fluctuation $\left[\frac{a}{n}, \frac{b}{n}\right]$ est sensiblement le même que l'intervalle $\left[p - \frac{1}{\sqrt{n}} \ , \ p + \frac{1}{\sqrt{n}}\right]$ proposé dans le programme de seconde.

Pour $n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$, on observe que l'intervalle de fluctuation $\left[\frac{a}{n}, \frac{b}{n}\right]$ est sensiblement le même que l'intervalle $\left[p - \frac{1}{\sqrt{n}} \ , \ p + \frac{1}{\sqrt{n}}\right]$ proposé dans le programme de seconde.

Exemple d'exercice

Monsieur Z, chef du gouvernement d'un pays lointain, affirme que 52 % des électeurs lui font confiance. On interroge 100 électeurs au hasard (la population est suffisamment grande pour considérer qu'il s'agit de tirages avec remise) et on souhaite savoir à partir de quelles fréquences, au seuil de 5 %, on peut mettre en doute le pourcentage annoncé par Monsieur Z, dans un sens, ou dans l'autre.

1. On fait l'hypothèse que Monsieur Z dit vrai et que la proportion des électeurs qui lui font confiance dans la population est $p = 0,52$. Montrer que la variable aléatoire X, correspondant au nombre d'électeurs lui faisant confiance dans un échantillon de 100 électeurs, suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

1. On fait l'hypothèse que Monsieur Z dit vrai et que la proportion des électeurs qui lui font confiance dans la population est $p = 0,52$. Montrer que la variable aléatoire X, correspondant au nombre d'électeurs lui faisant confiance dans un échantillon de 100 électeurs, suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

2. On donne ci-après un extrait de la table des probabilités cumulées $P(X \leq k)$ où X suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.
Déterminer a et b tels que définis précédemment et comparer les intervalles de fluctuation à 95 % $\left[\frac{a}{n}, \frac{b}{n}\right]$ et $\left[p - \frac{1}{\sqrt{n}} \ , \ p + \frac{1}{\sqrt{n}}\right]$.

k	$P(X \leq k)$
40	0,0106
41	0,0177
42	0,0286
43	0,0444
...	
61	0,9719
62	0,9827
63	0,9897
64	0,9941

k	$P(X \leq k)$
40	0,0106
41	0,0177
42	0,0286
43	0,0444
...	
61	0,9719
62	0,9827
63	0,9897
64	0,9941

3. Énoncer la règle décision permettant de rejeter ou non l'hypothèse $p = 0,52$, selon la valeur de la fréquence f des électeurs favorables à Monsieur Z obtenue sur l'échantillon.

4. Sur les 100 électeurs interrogés au hasard, 43 déclarent avoir confiance en Monsieur Z. Peut-on considérer, au seuil de 5 %, l'affirmation de Monsieur Z comme exacte ?

Remarque : la recherche de l'intervalle de fluctuation peut-être illustrée par le diagramme en bâton de la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

Remarque : la recherche de l'intervalle de fluctuation peut-être illustrée par le diagramme en bâton de la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

Utilisation du tableur Excel
1) Construire la table des probabilités et des probabilités cumulées de la loi Binomiale de paramètres $n = 100$ et $p = 0,52$.
2) Construire le diagramme en bâton de cette loi.

Loi d'Arcsinus avec un tableur ([LES])

- Deux joueurs A et B. Lancers d'une pièce équilibrée.
  Si Pile, A marque 1 point, si Face, B marque 1 point.
  Si le jeu (équitable) dure assez longtemps, on pourrait s'attendre à ce que A soit en avance sur B à peu près la moitié du temps, que les instants d'égalité reviennent régulièrement. C'est faux !
- Après n lancers,
  - S_n = nombre de Pile
  - $M_n = S_n - (n - S_n) = 2S_n - n$ = avance de A sur B au bout de n lancers (marche aléatoire simple)
  - $T_n = card\{k \ / \ 0 \leq k \leq n \text{ et } M_k > 0\}$ = nombre de fois où A est en avance
  - $F_n = \dfrac{T_n}{n}$ = fréquence de "A est en avance sur B"
  - $U_n = card\{k \ / \ 0 < k \leq n \text{ et } M_k = 0\}$ = nombre de fois où A et B sont à égalité

Loi d'Arcsinus avec un tableur ([LES])

- Deux joueurs A et B. Lancers d'une pièce équilibrée.
  Si Pile, A marque 1 point, si Face, B marque 1 point.
  Si le jeu (équitable) dure assez longtemps, on pourrait s'attendre à ce que A soit en avance sur B à peu près la moitié du temps, que les instants d'égalité reviennent régulièrement. C'est faux !
- Après n lancers,

  - S_n = nombre de Pile

  - $M_n = S_n - (n - S_n) = 2S_n - n$ = avance de A sur B au bout de n lancers (marche aléatoire simple)

  - $T_n = card\{k \ / \ 0 \leq k \leq n \text{ et } M_k > 0\}$ = nombre de fois où A est en avance

  - $F_n = \dfrac{T_n}{n}$ = fréquence de "A est en avance sur B"

  - $U_n = card\{k \ / \ 0 < k \leq n \text{ et } M_k = 0\}$ = nombre de fois où A et B sont à égalité

Théorème

Pour tout réel $\alpha \in]0,1[$, $\lim\limits_{n\to+\infty} P(F_n < \alpha) =$

$$\lim_{n\to+\infty} P(T_n < \alpha n) = \frac{1}{\pi}\int_0^\alpha \frac{1}{\sqrt{x(1-x)}}dx = \frac{2}{\pi}\arcsin\sqrt{\alpha}$$

Pour tout réel $\alpha > 0$, $\lim\limits_{n\to+\infty} P\left(U_n < \alpha\sqrt{n}\right) = \sqrt{\frac{2}{\pi}}\int_0^\alpha e^{-x^2/2}dx$

Application pour $n = 10000$ lancers

Pour $\alpha = 0.75$,

$$P(F_n < 0.75) = P(T_n < 7500) \simeq \frac{2}{\pi}\arcsin\sqrt{0.75} \simeq \frac{2}{3}$$

donc $P(F_n \geqslant 0.75) = P(T_n \geqslant 7500) \simeq \frac{1}{3}$.

Pour $\alpha = 3$, $P(U_n < 300) \simeq \sqrt{\frac{2}{\pi}}\int_0^3 e^{-x^2/2}dx \simeq 0.9973$.

Simulation : voir fichier excel - feuille Marche aléatoire

Références - Ouvrages

[CHV] Gérard Chauvat et al, Mathématiques BTS/DUT, Probabilités et statistique

[DEH] Catherine Déhon et al, Eléments de statistique

[LES] Emmanuel Lesigne, Pile ou Face, une introduction aux théorèmes limites du Calcul des Probabilités

Références - En ligne

[DUT] Fluctuation d'échantillonnage, Philippe Dutarte (académie de Créteil), http ://www3.ac-clermont.fr/pedago/maths-sciences-LP/beespip192$_3$22/*spip.php*?*article*367

[IRE] Discrimination et statistique, par les professeurs du groupe « Statistique et Citoyenneté » de l'IREM de Paris-Nord (et leurs élèves)

[SN] St@tNet - Les techniques de la statistique - http ://www.agro-montpellier.fr/cnam-lr/statnet/index.htm

[SUQ] Charles Suquet, Illustration de convergences de suites de variables aléatoires, Atelier aux Journées Académiques sur le Hasard, Lille 16 avril 2004, http ://math.univ-lille1.fr/ suquet/ens/IREM/AtelierSuquet-rev.pdf

[CHA] Brigitte Chaput et Stéphane Ducay, Enseignement de Statistique et Probabilités en 2ème année d'IUP MIAGE