Panorama
Probabilités - Statistiques

PROBABILITES :
- Modélisation d'une expérience aléatoire
- Loi des grands nombres
- Théorème de Moivre-Laplace
- Loi Normale, Intervalle de fluctuation

STATISTIQUES :
- Estimateur ponctuel
- Intervalle de confiance
- Test sur le paramètre p d'une loi Binomiale

Janvier 2014

N. Gozlan, P-M. Samson, Université Paris-Est Marne-la-Vallée

Document disponible sur : `http://zitt.perso.math.cnrs.fr/journee_formation.html` .1

1 PROBABILITES – Modélisation d'une expérience aléatoire

1.1 Espaces de probabilité

Définition : Espace de probabilité
Une expérience aléatoire est modélisée par la donnée :

- d'un ensemble Ω
  - $\rightsquigarrow$ permet de décrire les issues possibles de l'expérience.
  - $\rightsquigarrow$ les sous-ensembles de Ω sont appelés *événements*.
- d'une fonction $\mathbb{P} : \mathscr{P}(\Omega) \to [0,1]$, appelée *mesure de probabilité*, vérifiant les propriétés suivantes :
  - $\mathbb{P}(\Omega) = 1$,
  - Si $A \cap B = \emptyset$ alors $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
    $\rightsquigarrow$ permet de quantifier le caractère plus ou moins certain des événements.

Remarque : En réalité la probabilité $\mathbb{P}$ peut n'être définie que sur une sous-classe d'ensembles. De plus, on impose une propriété d'additivité valable pour des unions dénombrables d'événements. .2

Quelques exemples

Si
$$\Omega = \{\omega^{(1)}, \dots, \omega^{(N)}\}$$
est un ensemble fini, $\mathbb{P}$ est complètement déterminée par la donnée des nombres $\mathbb{P}(\{\omega^{(j)}\})$. Si aucune issue n'est privilégiée, on adopte la *probabilité uniforme* sur Ω définie par
$$\mathbb{P}(\{\omega^{(j)}\}) = \frac{1}{N}, \qquad \forall j \in \{1, \dots, N\}.$$

Exemples :

– Jeu de pile ou face.
$$\Omega = \{0,1\}, \qquad \mathbb{P}(\{1\}) = p, \qquad \mathbb{P}(\{0\}) = 1-p.$$
– Lancer de dé.
$$\Omega = \{1,2,3,4,5,6\}, \qquad \mathbb{P}(\{j\}) = \frac{1}{6}$$
– Tirage de carte.
$$\Omega = \text{ Jeu de carte}, \qquad \mathbb{P}(\{\text{ carte }\}) = 1/32.$$

.3

1.2 Schémas de Bernoulli

Schéma de Bernoulli

Définition informelle : Schéma de Bernoulli
Il s'agit de l'expérience aléatoire consistant à *répéter n fois de suite*, de façon identique, *une épreuve à deux issues* possibles, *le Succès* ou *l'Échec*.

Exemple concret : Une urne contient des bonbons et des cailloux. Il y a succès lorsqu'on tire un bonbon et échec sinon. On effectue n tirages, en n'oubliant pas, après chaque tirage, de remettre dans l'urne ce que l'on a tiré (*tirage avec remise*).

.4

Modélisation du Schéma de Bernoulli

– Le résultat de cette expérience aléatoire peut être modélisé par une suite de longueur n de 0 et de 1,
$$\omega = (\omega_1, \omega_2, \ldots, \omega_n),$$
avec
$$\omega_i = \begin{cases} 1 & \text{si le resultat de la } i\text{-ème épreuve est un Succès,} \\ 0 & \text{sinon.} \end{cases}$$
– L'ensemble des résultats est donc $\Omega = \{0,1\}^n$.
– Pour tout $\omega \in \Omega$, on définit
$$\begin{aligned} \mathbb{P}(\{\omega\}) &= p^{\text{Nombre de 1 dans } \omega}(1-p)^{\text{Nombre de 0 dans } \omega} \\ &= p^k(1-p)^{n-k}, \end{aligned}$$
où
– $p \in [0,1]$ est la probabilité de succès
– k est le nombre de coordonnées de ω égales à 1.

On a bien
$$\mathbb{P}(\Omega) = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} = (p+(1-p))^n = 1.$$

.5

Justification du modèle précédent

On se place dans la situation des n tirages avec remise de cailloux / bonbons. On note :
– m_0 le nombre de cailloux, m_1 le nombre de bonbons
– $m = m_0 + m_1$ et $p = m_1/m$.

On "colle" sur chaque objet un numéro pour les individualiser : les cailloux sont numérotés de 1 à m_0 et les bonbons de m_0+1 à m. On s'est ainsi ramené à un tirage avec remise de boules numérotées. Etant données les conditions dans lesquelles l'expérience est menée, on peut faire l'hypothèse d'*équiprobabilité des séquences de numéros*. Soit $\omega \in \{0,1\}^n$ et k le nombre de 1 dans ω ; on peut supposer sans perte de généralité que les k succès ont tous lieu lors des k premiers tirages.

$$\begin{aligned} &\mathbb{P}(\{\omega\}) = \\ &\frac{\text{Card}\{(x_1,\ldots,x_n); x_1,\ldots,x_k \in \{m_0+1,\ldots,m\} \text{ et } x_{k+1},\ldots,x_n \in \{1,\ldots m_0\}\}}{m^n} \\ &= \frac{m_1^k m_o^{n-k}}{m^n} = p^k(1-p)^{n-k}, \end{aligned}$$

avec $p = m_1/m$.

.6

Variables aléatoires

Définition : Variable aléatoire
Soit $(\Omega, \mathbb{P})$ un espace de probabilité. Une variable aléatoire X est une fonction définie sur Ω et à valeurs dans un autre ensemble E.

Une variable aléatoire focalise sur un certain aspect de l'expérience aléatoire. Lorsque E est un ensemble fini ou dénombrable, on parle de v.a. discrètes. Exemples :

- $\Omega = \{1,2,3,4,5,6\}^2$ et $X(\omega) = \omega_1 + \omega_2$ (score lors d'un lancer de deux dés.)
- $\Omega =$ Jeu de cartes et $X(\omega) =$ couleur de la carte.

Définition : Indépendance de variables aléatoires
Deux variables aléatoires X_1, X_2 définies sur un même espace de probabilité $(\Omega, \mathbb{P})$ sont dites *indépendantes* si

$$\mathbb{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\}) = \mathbb{P}(\{X_1 \in A_1\})\mathbb{P}(\{X_2 \in A_2\}), \qquad \forall A_1, A_2.$$

Plus généralement, l'indépendance de n variables aléatoires s'écrit

$$\begin{aligned}\mathbb{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \cdots \cap \{X_n \in A_n\}) \\ = \mathbb{P}(\{X_1 \in A_1\})\mathbb{P}(\{X_2 \in A_2\}) \cdots \mathbb{P}(\{X_n \in A_n\}), \qquad \forall A_1, A_2, \ldots, A_n.\end{aligned}$$

.7

Echantillon de Bernoulli

On revient à l'espace $\Omega = \{0,1\}^n$ muni de la probabilité

$$\mathbb{P}(\{\omega\}) = p^{\text{Nombre de 1 dans } \omega}(1-p)^{\text{Nombre de 0 dans } \omega}.$$

- On définit sur cet espace les variables aléatoires X_i, $i \in \{1, \ldots, n\}$, donnant le résultat de la i-ème épreuve :

$$X_i : \begin{array}{ccc} \Omega & \to & \{0,1\} \\ \omega & \mapsto & \omega_i. \end{array}$$

- Les variables aléatoires X_i sont *indépendantes* et suivent *une loi de Bernoulli de paramètre* p, où p est la probabilité d'obtenir un Succès lors d'une épreuve. (Dans le modèle de l'urne, p représente aussi la proportion de bonbons dans l'urne.)

$$\mathbb{P}(X_i = 1) = p, \qquad \mathbb{P}(X_i = 0) = 1 - p.$$

- Le n-uplet

$$\omega = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$$

est appelé *échantillon* de variables aléatoires de loi de Bernoulli de paramètre p.

.8

Nombre de succès - Loi binomiale

- Soit N le nombre de succès lors de l'expérience aléatoire,

$$N = N_n = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i.$$

- La variable aléatoire N suit une *loi binômiale de paramètres* n, p,

$$\mathbb{P}(N = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, \ldots, n\}.$$

- La fréquence de Succès est $F = F_n = N_n/n$. Dans ce cas il s'agit aussi de la moyenne empirique de l'échantillon $(X_1, X_2, \ldots, X_n)$

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

.9

2 Loi des grands nombres

2.1 Intuition

Intuition de la loi des grands nombres
Reprenons le modèle de l'urne. Imaginons qu'elle contienne une proportion $1/3$ de bonbons. Si on effectue un grand nombre n de tirages avec remise, alors on imagine que la fréquence F_n de succès obtenus, c'est à dire la fréquence de tirages d'un bonbon plutôt qu'un caillou, sera elle aussi proche de $1/3$. Autrement dit, plus généralement, si p est la proportion de bonbons dans l'urne, c'est à dire la probabilité de Succès,

$$F_n = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow[n\to+\infty]{} p \quad !$$

Question : Soyons plus précis, de quel type de convergence s'agit-il ici ?
– En fait, dans un cadre mathématique rigoureux qui fait appel à la théorie de la mesure, on peut démontrer que

$$\mathbb{P}\left(\lim_{n\to+\infty} F_n = p\right) = 1.$$

Ce résultat est connu sous le nom de loi forte des grands nombres.
– Nous allons nous contenter de démontrer la loi faible des grands nombres : pour tout $\varepsilon > 0$

$$\mathbb{P}\left(|F_n - p| \leq \varepsilon\right) = \mathbb{P}\left(F_n \in [p-\varepsilon, p+\varepsilon]\right) \xrightarrow[n\to+\infty]{} 1.$$

.10

2.2 Démonstration de la loi faible des grands nombres

Vers la démonstration de la loi faible des grands nombres

La loi faible des grands nombres énoncée précédemment est une conséquence de l'inégalité de Bienaymé-Tchebychev, qui est elle même une conséquence de l'inégalité de Markov.

Proposition : Inégalité de Markov
Soit Y une variable aléatoire discrète qui prend des valeurs strictement positives $y_1, y_2, \ldots, y_p$, pour tout $a > 0$,

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}.$$

Démonstration :

$$\mathbb{E}(Y) = \sum_{k=1}^{p} y_k \mathbb{P}(Y = y_k) \geq \sum_{y_k, y_k \geq a} y_k \mathbb{P}(Y = y_k) \geq a \left(\sum_{y_k, y_k \geq a} \mathbb{P}(Y = y_k) \right) = a\,\mathbb{P}(Y \geq a).$$

.11

Inégalité de Bienaymé-Tchebychev

Soit Z une variable aléatoire prenant un nombre fini de valeurs et $\varepsilon > 0$. En choisissant

$$Y = (Z - \mathbb{E}(Z))^2, \quad a = \varepsilon^2,$$

on a $\mathbb{E}(Y) = \mathrm{Var}(Z)$ et

$$\mathbb{P}(Y \geq a) = \mathbb{P}((Z - \mathbb{E}(Z))^2 \geq \varepsilon^2) = \mathbb{P}(|Z - \mathbb{E}(Z)| \geq \varepsilon).$$

L'inégalité de Markov donne ainsi l'inégalité de Bienaymé-Tchebychev.

Proposition : Inégalité de Bienaymé-Tchebychev
Si Z est une variable aléatoire prenant un nombre fini de valeurs, alors pour tout $\varepsilon > 0$

$$\mathbb{P}(|Z - \mathbb{E}(Z)| \geq \varepsilon) \leq \frac{\mathrm{Var}(Z)}{\varepsilon^2}.$$

.12

Démonstration de la loi faible des grands nombres

La loi faible des grands nombres s'obtient en appliquant l'inégalité de Bienaymé-Tchebychev à la variable aléatoire

$$Z = F_n = \frac{\sum_{i=1}^n X_i}{n}.$$

On a alors par linéarité de l'espérance

$$\mathbb{E}(Z) = \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = p,$$

et puisque la variance est homogène d'ordre 2 et que les v. a. X_i sont indépendantes

$$\mathrm{Var}(Z) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{p(1-p)}{n}.$$

(Rappelons que la variance d'une loi de Bernoulli de paramètre p est $p(1-p)$.) D'après Bienaymé-Tchebychev,

$$\mathbb{P}(|Z - \mathbb{E}(Z)| > \varepsilon) \leq \frac{\mathrm{Var}(Z)}{\varepsilon^2},$$

c'est à dire

$$\mathbb{P}(|F_n - p| > \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0,$$

donc

$$\mathbb{P}(|F_n - p| \leq \varepsilon) \xrightarrow[n\to\infty]{} 1.$$

La loi faible des grands nombres est donc une conséquence du fait que la variance de F_n décroît en $1/n$ lorsque n tend vers l'infini.

.13

Enoncé général de la loi des grands nombres

Plus généralement,

Théorème : Loi forte des grands nombres
Si $X_1, X_2, \ldots$ est une suite de variables aléatoires à valeurs réelles indépendantes et de même loi telles que $\mathbb{E}[|X_1|] < +\infty$, alors en notant $m = \mathbb{E}[X_1]$

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} = m\right) = 1$$

Théorème : Loi faible des grands nombres
Si $X_1, X_2, \ldots$ est une suite de variables aléatoires à valeurs réelles indépendantes et de même loi telles que $\mathbb{E}[|X_1|^2] < +\infty$, alors en notant $m = \mathbb{E}[X_1]$, pour tout $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - m\right| \leq \varepsilon\right) \to 1 \text{ lorsque } n \to \infty.$$

Si les variables sont discrètes la preuve précédente de la loi faible s'adapte immédiatement.

.14

Variables aléatoires à densité

Définition : Variable à densité
Soit $h : \mathbb{R} \to \mathbb{R}^+$ une fonction continue par morceaux telle que

$$\int_{-\infty}^{+\infty} h(x)\,dx = 1.$$

On dit qu'une variable aléatoire X à valeurs dans $\mathbb{R}$ a pour densité h si pour tout $a < b$ on a

$$\mathbb{P}(a \leq X \leq b) = \int_a^b h(x)\,dx.$$

Exemples :

- Loi uniforme $\mathscr{U}[a,b]$: $h(x) = \frac{1}{b-a}$ si $x \in [a,b]$ et $h(x) = 0$ sinon.
- Loi normale $\mathscr{N}(m, \sigma^2)$:

$$h(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

- Loi exponentielle $\mathscr{E}(\lambda)$: $h(x) = \lambda e^{-\lambda x}$, si $x \geq 0$ et $h(x) = 0$ sinon.

.15

Moments d'une variable à densité

Si X a une densité h telle que $\int_{-\infty}^{+\infty} |x|h(x)\,dx < +\infty$, on pose

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xh(x)\,dx$$

et si $\int_{-\infty}^{+\infty} x^2 h(x)\,dx < \infty$, on pose

$$\mathbb{E}[X^2] = \int x^2 h(x)\,dx \quad \text{et} \quad \mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

L'inégalité de Bienaymé-Tchebychev reste valable pour les variables aléatoires à densité ainsi que la preuve de la loi faible des grands nombres.

Exemples :

- Si $X \sim \mathscr{U}[a,b]$, $\mathbb{E}[X] = \frac{a+b}{2}$ et $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.
- Si $X \sim \mathscr{N}(m,\sigma^2)$, $\mathbb{E}[X] = m$, $\mathrm{Var}(X) = \sigma^2$.
- Si $X \sim \mathscr{E}(\lambda)$, $\mathbb{E}[X] = 1/\lambda$, $\mathrm{Var}(X) : 1/\lambda^2$.

.16

3 Théorème de Moivre-Laplace

Le nombre de succès N_n lors des n épreuves identiques successives suit une loi binomiale $B(n,p)$. On observe le phénomène suivant, lorsqu'on centre puis qu'on renormalise le nombre de succès N_n, la loi de la variable Z_n ainsi obtenue est proche de la loi d'une variable aléatoire Z normale centrée réduite lorsque n est grand.

$$N_n \xrightarrow{centrage} N_n - \mathbb{E}(N_n) \xrightarrow{renormalisation} Z_n = \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathrm{Var}(N_n)}} = \frac{N_n - np}{\sqrt{np(1-p)}} = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Théorème de Moivre-Laplace
Pour toutes valeurs réelles $a < b$,

$$\lim_{n\to+\infty} \mathbb{P}(a \le Z_n \le b) = \mathbb{P}(a \le Z \le b) = \int_a^b e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

Ce résultat est connu sous le nom de Théorème de Moivre-Laplace, c'est un cas particulier du théorème de la limite centrale.

Simulations, observations : TestCabri2

.17

Théorème de la limite centrale
Soit $X_1, X_2, \ldots$ une suite de v.a. indépendantes et de même loi telle que $\mathbb{E}[X_1^2] < \infty$. Posons $m = \mathbb{E}[X_1]$, $\sigma^2 = \mathrm{Var}(X_1)$ et

$$Z_n = \frac{X_1 + \cdots + X_n - nm}{\sqrt{n}\sigma}.$$

Pour tous $a < b$,

$$\mathbb{P}(a \le Z_n \le b) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx,$$

lorsque $n \to \infty$.

Remarque : Si on pose

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \qquad (\textit{moyenne empirique de l'échantillon}),$$

alors

$$Z_n = \frac{\overline{X} - \mathbb{E}[\overline{X}]}{\sqrt{\mathrm{Var}(\overline{X})}}.$$

.18

3.1 Quelle information apporte le Théorème de Moivre-Laplace ?

La loi normale centrée réduite est une loi concentrée entre les valeurs -4 et 4, en effet

$$\mathbb{P}(-3,6 \leq Z \leq 3,6) \geq 0.9996.$$

En conséquence, d'après le théorème de Moivre-Laplace, il en est de même pour Z_n lorsque n est grand. Avec grande probabilité,

$$-3,6 \leq \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 3,6,$$

c'est à dire,

$$|F_n - p| \leq \frac{3,6\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{1,8}{\sqrt{n}}.$$

Cette inégalité montre qu'avec grande probabilité F_n tend vers p avec une vitesse de l'ordre de $1,8/\sqrt{n}$. Le Théorème de Moivre-Laplace donne donc une vitesse de convergence en $1/\sqrt{n}$ dans la loi faible des grands nombres. Cette information est cruciale, elle va permettre de construire des *intervalles de fluctuation* de la fréquence observée F_n lorsque n est grand.

.19

3.2 Approximation de la loi binomiale par la loi normale

Notons $m_n = \mathbb{E}(N_n) = np$ et $\sigma_n^2 = \mathrm{Var}(N_n) = np(1-p)$, on a alors

$$Z_n = \frac{N_n - m_n}{\sigma_n}.$$

Le Théorème de Moivre-Laplace indique que pour pour tous réels $a < b$

$$\mathbb{P}\left(a \leq \frac{N_n - m_n}{\sigma_n} \leq b\right) = \mathbb{P}(a \leq Z_n \leq b) \simeq \mathbb{P}(a \leq Z \leq b).$$

Autrement dit

$$\mathbb{P}(\sigma_n a + m_n \leq N_n \leq \sigma_n b + m_n) \simeq \mathbb{P}(\sigma_n a + m_n \leq \sigma_n Z + m_n \leq \sigma_n b + m_n).$$

En posant $a' = \sigma_n a + m_n$ et $b' = \sigma_n b + m_n$, on obtient pour tous réels $a' < b'$,

$$\mathbb{P}(a' \leq N_n \leq b') \simeq \mathbb{P}(a' \leq \sigma_n Z + m_n \leq b').$$

Cette égalité signifie que la loi binomiale $B(n,p)$ de N_n est proche de la loi de $\sigma_n Z + m_n$, qui est la loi normale de moyenne $m_n = \mathbb{E}(N_n)$ et de variance $\sigma_n^2 = \mathrm{Var}(N_n)$. En effet, on a le résultat suivant,

Lemme
Si Z suit une loi normale $\mathcal{N}(0,1)$ alors $\sigma Z + m$ suit la loi normale $\mathcal{N}(m, \sigma^2)$.

On obtient ainsi le résultat suivant.

Approximation de la loi binomiale par la loi normale.
Lorsque n est grand, la loi binomiale $B(n,p)$ est proche de la loi normale $\mathcal{N}(np, np(1-p))$.

.20

4 Intervalle de fluctuation

4.1 Fluctuation d'échantillons

En pratique, un même schéma de Bernoulli répété deux fois de suite donne, au gré du hasard, deux résultats différents ω et ω', c'est à dire deux échantillons différents dans $\{0,1\}^n$,

$$\omega = (X_1(\omega), \ldots, X_n(\omega)) \qquad \text{et} \qquad \omega' = (X_1(\omega'), \ldots, X_n(\omega')).$$

En fait, chaque simulation du schéma de Bernoulli produit un nouvel échantillon, c'est le phénomène de *fluctuation des échantillons*. Par suite, la fréquence observée $F_n(\omega)$ fluctue elle aussi en fonction de l'échantillon ω.

Cependant, puisque la loi forte des grands nombres indique que

$$\mathbb{P}\left(\lim_{n \to +\infty} F_n = p\right) = 1,$$

les fluctuations de $F_n(\omega)$ tournent autour de p, si n a été choisi suffisamment grand.

.21

4.2 Intervalle de fluctuation de la fréquence de succès

Imaginons que la probabilité de succès p à chaque épreuve soit connue (par exemple $p = 1/4$).

Pour préciser l'idée précédente, considérons un intervalle centré en p,

$$I = [p-\varepsilon, p+\varepsilon]$$

(par exemple $I = [0,24; 0,26]$ pour $\varepsilon = 0,01$). Nous pouvons vérifier que

$$\mathbb{P}(F_n \in I) \text{ est proche de } 1.$$

L'intervalle I est appelé *intervalle de fluctuation* de la fréquence F_n.

1ère méthode. Lorsque p, ε, n sont fixés, on calcule exactement $\mathbb{P}(F_n \in I)$ en écrivant un algorithme,

$$\mathbb{P}(F_n \in I) = \mathbb{P}(N_n \in [n(p-\varepsilon), n(p+\varepsilon)]) = \sum_{k, n(p-\varepsilon) \leq k \leq n(p+\varepsilon)} \mathbb{P}(N_n = k),$$

avec $\mathbb{P}(N_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$. .22

2ème méthode. On peut aussi estimer $\mathbb{P}(F_n \in I)$ grâce à l'inégalité de Bienaymé-Tchebichev.

$$\begin{aligned}
\mathbb{P}(F_n \in I) &= \mathbb{P}(|N_n - np| \leq n\varepsilon) = \mathbb{P}(|N_n - \mathbb{E}(N_n)| \leq n\varepsilon) \\
&= 1 - \mathbb{P}(|N_n - \mathbb{E}(N_n)| > n\varepsilon) \\
&\geq 1 - \frac{\mathrm{Var}(N_n)}{n^2\varepsilon^2} = 1 - \frac{p(1-p)}{n\varepsilon^2},
\end{aligned}$$

car $\mathrm{Var}(N_n) = np(1-p)$. .23

3ème méthode. On peut aussi estimer $\mathbb{P}(F_n \in I)$ en utilisant le Théorème de Moivre-Laplace. Rappelons que $Z_n = \frac{N_n - np}{\sigma_n}$, avec $\sigma_n^2 = np(1-p)$,

$$\begin{aligned}
\mathbb{P}(F_n \in I) &= \mathbb{P}\left(|Z_n| \leq \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}} \right) \\
&\simeq \mathbb{P}\left(|Z| \leq \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}} \right) \\
&= \int_{-\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}}^{\frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}
\end{aligned}$$

Posons $a = \frac{\sqrt{n}\varepsilon}{\sqrt{p(1-p)}}$, on a alors $\varepsilon = a\frac{\sqrt{p(1-p)}}{\sqrt{n}}$, on vient d'établir que

$$\mathbb{P}(F_n \in I) = \mathbb{P}\left(p - a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right) \simeq \mathbb{P}(-a \leq Z \leq a).$$

Remarques :

- pour $a = 3,6$, $\mathbb{P}(-a \leq Z \leq a) \simeq 1$.
- Cette expression montre qu'un échantillon 100 fois plus important donnera, avec la même probabilité, une fluctuation 10 fois plus petite.
- On peut grâce à ces outils, répondre par exemple à la question, pour quelle taille minimum d'échantillon la fréquence d'observation de succès sera comprise entre (p-0,01) et (p+0,01) avec 95 % de chance... .24

5 STATISTIQUES – Introduction : question statistique / question probabiliste

Problème de probabilité

Une urne contient 1/4 de bonbons. Quelle est la probabilité qu'au cours de n tirages avec remise, la fréquence d'apparition d'un bonbon soit comprise entre 0,2 et 0,3 ?

$n = 50$,

$$\begin{aligned}
\mathbb{P}(0,2 \leq F_n \leq 0,3) &= \mathbb{P}\left(\frac{0,2-0,25}{\sqrt{\frac{0,25(1-0,25)}{50}}} \leq Z_n \leq \frac{0,3-0,25}{\sqrt{\frac{0,25(1-0,25)}{50}}} \right) \\
&\simeq \mathbb{P}(-0,82 \leq Z_n \leq 0,82) \\
&\simeq \mathbb{P}(|Z| \leq 0,82) \simeq 0,59 \qquad \text{où } Z \sim \mathcal{N}(0,1).
\end{aligned}$$

Problème de statistique

On effectue 50 tirages avec remise dans une urne contenant une proportion de bonbons *inconnue* p. Suite à ces tirages, la fréquence d'apparition de bonbon observée est $F_n = 0,3$.

- Peut-on donner *une estimation de* p ?
- Dans quelle mesure peut-on affirmer que $p = 1/4$? .25

6 Estimateur ponctuel du paramètre p d'une loi binomiale

Rappelons que

Loi forte des grands nombres

$$\mathbb{P}\left(\left\{\omega\in\Omega,\ \lim_{n\to+\infty} F_n(\omega)=p\right\}\right)=1.$$

Autrement dit, quel que soit l'échantillon $\omega=(X_1(\omega),\dots,X_n(\omega))$ observé, s'il est de grande taille n, la fréquence de succès $F_n(\omega)$ issue de cet échantillon *est proche de* p.

Définition

- F_n est appelé *estimateur asymptotique* de p.
- F_n est un estimateur de p dit *"sans biais"* car $\mathbb{E}(F_n)=p$. ($\mathbb{E}(F_n)$ s'interprète comme la moyenne sur toutes les observations ω de probabilité $\mathbb{P}(\omega)$.)

Cette *estimation ponctuelle* $F_n(\omega)$ de p, qui correspond à la fréquence issue d'un échantillon, *n'est pas satisfaisante*. En effet, un autre échantillon ω' obtenu lors d'une autre observation donnerait une autre fréquence $F_n(\omega')$.

Plutôt qu'une valeur unique qui estime p, nous allons construire, à partir de cette fréquence observée $F_n(\omega)$, *un intervalle de valeurs* qui contient p avec une forte probabilité.
Cet intervalle est appelé *intervalle de confiance* pour le paramètre p.

.26

7 Intervalle de confiance pour le paramètre p d'une loi binomiale

La fréquence de succès F_n varie en fonction des observations autour de la valeur p.
Rappel : (à l'aide du Théorème de Moivre-Laplace, $Z\sim\mathcal{N}(0,1)$)

$$\mathbb{P}\left(p-a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\le F_n\le p+a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\simeq\mathbb{P}(-a\le Z\le a),\quad a\ge 0.$$

Cette expression s'écrit encore

$$\mathbb{P}\left(F_n-a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\le p\le F_n+a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\simeq\mathbb{P}(-a\le Z\le a).$$

- Première approximation : Puisque $p(1-p)\le 1/4$ pour $p\in[0,1]$, on obtient

$$\mathbb{P}\left(F_n-\frac{a}{2\sqrt{n}}\le p\le F_n+\frac{a}{2\sqrt{n}}\right)\gtrsim\mathbb{P}(-a\le Z\le a).$$

Soit $\alpha\in]0,1[$ et $a(\alpha)>0$ tel que : $\mathbb{P}(-a(\alpha)\le Z\le a(\alpha))=1-\alpha$ (par exemple : $a(0,05)=1,96$).
L'intervalle (*aléatoire*)

$$J(\omega)=\left[F_n(\omega)-\frac{a(\alpha)}{2\sqrt{n}},F_n(\omega)+\frac{a(\alpha)}{2\sqrt{n}}\right]$$

est tel que $\mathbb{P}(p\in J)\gtrsim 1-\alpha$.

Définition
L'intervalle $J=\left[F_n-\frac{a(\alpha)}{2\sqrt{n}},F_n+\frac{a(\alpha)}{2\sqrt{n}}\right]$ satisfait $\mathbb{P}(p\in J)\gtrsim 1-\alpha$. C'est *un intervalle de confiance* pour le paramètre *inconnu* p. La probabilité $(1-\alpha)$ est le *niveau de confiance* (proche de 1).

Chaque échantillon observé permet de construire un tel intervalle de confiance au niveau $(1-\alpha)$ souhaité.

.27

- Seconde approximation :
Rappel :

$$\underbrace{\mathbb{P}\left(F_n-a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\le p\le F_n+a\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)}_{(e)}\simeq\mathbb{P}(-a\le Z\le a).$$

Observons que $(e)\Leftrightarrow(F_n-p)^2\le a^2\frac{p(1-p)}{n}\Leftrightarrow P_0\le p\le P_1$, avec

$$P_0=F_n-a\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}(1+\varepsilon(n)),\qquad P_1=F_n+a\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}(1+\varepsilon(n))$$

et $\lim_{n\to\infty}\varepsilon(n)=0$.

Définition

L'intervalle $J = \left[F_n - a(\alpha)\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}, F_n + a(\alpha)\frac{\sqrt{F_n(1-F_n)}}{\sqrt{n}}\right]$, est *un intervalle de confiance* pour p au niveau de confiance $(1-\alpha)$.

En effet, lorsque n est grand, on a

$$\mathbb{P}(p \in J) \simeq \mathbb{P}(P_0 \leq p \leq P_1) \simeq \mathbb{P}(-a(\alpha) \leq Z \leq a(\alpha)) = 1-\alpha.$$

Dans (e), $p(1-p) \rightsquigarrow F_n(1-F_n)$.

.28

8 Test d'hypothèse sur le paramètre p d'une loi binomiale

8.1 Introduction

La construction de ce test d'hypothèse a pour but de répondre à la question suivante.

Question :
Suite à 50 tirages avec remise dans une urne contenant une proportion de bonbons *inconnue* p, on observe une fréquence de succès $F_n = 0,3$. *Est-ce-que* $p = 1/4$ *?* Ou plus généralement, *est-ce-que* $p = p_0$ *?* (où p_0 est une proportion attendue fixée).

Considérons la variable aléatoire

$$W_n(\omega) = \frac{F_n(\omega) - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Remarque : Etant donné un échantillon ω, *cette variable est calculable*. Dans l'exemple précédent, $W_n = \frac{0,3-0,25}{\sqrt{\frac{0,25(1-0,25)}{50}}} \simeq 0,82$.

.29

Comment se comporte cette variable aléatoire W_n ?

$$W_n = \frac{F_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \underbrace{\frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}}_{=Z_n} \cdot \frac{\sqrt{p(1-p)}}{\sqrt{p_0(1-p_0)}} + \underbrace{\sqrt{n}\frac{p-p_0}{\sqrt{p_0(1-p_0)}}}_{=m_n}.$$

Rappelons que, d'après le Théorème de Moivre-Laplace,

$$Z_n \xrightarrow[n\to\infty]{} Z \sim \mathcal{N}(0,1).$$

Par conséquent, *avec grande probabilité, pour* $n \gg 1$ (n suffisamment grand),

$$|Z_n| \leq 3,6$$

car $\mathbb{P}(|Z| \leq 3,6) \simeq 0,9996$. Par suite,

- Si $p = p_0$ alors $W_n(\omega) = Z_n(\omega)$ et avec grande probabilité, pour $n \gg 1$,

$$|W_n(\omega)| \leq 3,6.$$

- Si $p \neq p_0$ alors puisque $\lim_{n\to\infty} m_n = +/- \infty$, avec grande probabilité,

$$\lim_{n\to+\infty} |W_n(\omega)| = +\infty.$$

Simulations : TestCabri2

.30

Conclusion :
Si on observe que

- $|W_n(\omega)| \leq 3,6$ alors il y a de *fortes chances* pour que $p = p_0$,
- $|W_n(\omega)| \gg 3,6$ alors il y a de *fortes chances* pour que $p \neq p_0$.

Pour être plus précis et mesurer cette *"forte chance"*, on construit un test, dit

$$\textit{test d'hypothèse de } H_0 = \{p = p_0\} \textit{ contre } H_1 = \{p \neq p_0\},$$

en établissant *un seuil* a tel que

- si $|W_n(\omega)| \leq a$ alors *on accepte* l'hypothèse $H_0 = \{p = p_0\}$,
- si $|W_n(\omega)| > a$ alors *on rejette* l'hypothèse $H_0 = \{p = p_0\}$, on accepte donc l'hypothèse $H_1 = \{p \neq p_0\}$.

.31

8.2 Construction du test

Construction du test

Comment choisir le seuil a ?

– 1ère remarque : *Supposons* qu'en effet $p = p_0$, alors $W_n = Z_n$.Cependant, quelle que soit le seuil a, si $n \gg 1$,

$$\mathbb{P}_{\{p=p_0\}}(|W_n| > a) \neq 0.$$

Donc, il se peut qu'*on observe* $|W_n(\omega)| > a$. *On rejettera* alors l'hypothèse $H_0 = \{p = p_0\}$ *à tort !*

Définition
$\mathbb{P}_{\{p=p_0\}}(|W_n| > a)$ est le *risque d'erreur de première espèce*.

On contrôle ce risque d'erreur en choisissant *un seuil* $a = a(\alpha)$ tel que

$$\mathbb{P}_{\{p=p_0\}}(|W_n| > a(\alpha)) \simeq \alpha, \qquad (\alpha \text{ proche de } 0).$$

En effet puisque $n \gg 1$,

$$\mathbb{P}_{\{p=p_0\}}(|W_n| > a(\alpha)) = \mathbb{P}(|Z_n| > a(\alpha)) \simeq \mathbb{P}(|Z| > a(\alpha)), \quad \text{où } Z \sim \mathcal{N}(0,1).$$

Le seuil $a(\alpha)$ est donc obtenu à partir des *tables de la loi Gaussienne standard* de telle sorte que

$$\mathbb{P}(|Z| > a(\alpha)) = \alpha.$$

Exemple : pour $\alpha = 0,05 = 5\%$, le seuil est $a(\alpha) = a(0,05) \simeq 1,96$.

Simulations : TestCabri2, TestCabri2 .32

– Une fois le seuil $a(\alpha)$ établi, le test est entièrement construit ! Dans l'exemple précédent ($F_n = 0,3$, $p_0 = 1/4$),

$$|W_n(\omega)| \simeq 0,82 \leq 1,96.$$

Conclusion : Avec un risque d'erreur de première espèce de 5%, la proportion de bonbons dans l'urne est $p = 1/4$.

– 2ème remarque : *Supposons* qu'en fait $p \neq p_0$,

$$W_n = Z_n \frac{\sqrt{p(1-p)}}{\sqrt{p_0(1-p_0)}} + \sqrt{n}\frac{p-p_0}{\sqrt{p_0(1-p_0)}},$$

et par conséquent, pour $n \gg 1$,

$$\mathbb{P}_{p\neq p_0}(|W_n| \leq a(\alpha)) = \mathbb{P}(a_n \leq Z_n \leq b_n) \simeq \mathbb{P}(a_n \leq Z \leq b_n) \neq 0,$$

où

$$a_n = -a(\alpha)\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} - \sqrt{n}\frac{p-p_0}{\sqrt{p(1-p)}},$$

et

$$b_n = a(\alpha)\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} - \sqrt{n}\frac{p-p_0}{\sqrt{p(1-p)}}.$$

.33

Puisque

$$\mathbb{P}_{p\neq p_0}(|W_n| \leq a(\alpha)) \neq 0,$$

on peut avoir observé que $|W_n| \leq a(\alpha)$ alors que $p \neq p_0$, on a alors *accepté* l'hypothèse $H_0 = \{p = p_0\}$ *à tort !*

Définition
$\mathbb{P}_{p\neq p_0}(|W_n| \leq a(\alpha))$ est le *risque d'erreur de seconde espèce*.

Remarque : Le test est entièrement construit, on ne peut donc pas contrôler ce risque d'erreur. Cependant,

$$\lim_{n\to\infty} \mathbb{P}_{p\neq p_0}(|W_n| \leq a(\alpha)) = 0,$$

le risque diminue avec la taille de l'échantillon n. En effet, quand $n \to \infty$,

– si $p > p_0$ alors

$$a_n = -a(\alpha)\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} - \sqrt{n}\frac{p-p_0}{\sqrt{p(1-p)}} \to -\infty,$$

$$b_n = a(\alpha)\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}} - \sqrt{n}\frac{p-p_0}{\sqrt{p(1-p)}} \to -\infty,$$

– si $p < p_0$ alors $a_n \to +\infty$ et $b_n \to +\infty$.

Donc, lorsque $p \neq p_0$ et $n \to \infty$,

$$\mathbb{P}_{p\neq p_0}(|W_n| \leq a(\alpha)) \simeq \mathbb{P}(a_n \leq Z \leq b_n) \to 0, \quad Z \sim \mathcal{N}(0,1).$$

Simulations : TestCabri2 .34