Intervalle de fluctuation intervalle de confiance urnes et M&M's

A. Rolland, IUT LUMIERE LYON II

Séminaire IREM Lyon - 21 juin 2019

QUI SUIS-JE ?

Antoine ROLLAND

- Etudes : agrégation de mathématique, thèse en informatique théorique
- Poste : Maître de conférence en statistique à l'IUT Lumière Lyon II
  DUT Statistique et Informatique Décisionnelle (STID)
- liens entre stats et décision multicritère

`http://eric.univ-lyon2.fr/~arolland/`

SOMMAIRE

Rappel du Sommaire

Théorème
Soient $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes suivant toutes la même loi $\mathcal{L}$ de moyenne m et d'écart-type σ.
Soit $\overline{X_n}$ la moyenne de $X_1, X_2, \ldots, X_n$

la variable aléatoire $\overline{X_n}$ converge en loi vers la loi normale $\mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$

Processus d'échantillonnage

- Cadre : on connaît la **population**, on veut deviner ce qui se passe dans **l'échantillon**.
- Données : une population sur laquelle on observe une information (=une variable statistique), qui peut être numérique (on s'intéresse alors à la somme ou la moyenne dans l'échantillon) ou booléenne (on s'intéresse alors à la proportion dans l'échantillon). c'est la même chose car une proportion est juste une moyenne sur une variable booléenne.

Processus d'estimation

- Cadre : on connaît un **échantillon**, on veut deviner la valeur du paramètre dans la **population**.
- Données : un échantillon de taille n sur laquelle on observe une information (=une variable statistique). On peut s'intéresser à la loi de distribution de la variable, sa moyenne, son écart-type ou...

Rappel du Sommaire

Processus d'échantillonnage

- On connait la distribution de la population
- On tire un échantillon de taille n
- On connait la loi de probabilité de la moyenne de l'échantillon
- On peut donc calculer la probabilité d'obtenir une valeur ou un intervalle de valeurs particulier

Intervalle de fluctuation

Intervalle de fluctuation pour un risque α : c'est l'intervalle qui contient la moyenne dans $(1 - \alpha)$ des cas.

- Il n'y a pas unicité, mais deux types sont "logiques"
- intervalle centré sur la moyenne
- intervalle unilatéral (non borné d'un côté)

Dans les programmes de lycée (jusqu'à maintenant...) : approximation de l'IF par utilisation du TCL plus ou moins brutal.

Pourquoi utiliser des outils approximatifs et faux alors qu'on peut utiliser des outils exacts et vrais ?

Quelques interrogations du professeur de lycee autour des intervalles de fluctuation, Véronique Cerclé, repère IREM 91, 2013
Remarques sur l'enseignement des probabilités et de la statistique au Lycée, Daniel Perrin, S&E, 2015

Exemple de l'urne

- 1800 boules blanches et 400 boules rouges.
- variable d'intérêt : la proportion de boules rouges
- échantillons de taille $n \in \{25, 50, 100\}$

INTERVALLE DE FLUCTUATION

La proportion de boules rouges se situe 95% des fois dans l'intervalle

$$]p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}[$$

D'où ça vient ? 3 approximations :

- on approche une loi binomiale (discrète) par une loi normale (continue)
- on approche 1,96 par 2
- on approche $\sqrt{p(1-p)}$ par 0,5

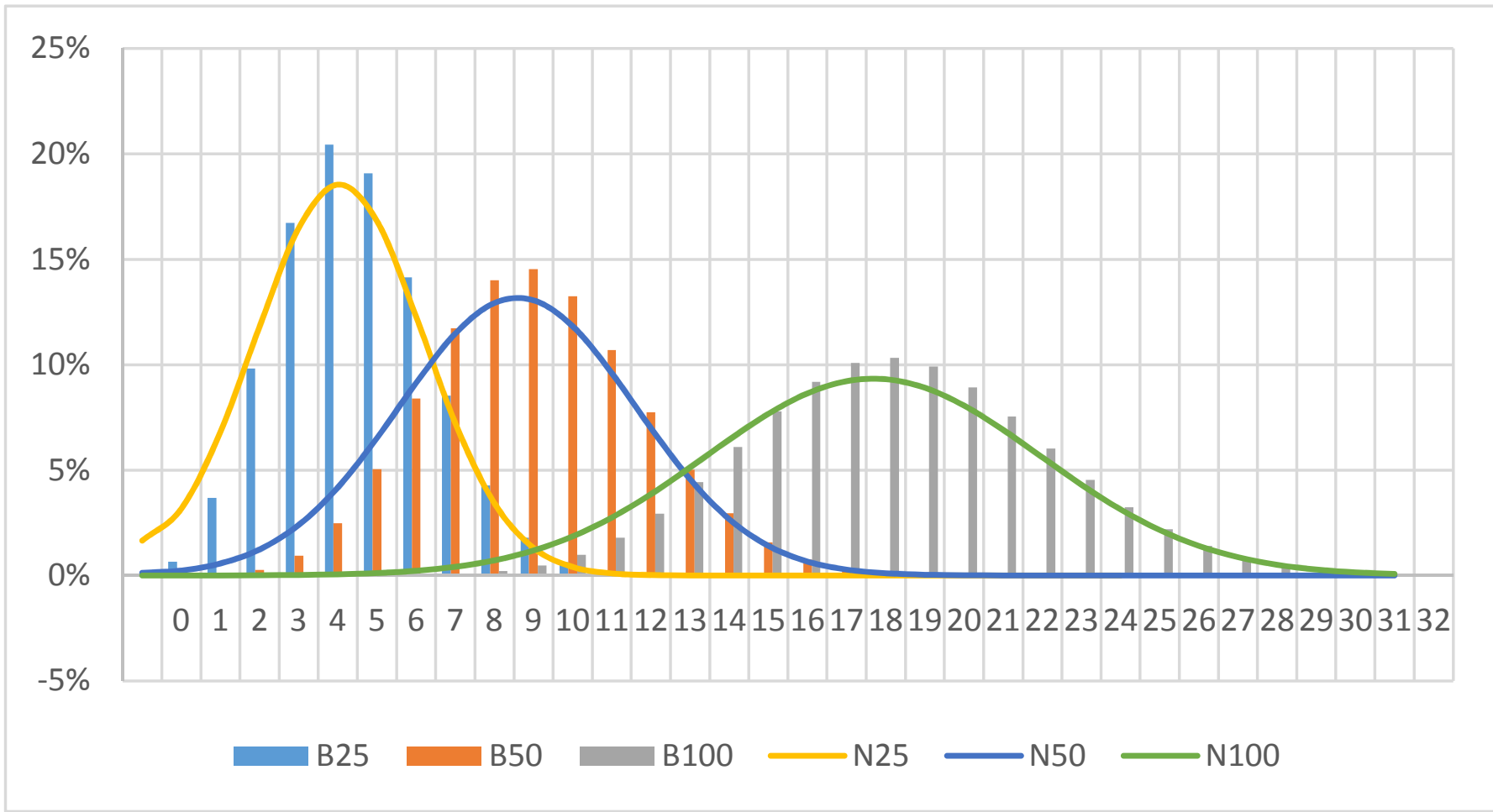
INTERVALLE DE FLUCTUATION

La proportion de boules rouges se situe 95% des fois dans l'intervalle

$$]p - 1,96\sqrt{\frac{p(1-p)}{n}}; p + 1,96\sqrt{\frac{p(1-p)}{n}}[$$

Mais toujours le problème d'approximer une loi discrète par une loi continue . . . alors qu'Excel donne instantanément la bonne valeur !

```
LOI.BINOMIALE.N(nombre_succès;tirages;
    probabilité_succès;cumulative)
```


25%
20%
15%
10%
5%
0%
-5%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
B25
B50
B100
N25
N50
N100

Calculs exacts

Intervalle de confiance à 95% pour la pelle de 25 alvéoles

- intervalle “seconde” :]-0,018 ;0.381[: ie]0 ;9,52[
- $1,96\sqrt{\frac{p(1-p)}{n}} = 0.1511923$
  intervalle “loi normale” :]0,030 ;0,333[ie]0,76 ;8,32[
- “vrai intervalle centré” : [2 ;8] à 93%, [2 ;9] à 95% ou [1 ;9] à 98%

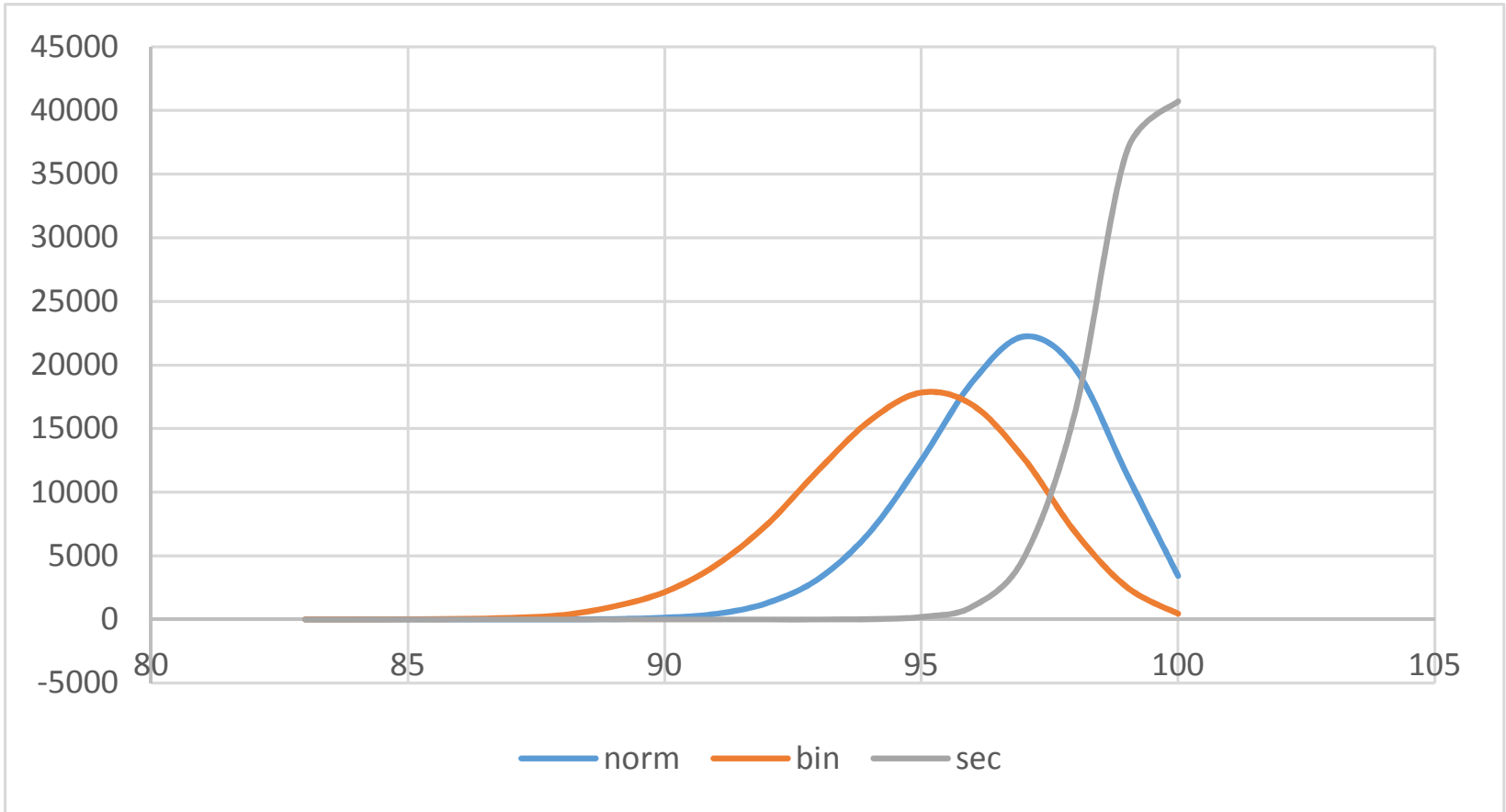
ESSAIS !

Allez on essaie pour de vrai

Simulations (pelle 25)

A l'échelle d'une classe : 100 essais. Simulation du tirage de 100 essai (sur R) répété 100000 fois. combien de fois est-on dans l'intervalle de fluctuation ?

- approximation normale : 96,6%
- approximation "seconde" : 99,1%
- calcul exact : 94,7%


45000
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
80
85
90
95
100
105
norm
bin
sec

Calculs exacts

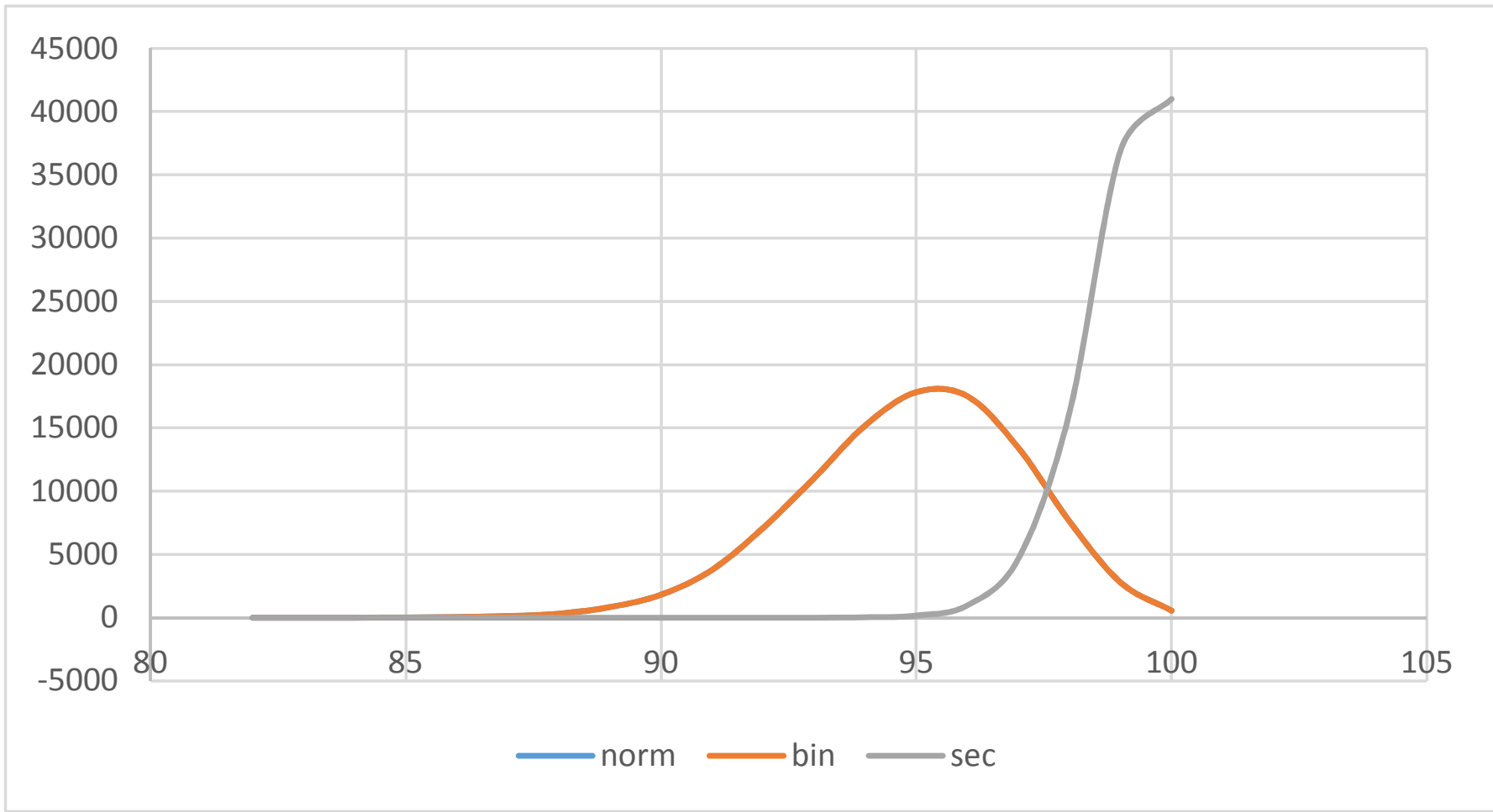
Intervalle de confiance à 95% pour la pelle de 100 alvéoles

- intervalle “seconde” :]-0,08 ;0.281[: ie]8,1 ;28,1[
- $1,96\sqrt{\frac{p(1-p)}{n}} = 0.075$
  intervalle “loi normale” :]0,106 ;0,257[ie]10,6 ;25,7[
- “vrai intervalle centré” : [11 ;25] à 95%

Simulations (pelle 100)

A l'échelle d'une classe : 100 essais. Simulation du tirage de 100 essai (sur R) répété 100000 fois. combien de fois est-on dans l'intervalle de fluctuation ?

- approximation normale : 94,89%
- approximation "seconde" : 99,1%
- calcul exact : 94,89%


45000
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
80
85
90
95
100
105
norm
bin
sec

Rappel du Sommaire

Tirage : on tire un échantillon de n individus dans une population de moyenne m et d'écart type σ pour le caractère étudié. On considère les tirages indépendants. La variable $\overline{X_n} = \frac{X_1+X_2+...+X_n}{n}$ associe à tout échantillon de taille n la moyenne de cet échantillon.

PROPOSITION

$\overline{X_n}$ prend pour valeurs les moyennes $\overline{x_1}, \overline{x_2}, \ldots, \overline{x_j}, \ldots$ de tous les échantillons de même effectif n, prélevés dans la population. D'après le théorème de la limite centrée, $\overline{X_n}$ suit approximativement la loi $\mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$.

Estimation de la moyenne : on considère une population de moyenne m et d'écart-type σ inconnus. On prélève un échantillon de taille n de moyenne $\overline{x}$ et d'écart-type σ'. On suppose n suffisamment grand.

- On estime ponctuellement m par $\overline{x}$.
- On estime ponctuellement σ par $\sigma' \sqrt{\frac{n}{n-1}}$.
  Ce facteur permet d'éviter de sous-estimer la dispersion.
- Pour une proportion, on estime ponctuellement p par la fréquence f dans l'échantillon (cas particulier d'une moyenne).

L'intervalle

$$\left]\overline{X_n} - z\frac{\sigma}{\sqrt{n}}; \overline{X_n} + z\frac{\sigma}{\sqrt{n}}\right[$$

est appelé "intervalle de confiance pour m" à $1 - \alpha$.

Quelques valeurs de z :

- $\alpha = 0,1 \Rightarrow 1 - \alpha = 0,90 \Rightarrow z = 1,65$
- $\alpha = 0,05 \Rightarrow 1 - \alpha = 0,95 \Rightarrow z = 1,96$
- $\alpha = 0,01 \Rightarrow 1 - \alpha = 0,99 \Rightarrow z = 2,58$

Remarques

- l'intervalle donné ne contient m que dans une proportion $1 - \alpha$ des cas. Il est donc possible que m ne soit pas dans cet intervalle. Par exemple pour $\alpha = 0,05$ m n'appartient pas à l'intervalle de confiance une fois sur 20.
- si m appartient à cet intervalle, on ne peut pas dire si m est plutôt au centre de l'intervalle ou près de l'une ou l'autre des extrémités.
- Pour pouvoir calculer cet intervalle, il faut être dans les conditions d'application du théorème de la limite centrée. Plus précisément, on doit être dans un des cas suivants :
  1. Population normale et écart-type connu.
  2. Population quelconque avec $n \geq 30$ et on estime σ si on ne le connaît pas.

Estimation d'une proportion

Soit une population contenant une proportion inconnue p d'éléments possédant une certaine propriété. On se place dans le cas où la variable F qui à chaque échantillon de taille n fixée associe la proportion des éléments de cet échantillon possédant la propriété considérée, suit une loi que l'on peut approximer par la loi normale $N(p; \sqrt{\frac{p(1-p)}{n}})$. Soit un échantillon de taille n contenant une proportion f d'éléments possédant la propriété considérée (on suppose $n \geq 30$). L'intervalle $[f - z\sqrt{\frac{f(1-f)}{n}}; f + z\sqrt{\frac{f(1-f)}{n}}]$ est l'intervalle de confiance de la proportion inconnue p de la population avec le coefficient de confiance α tel que $\alpha = P(-z \leq Z \leq z)$.

ESSAIS !

Allez on essaie pour de vrai, avec des M&M's !

- à partir des moyennes
- à partir des proportions