Echantillonnage, estimation

Michel PUYOU
Pierre-Henri TERRACHER
(IREM d'Aquitaine)

« *La première question est de construire un schéma mathématique présentant avec la réalité d'assez étroits rapports* »

Emile Borel, Le Hasard (1914)

Quelques notions à préciser

1° Expérience aléatoire

2° Modèle probabiliste

3° Modélisation

4° Simulation

1. Expérience aléatoire

Morceaux choisis :

« Une expérience aléatoire est une expérience dont le résultat est incertain »

« Une expérience est aléatoire lorsqu'on ne peut pas en prévoir avec exactitude le résultat »

L’avis d’un spécialiste :

« Une expérience aléatoire est définie si l’on est en mesure de préciser :

- *les conditions de l’expérience*
- *l’ensemble des résultats possibles »*

Albert JACQUARD (Les probabilités)

Dans l’enseignement secondaire

expérience aléatoire
- protocole expérimental
- issues possibles identifiées
- intervention du hasard

2. Modèle probabiliste

Les modèles « théoriques »

ou, mieux dit : « les lois de probabilité »

Lois discrètes : équirépartie, de Bernoulli, binômiale, etc…

Lois continues : exponentielle, normale,…

Lois « transferts » : par variables aléatoires à partir de la loi équirépartie

« Modèles» de l’urne ?

ou modèle « pseudo concret »

Situations génériques décontextualisées

- Références pour les lois discrètes usuelles
- Outil de travail (échantillonnage …)

3. Modélisation

Modéliser une expérience aléatoire c’est associer à cette expérience une loi de probabilité Sur l’ensemble des issues possibles.

Et c’est tout !

Sur le choix du modèle

1° Considérations propres au protocole expérimental
exemple : équirépartition

2° A partir de données expérimentales
délicat (hors programme secondaire)

3° Des « complications » célèbres
- les paradoxes qui n'en sont pas (la corde de Bertrand, le problème des bancs…)
- la loi de Benford.
- « le » modèle introuvable.

M

Ni vrai ni faux

Un modèle n’est ni vrai ni faux.

Il peut être validé ou rejeté

La statistique inférentielle

- compatibilité avec les données expérimentale
- procédures de validation

(Justification:la loi des grands nombres)

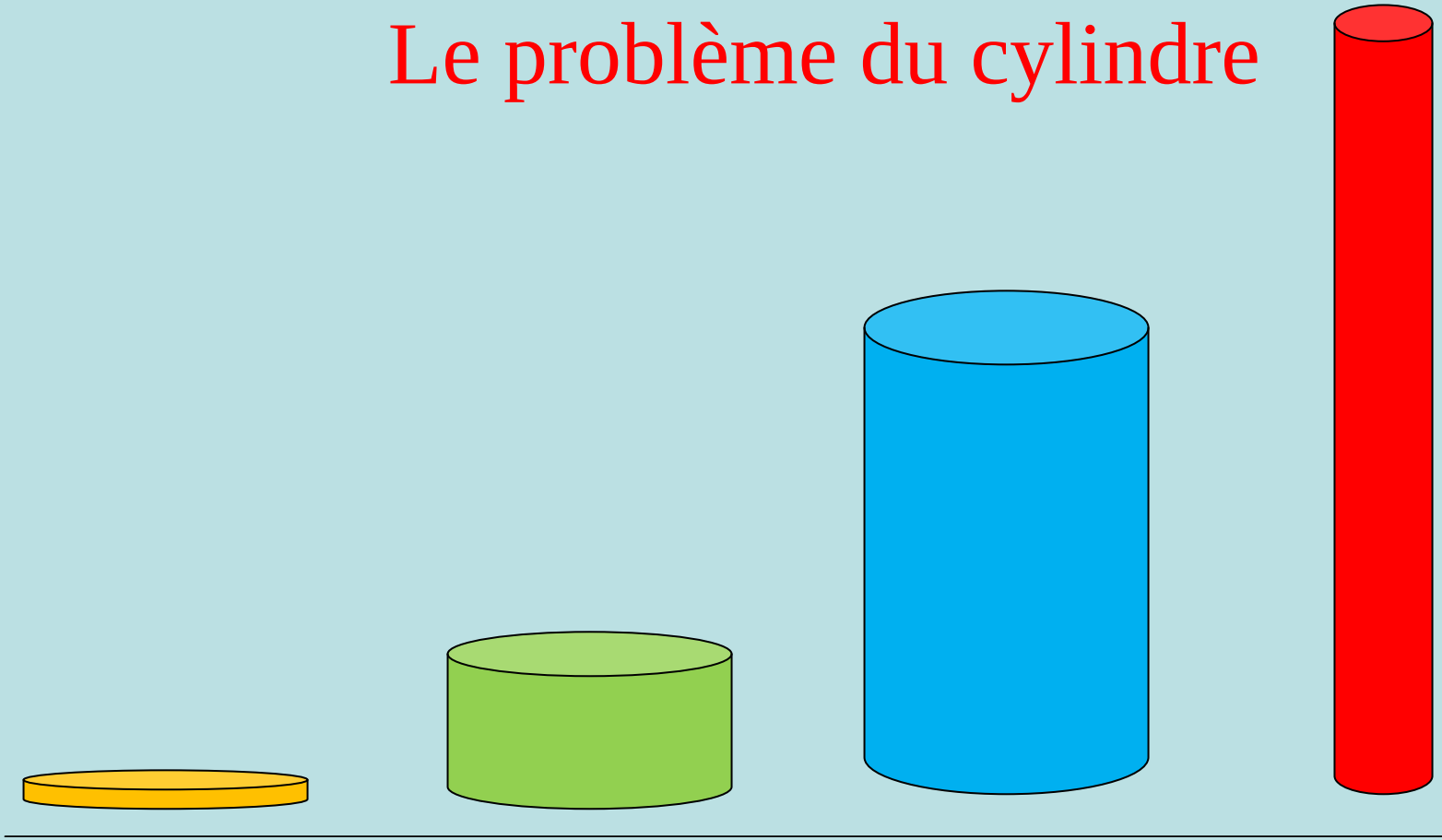
4. La simulation

Extrait manuel scolaire

«Lorsqu'on souhaite répéter une expérience aléatoire un grand nombre de fois (5000 fois le lancer d'un dé à 6 faces, par exemple), on peut faire soi-même l'expérience avec un dé (mais c'est long et fastidieux) ou on peut utiliser un simulateur(la calculatrice ou un tableur, par exemple). Ainsi, la simulation remplace l'expérience et permet d'étudier des séries statistiques comportant un grand nombre de données. »

Un cas édifiant s’il en est.

Le problème du cylindre



L’intérêt d’une définition n’est donc plus à deviner…

« Simuler une expérience, c’est choisir un modèle de cette expérience, puis simuler ce modèle »

Ainsi, sans modèle théorique, il ne saurait être question de simulation.

Simulation et conjecture

Exemple 1

On tire au hasard et indépendamment les uns des autres six réels entre 0 et 1.

Combien y a-t-il en moyenne d'intervalles parmi les six intervalles $I_k = \left[\frac{k}{6}, \frac{k+1}{6}\right[$ $(0 \leq k \leq 5)$ qui contiennent au moins l'un de ces réels ?

Simulation : environs 3,99 (sur 10000expériences)

Conjecture élève : 4 intervalles

Le calcul (facile) : $6\left(1-\left(\frac{5}{6}\right)^6\right)$ (environs 3,9906°)

M

Exemple 2

On tire au hasard des réels entre 0 et1 indépendamment les uns des autres jusqu'à ce que leur somme dépasse 1. Combien doit-on « tirer » de réels en moyenne ?

La simulation

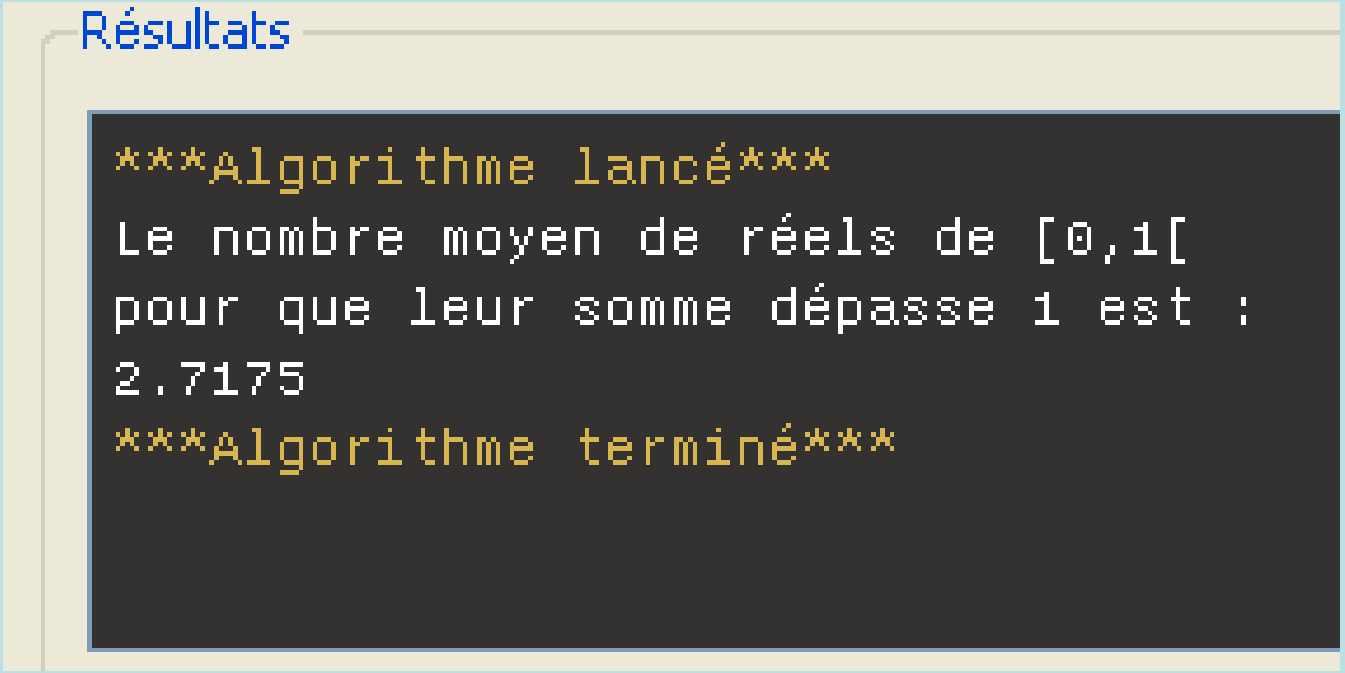
CODE DE L'ALGORITHME :

```
1     VARIABLES
2       S EST_DU_TYPE NOMBRE
3       N EST_DU_TYPE NOMBRE
4       M EST_DU_TYPE NOMBRE
5       C EST_DU_TYPE NOMBRE
6     DEBUT_ALGORITHME
7       M PREND_LA_VALEUR 0
8       POUR C ALLANT_DE 1 A 10000
9         DEBUT_POUR
10        N PREND_LA_VALEUR 0
11        S PREND_LA_VALEUR 0
12        TANT_QUE (S<1) FAIRE
13          DEBUT_TANT_QUE
14          S PREND_LA_VALEUR S+random()
15          N PREND_LA_VALEUR N+1
16          FIN_TANT_QUE
17        M PREND_LA_VALEUR M+N
18        FIN_POUR
19      M PREND_LA_VALEUR M/10000
20      AFFICHER "Le nombre moyen de réels de [0,1[ pour que leur somme dépasse 1 est : "
21      AFFICHER M
22    FIN_ALGORITHME
```

Cet algorithme réalisé avec Algobox
« simule » le problème de Halmos .

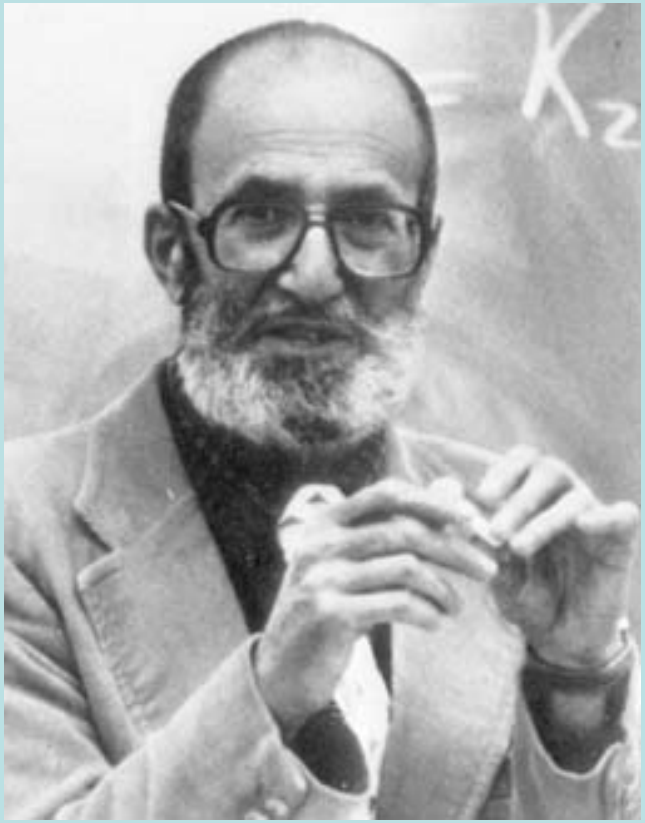
Il fournit la moyenne du nombre de tirages aléatoires dans [0 ;1] pour obtenir une somme supérieure ou égale à 1

Voici ce que ça donne :



Conjecture ?

Démonstration (Halmos)



Paul Halmos
Mathématicien américain (1916-2006)
« Les mathématiciens sont des artistes,
non pas des calculateurs »

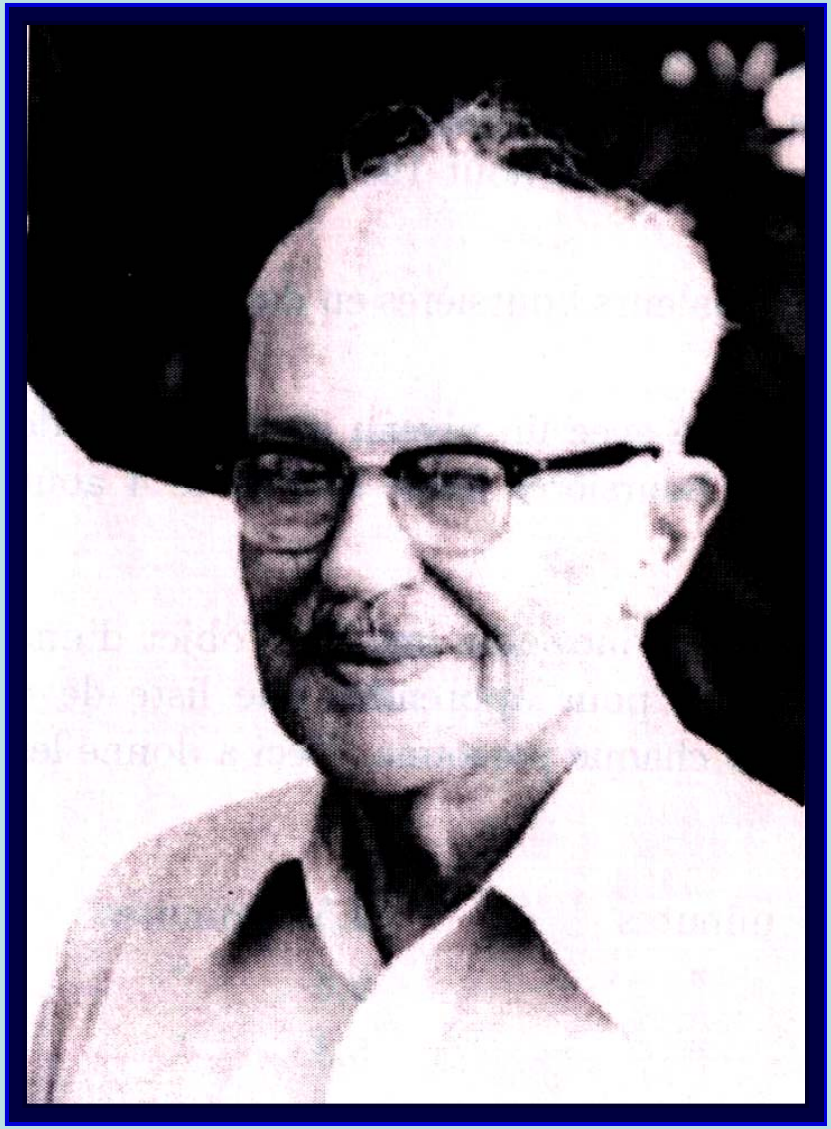
Échantillonnage

Estimation

(pour se faire une idée…)

« Quel danger, quelle folie de choisir sur des échantillons »

Nathalie Sarraute
Le Planétarium (1959)



Jerzy NEYMAN
1894 - 1981

1 - Introduction

1-1 Situation prototype

$\mathcal{P}$: population de boules blanches et noires dans une urne

p : proportion de blanches, *inconnue*

ε : échantillon de n boules extraites de $\mathcal{P}$

f : fréquence de blanches dans ε

Peut-on “déterminer” p à partir de f ?

1-2 Une première idée (naïve)

« Prendre $p = f$ »

- Illusoire
  - fluctuation d'échantillonnage
  - taille de l'échantillon

- Conséquences
  - renoncer à "déterminer" p
  - tenter plutôt d'estimer p

Signification ?

pour l'instant… AUCUNE

1-3 Une deuxième idée (plus élaborée)

Supposons que nous possédions des renseignements sur la distribution des fréquences de TOUS les échantillons aléatoires de taille n issus de $\mathcal{P}$, c'est-à-dire sur la **variable aléatoire** F

$$F : \mathfrak{E}_n \to [0,1]$$
$$\varepsilon \mapsto f_\varepsilon$$

$\mathfrak{E}_n$: ensemble de tous les échantillons de taille n issus de $\mathcal{P}$.

f_ε : fréquence de boules blanches dans l'échantillon ε.

Exemple

$n = 2000$ on suppose que “l’on sait que” :

$$\mathbb{P}\left(|F - p| < 0,025\right) = 0,95$$

Interprétation

L’intervalle aléatoire $[F - 0,025, F + 0,025]$ a 95% de chances de contenir p

ou encore :

Pour 95% des échantillons ε de taille $n = 2000$, l’intervalle $[f_\varepsilon - 0,025,\ f_\varepsilon + 0,025]$ contient p.

Pour la valeur observée f :

$[f - 0,025,\ f + 0,025]$ « fourchette » pour p au niveau de confiance $0,95$

1-4 D'où le problème

Renseignements sur *F* ?

- Réponse 1 : Examiner tous les échantillons(!)
- Réponse 2 : Tenter de VOIR si la distribution de *F* ne serait pas VOISINE d'une distribution CONNUE

VOIR ? Étude de la fréquence d'échantillon

VOISINE ? Théorème d'approximation

CONNUE ? Loi normale

1-5 Plan

- Loi normale
- Théorème d'approximation
- Fréquence d'échantillonnage
  Intervalle de confiance
- Quelques remarques pour conclure

2- Loi normale

Carl Friedrich Gauss
1777 - 1855

2-1 Notions de base : $\mathcal{N}(\mu,\sigma)$

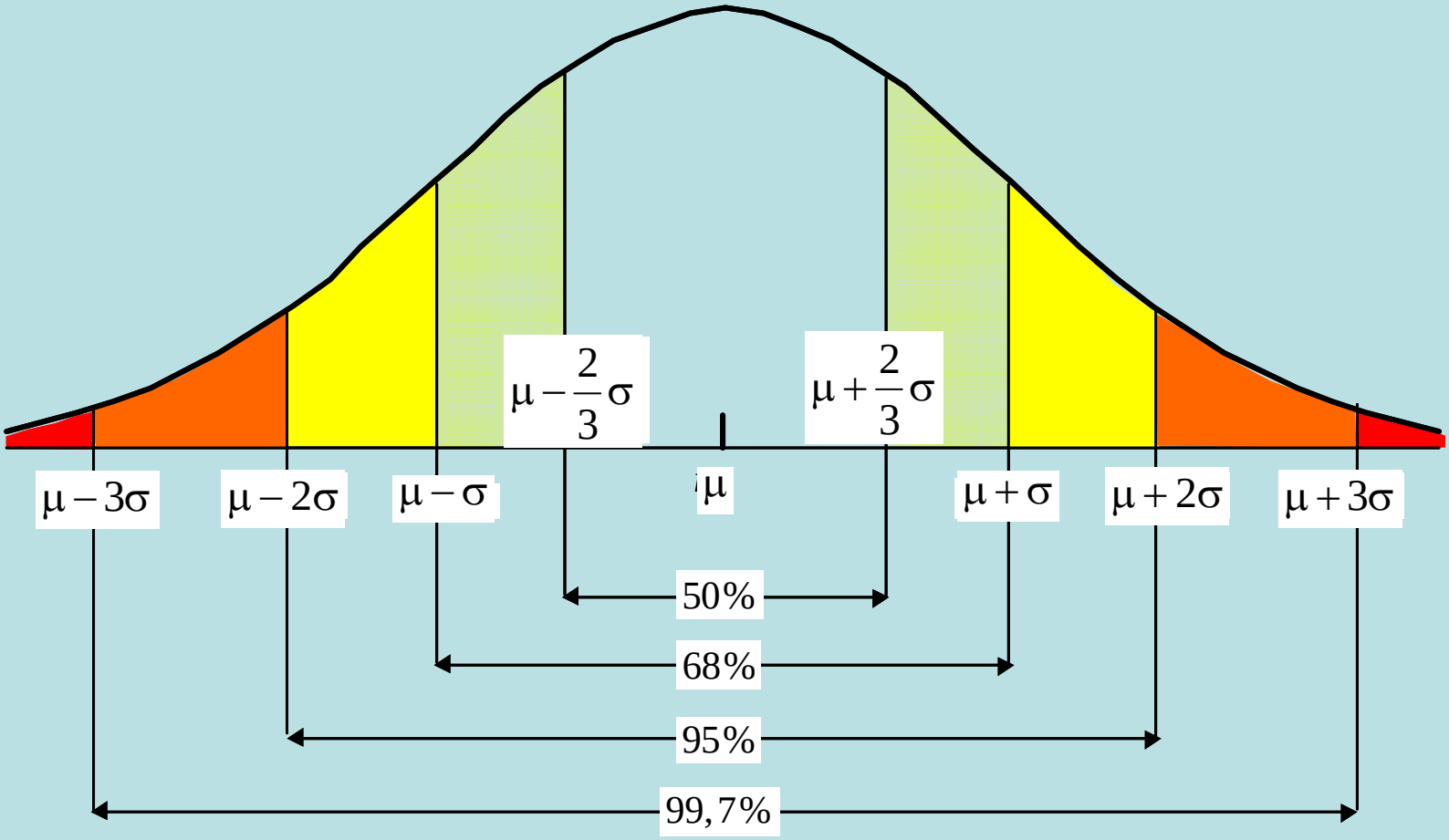
- Densité
  - $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - f continue, positive sur $\mathbb{R}$

    $\int_{\mathbb{R}} f(x)\,\mathrm{d}x=1$

    $\left(\int_0^{+\infty} e^{-x^2}\,\mathrm{d}x=\dfrac{\sqrt{\pi}}{2}\right)$

- Moyenne : μ
- Écart type : σ

Dispersion autour de la moyenne



2-2 La loi normale centrée réduite $\mathcal{N}(0,1)$

- La fonction de répartition : $\pi(t) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}s^2} \, ds$

- Essentiel

  X suit la loi normale $\mathcal{N}(\mu, \sigma)$

  $$\Leftrightarrow T = \frac{X - \mu}{\sigma} \text{ suit } \mathcal{N}(0,1)$$

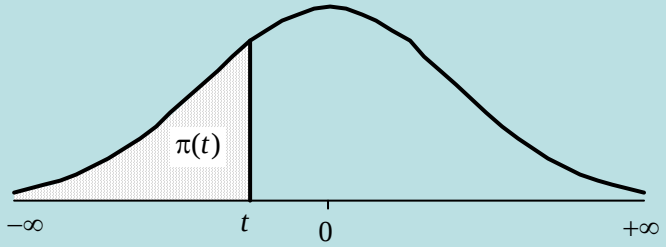
  ($E(T) = 0$; $\sigma(T) = 1$: centrée réduite associée à X)

- Les tables de la loi $\mathcal{N}(0,1)$

  Valeurs de π (fonction de répartition) pour $t \geq 0$.

Fonction de répartition de la loi de Laplace-Gauss

$$\pi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{t^2}{2}} dt$$



t	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de *t*.

t	3,0	3,1	3,2	3,3	3,4	3,5
$\pi(t)$	0,998650	0,999032	0,999313	0,999517	0,999663	0,999767

t	3,6	3,7	3,8	3,9	4,0	4,5
$\pi(t)$	0,999841	0,999892	0,999928	0,999952	0,999968	0,999997

- Propriétés

T suit $\mathcal{N}(0,1)$

$\pi(t) = \mathbb{P}(T \leq t)$

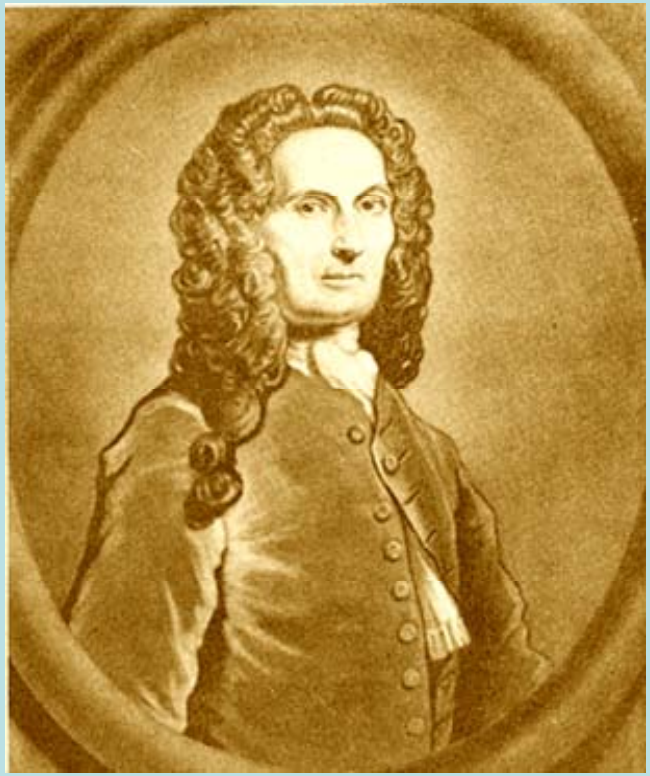
- $\mathbb{P}(t_1 \leq T \leq t_2) = \pi(t_2) - \pi(t_1)$
- $\mathbb{P}(T > t) = 1 - \pi(t)$
- $\mathbb{P}(T < -t) = 1 - \pi(t)$
- $\mathbb{P}(|T| < t) = 2\pi(t) - 1 \qquad (t \geq 0)$

(lecture graphique)

$$\mathbb{P}(|T| \leq 1{,}96) \approx 0{,}95$$

$$\mathbb{P}(|T| \leq 2{,}58) \approx 0{,}99$$

3- Le théorème central limite



Abraham de Moivre
1667 - 1754

3-1 Approximation d'une loi binomiale par une loi normale

- Un premier énoncé
  - S_n variable aléatoire suit la loi binomiale $\mathsf{B}(n, p)$
  - $T_n = \dfrac{S_n - np}{\sqrt{npq}}$ suit la loi $\mathcal{N}(0,1)$

$$\mathbb{P}(a < T_n < b) \approx \pi(b) - \pi(a)$$

(sous certaines conditions) (voir plus loin)

- Variantes

Sous les mêmes conditions :

- S_n suit approximativement la loi normale $\mathcal{N}\left(np, \sqrt{npq}\right)$
- $\dfrac{1}{n} S_n$ suit approximativement la loi normale $\mathcal{N}\left(p, \sqrt{\dfrac{pq}{n}}\right)$

- Les conditions d'approximation

1 En général

$n \geq 30$, $np \geq 5$, $nq \geq 5$

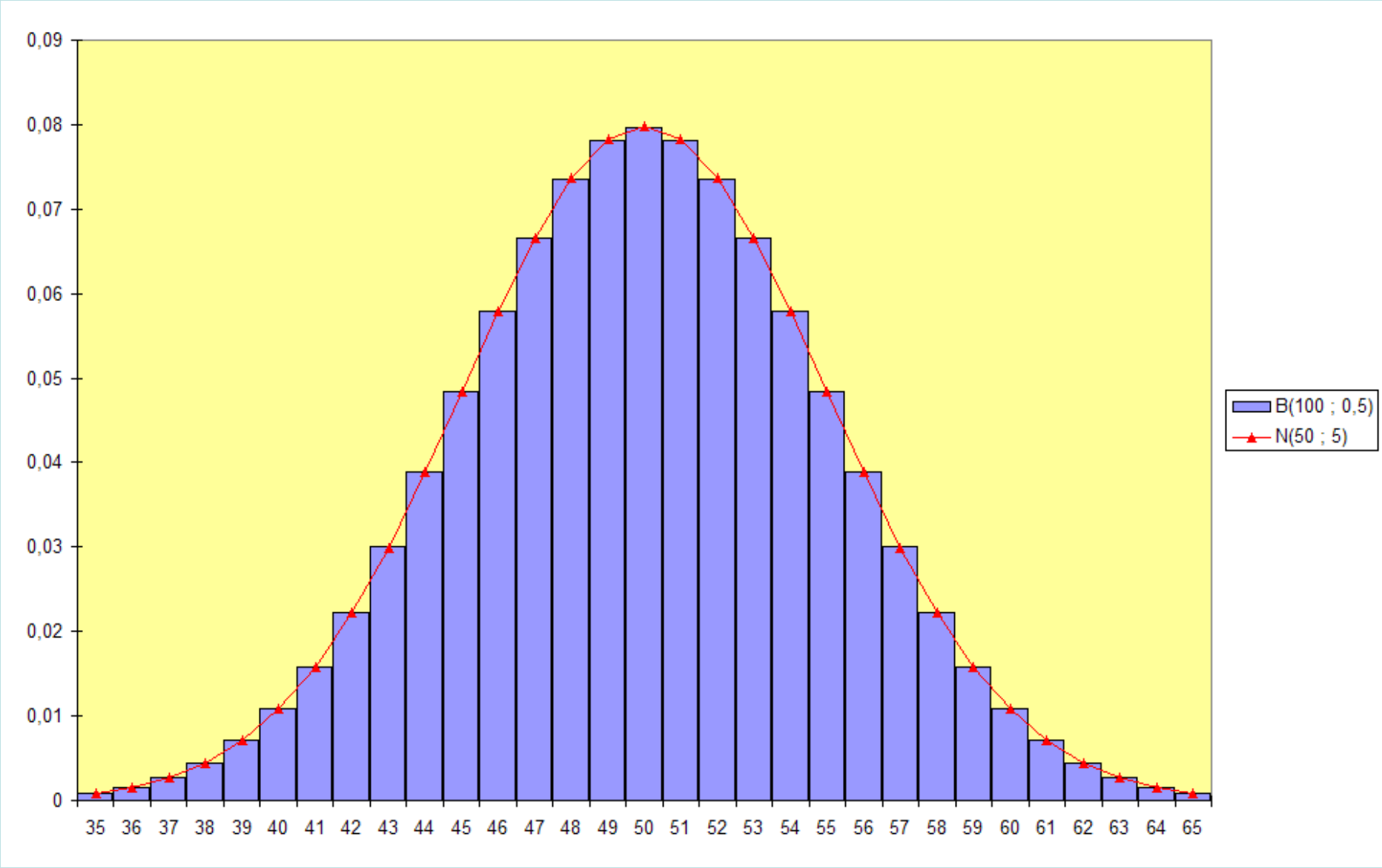
2 Très bonne

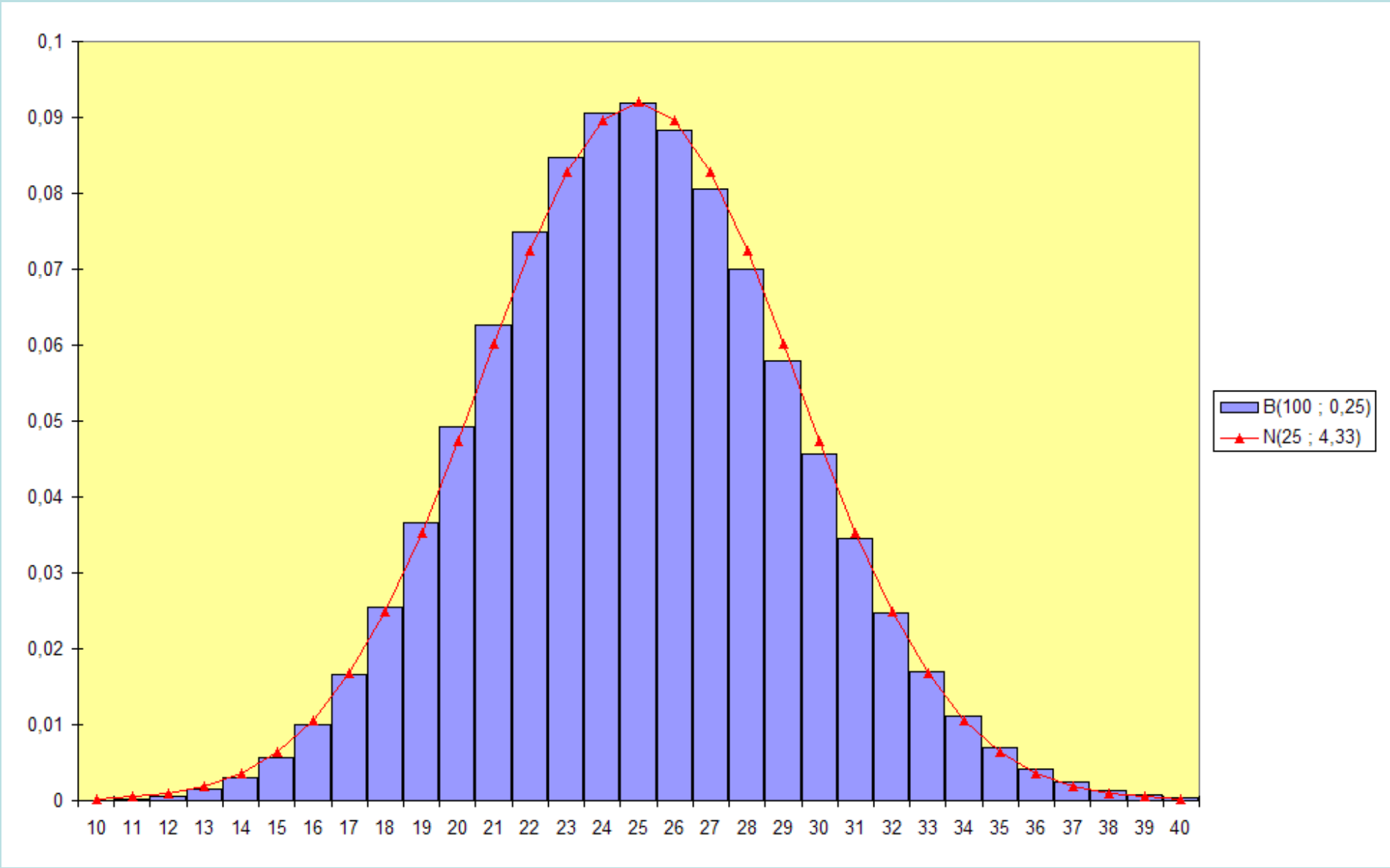
$n \geq 30$ $0,2 \leq p \leq 0,8$

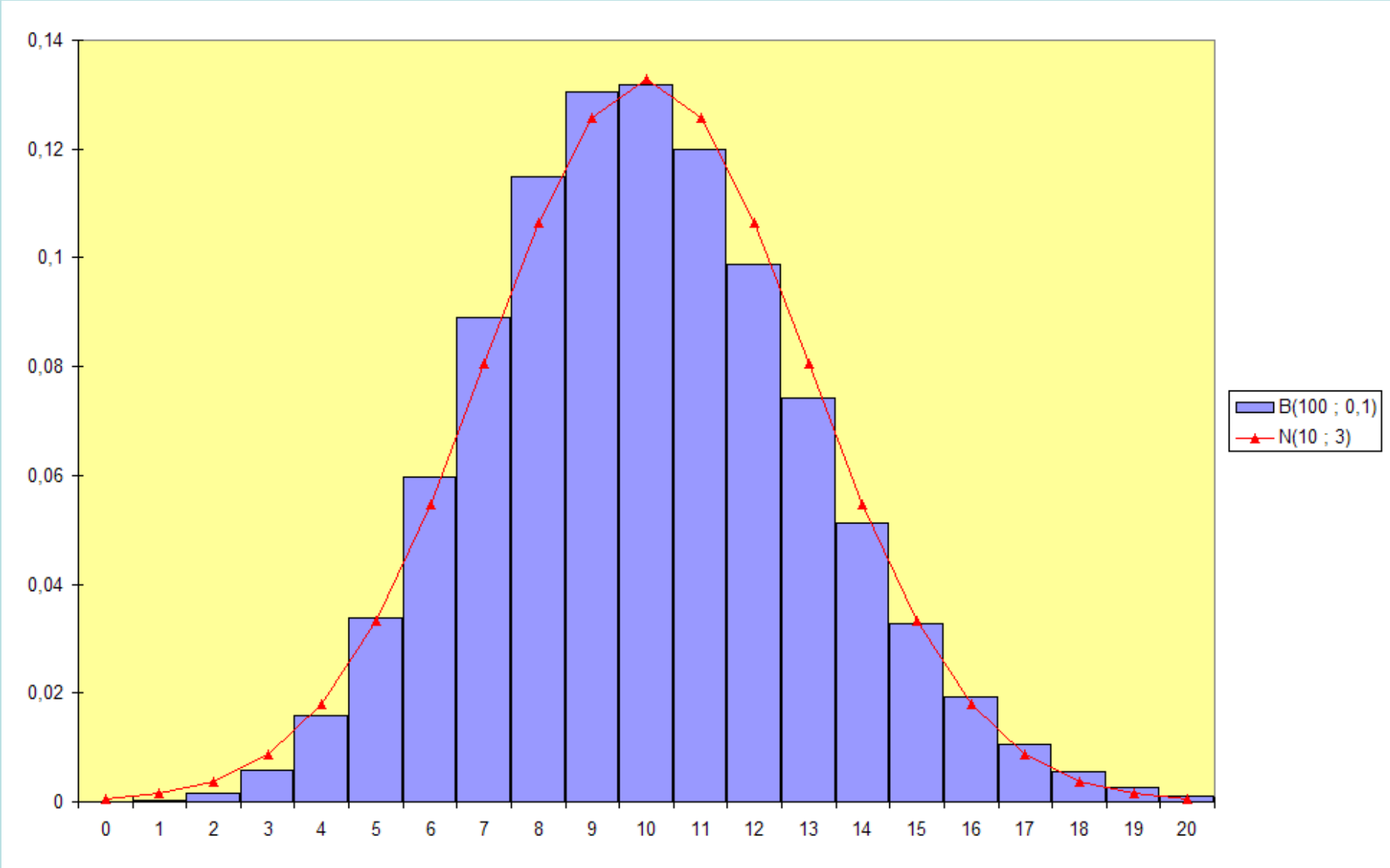
- Le point de vue graphique

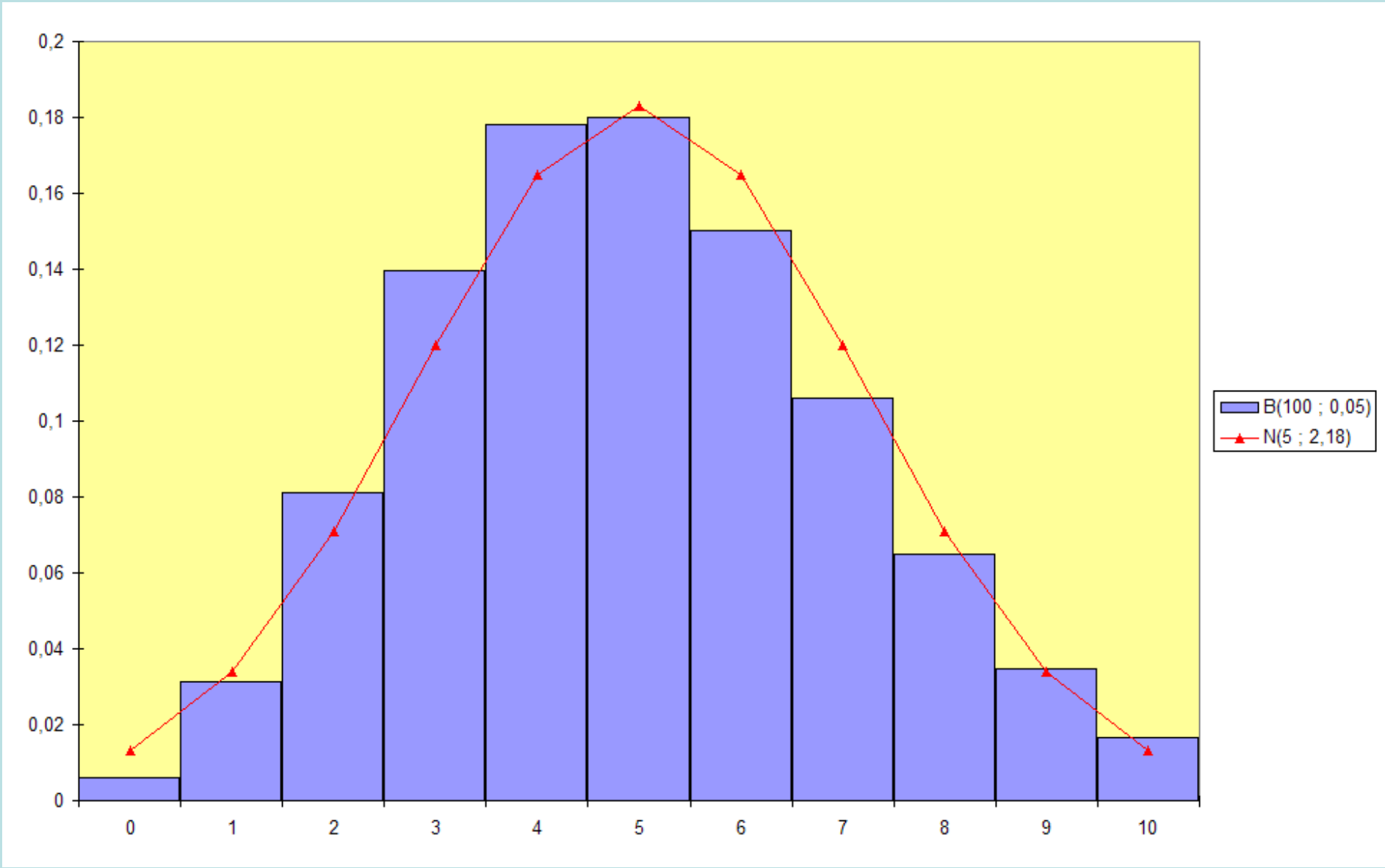
$$n = 100 \begin{cases} \bullet\ p = 0,5 \\ \bullet\ p = 0,25 \\ \bullet\ p = 0,10 \\ \bullet\ p = 0,05 \end{cases}$$

M


0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65
B(100 ; 0,5)
N(50 ; 5)


0,1
0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
B(100 ; 0,25)
N(25 ; 4,33)


0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
B(100 ; 0,1)
N(10 ; 3)


0,2
0,18
0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
0
1
2
3
4
5
6
7
8
9
10
B(100 ; 0,05)
N(5 ; 2,18)

3-2 Deux exemples

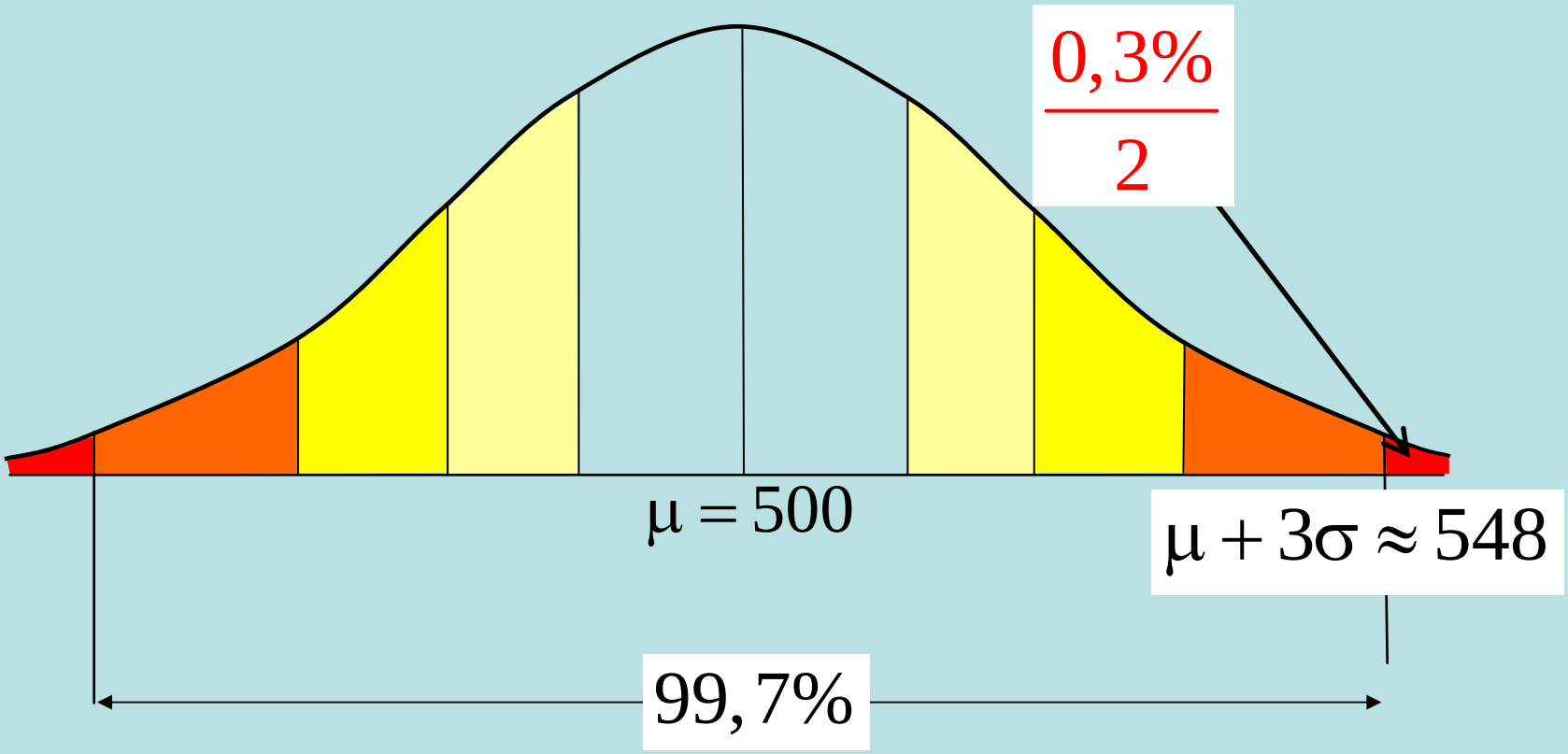
- Mille « PILE ou FACE »
  - 1000 lancers d'une pièce de monnaie déclarée "honnête"
  - probabilité d'obtenir 548 PILES ou plus

① Valeur exacte

$$p = \sum_{k=548}^{1000} \binom{1000}{k} \times \left(\frac{1}{2}\right)^{1000}$$

Calculatrice 0,00132

② Calcul de tête (!) Un peu moins de 0,0015


$\frac{0,3\%}{2}$
$\mu = 500$
$\mu + 3\sigma \approx 548$
$99,7\%$

③ Par approximation

- Conditions d'approximations : OK
- Calculs des paramètres $np = 500, \quad \sqrt{npq} = 5\sqrt{10}$
- Approximation

X suit $\mathsf{B}\left(1000, \frac{1}{2}\right)$

$T = \frac{X - 500}{5\sqrt{10}}$ suit approx^{t} $\mathcal{N}(0,1)$

$548 \leq X \leq 1000 \Leftrightarrow 3,03579 \leq T \leq 31,62$

$\pi(31,62) \approx 1$ (normal)

$\pi(3,03579) \approx 0,998664$ (interpolation)

Résultat : $\boxed{0,001336}$

Le Q. C. M.

- Q. C. M. : 100 questions
  3 réponses possibles (une seule correcte)
- Un élève répond au hasard à toutes les questions
- Trouver k entier tel que : **l'élève ait moins de 5% de chance d'avoir au moins *k* réponses justes**

Interprétation

X variable aléatoire = « nombre de réponses exactes "tout au hasard" »

Trouver k tel que $\begin{cases} P(X \geq k) < 0,05 \\ \text{ou} \\ P(X < k) \geq 0,95 \end{cases}$ X suit $\mathrm{B}\left(100, \frac{1}{3}\right)$

- Approximation

  1- Conditions : OK

  2- Paramètres $\mu = np = \frac{100}{3}$, $\sigma = \sqrt{npq} = \frac{10}{3}\sqrt{2}$

P

3- Approximation

$$T = \frac{X - \mu}{\sigma} \text{ suit approximativement } \mathcal{N}(0,1)$$

$$X < k \Leftrightarrow T < \frac{k - \mu}{\sigma} = t_0$$

$$\pi(t_0) = P(T < t_0) \geq 0,95 = \pi(1,645) \quad \text{(table, calculatrice)}$$

signifie $t_0 \geq 1,645$ (croisssance)

Soit : $k \geq \sigma \times 1,645 + \mu$ i. e. $k \geq \left(\frac{10}{3}\sqrt{2}\right) \times 1,645 + \frac{100}{3}$

$k \geq 41,087$

$k = 42$

3-3 “Fourchette” d’une loi binomiale

Théorème

- S_n var aléatoire suit $\mathsf{B}(n,p)$
- $F_n = \dfrac{S_n}{n}$

Sous certaines conditions $\mathbb{P}\left(|F_n - p| < \dfrac{1}{\sqrt{n}}\right) \geq 0{,}95$

$T_n = \dfrac{S_n - np}{\sqrt{npq}}$ suit $\approx \mathcal{N}(0,1)$ $P(|T_n| \leq 1{,}96) \approx 0{,}95$

$|T_n| \leq 1{,}96 \Leftrightarrow |F_n - p| \leq \dfrac{1}{\sqrt{n}} \times 1{,}96 \times \sqrt{pq}$

or $1{,}96 \times \sqrt{p(1-p)} \leq 1{,}96 \times \sqrt{\dfrac{1}{4}} \leq 1$

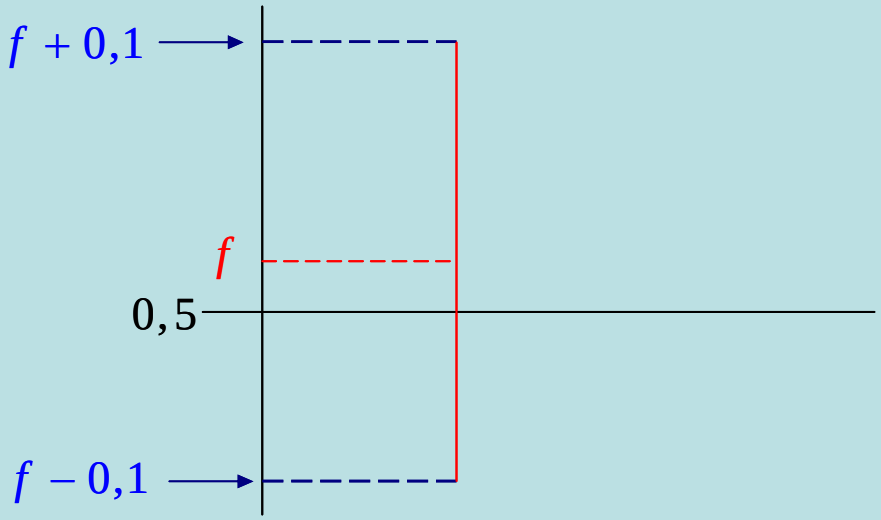
D’où l’inclusion $(|T_n| \leq 1{,}96) \subset \left(|F_n - p| \leq \dfrac{1}{\sqrt{n}}\right)$

M

- Illustration graphique

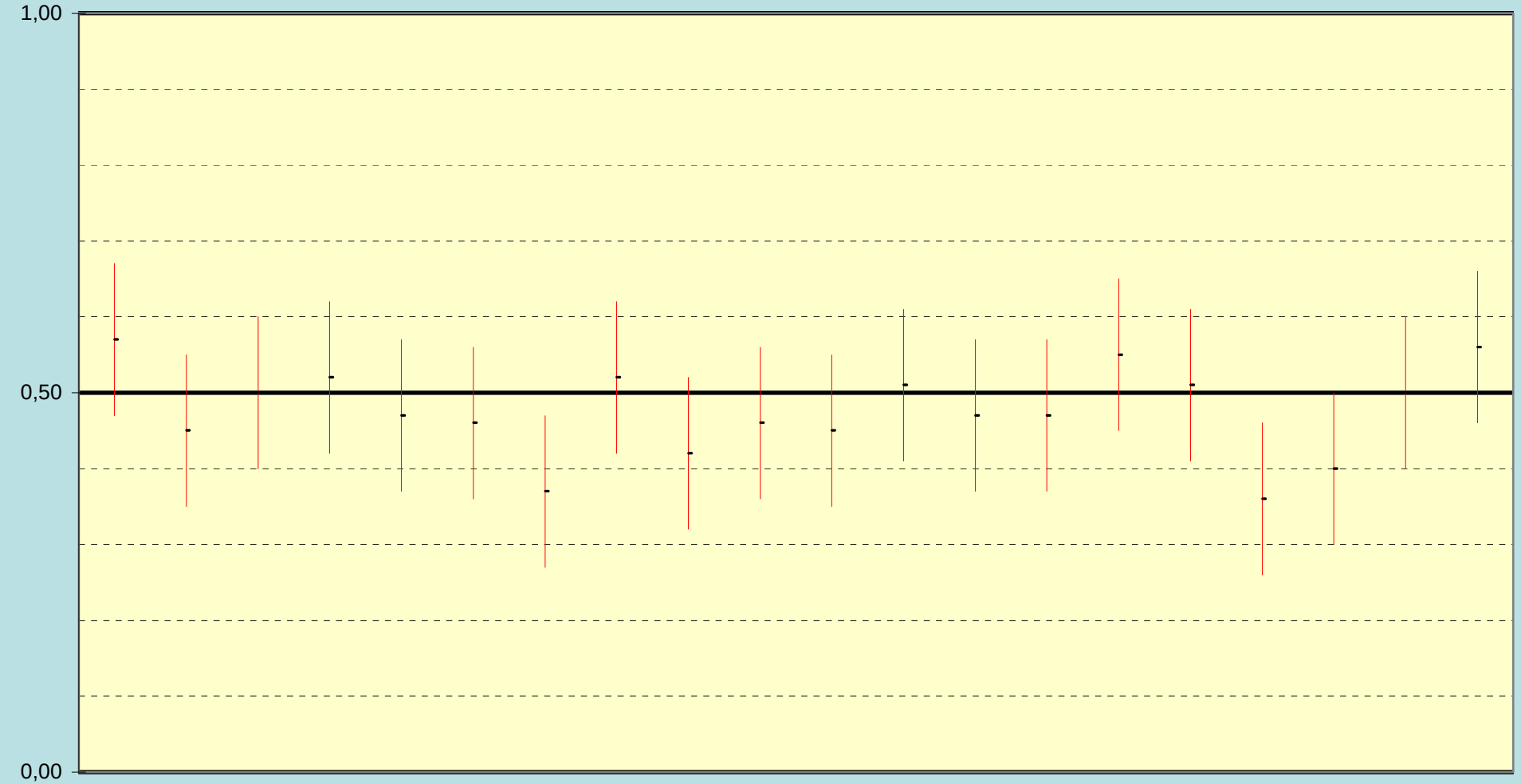
$$\begin{cases} n = 100 \\ p = 0,5 \end{cases}$$

simulation $\rightarrow$ f fréquence observée

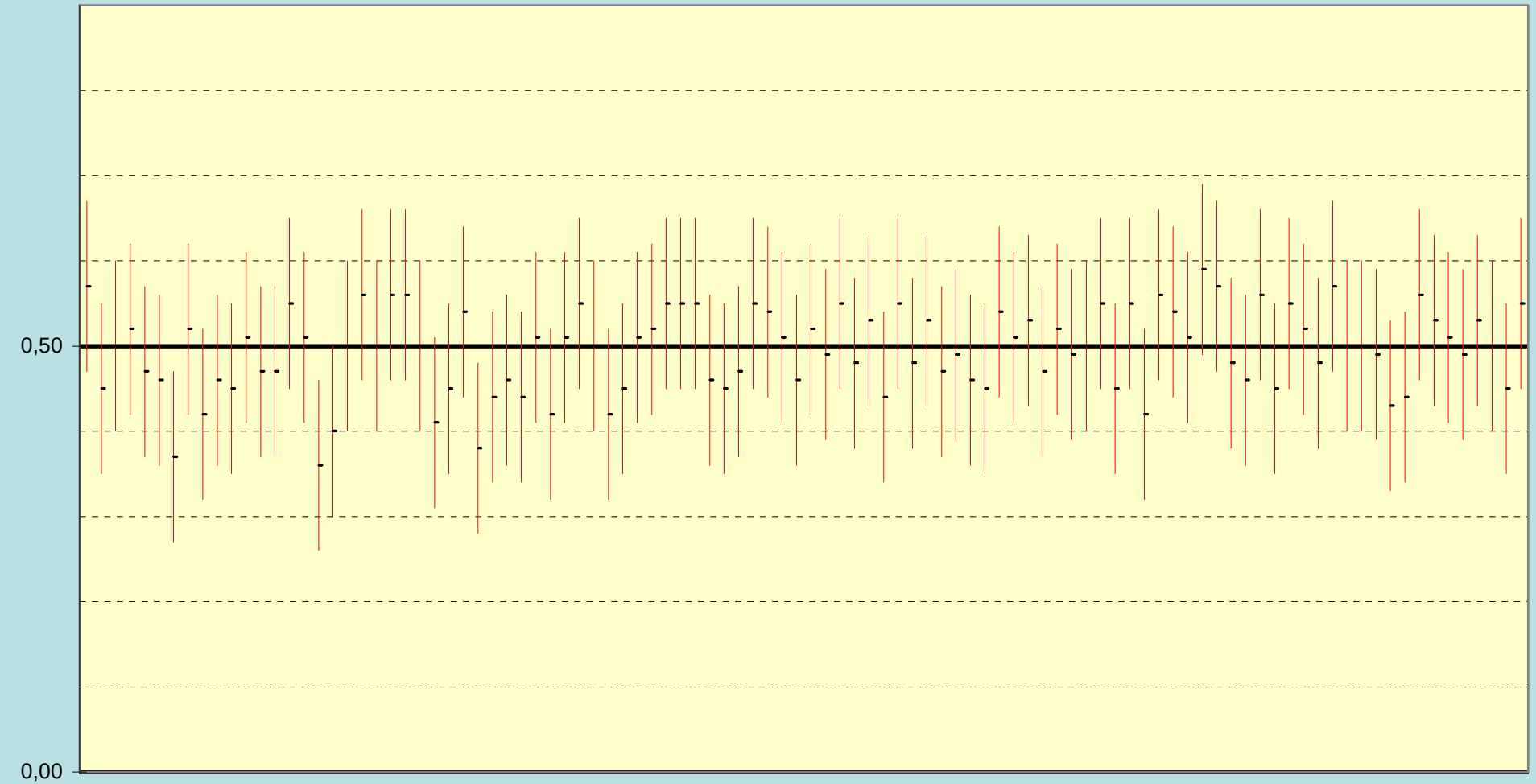


- 20 simulations
- 100 simulations
- 1000 simulations

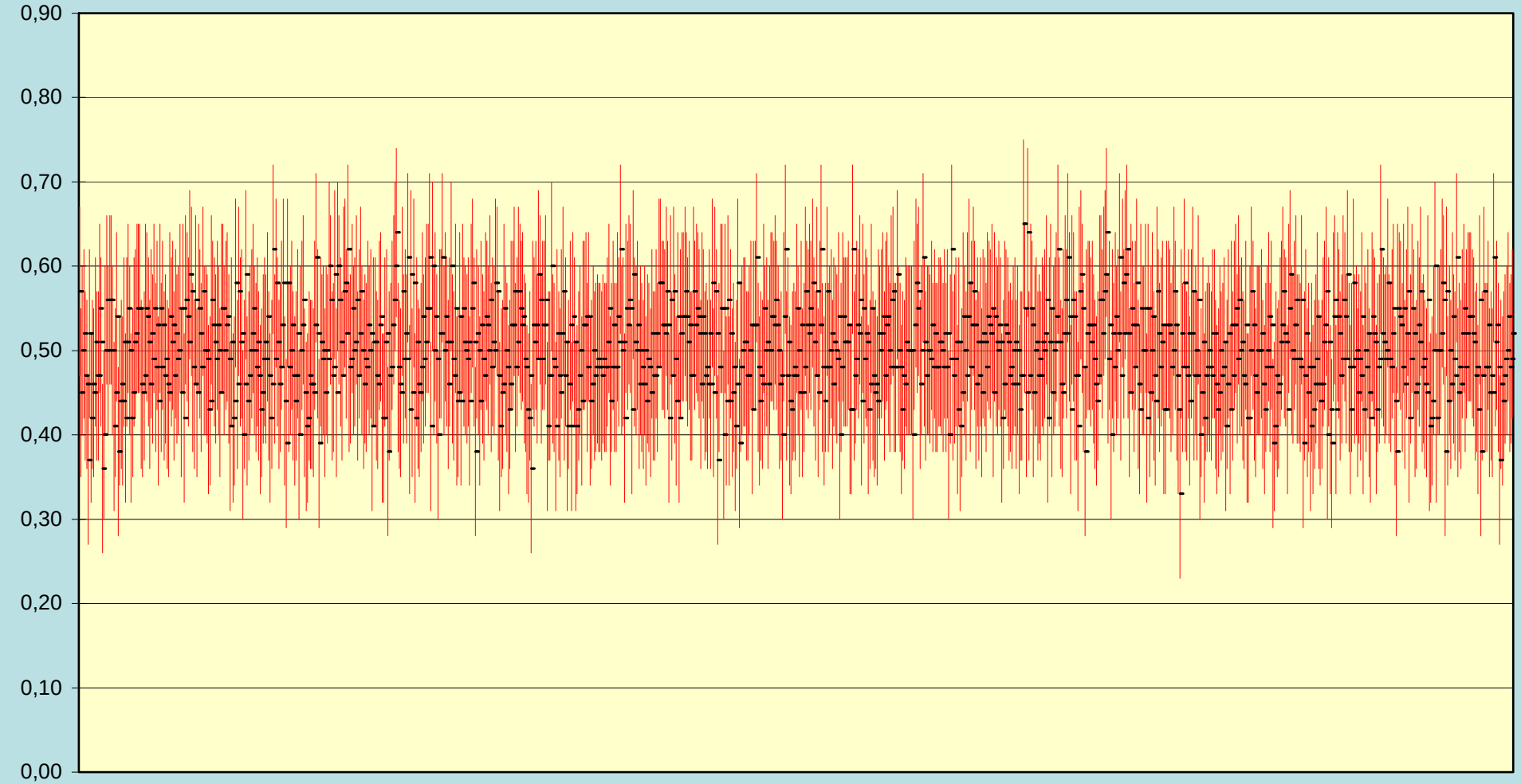
20 séries de 100 lancers



100 séries de 100 lancers



1000 séries de 100 lancers



Remarque

Avec le même raisonnement on a

$$\mathbb{P}\left(|F_n - p| < \frac{1,3}{\sqrt{n}} \right) \geq 0,99$$

Explications

- $\mathbb{P}\left(|T_n| \leq 2,58\right) \approx 0,99$
- $2,58 \times \sqrt{p(1-p)} \leq \frac{2,6}{2} = 1,3$

◆ Exemple : le problème du dé

- n lancers d'un dé parfait
- F_n fréquence de sorties de l'As

Trouver n_0 à partir duquel $\mathbb{P}\left(\left|F_n - \frac{1}{6}\right| < 0,01\right) \geq 95\%$

➢ Simulation

La feuille de calcul suivante permet de simuler 10000 séries de n lancers d'un dé pour diverses valeurs de n.

Une macro permet de calculer, pour chaque série, la fréquence F_n d'apparition de l'As et de tester si l'événement « $\left|F_n - \frac{1}{6}\right| < 0,1$ » est réalisé

```
Sub Tche()

Dim stat, compt, Nbr, X As Integer
Dim Fn As Variant
Nbr = Cells(5, 6).Value
Range("A5:A10005").Select
Selection.Clear
Randomize
For stat = 1 To 10000
        Fn = 0
        For compt = 1 To Nbr
            X = Int(Rnd * 6 + 1)
            If X = 6 Then
                Fn = Fn + 1
            End If
        Next compt
        Fn = Fn / Nbr
        Cells(4 + stat, 1).Value = Fn
  Next stat

End Sub
```

Simulation de la fréquence d'apparition du 6 sur n lancers de dé

Fn	test	P(|Fn -1/6|< 0,01)
0,1628	1	0,992000000
0,1664	1	
0,1631	1	
0,1687	1	
0,1655	1	
0,1691	1	
0,1661	1	
0,1655	1	
0,1618	1	
0,1688	1	
0,1665	1	
0,1574	1	
0,1631	1	
0,1733	1	
0,166	1	
0,1681	1	
0,1621	1	
0,1721	1	
0,1644	1	
0,1665	1	
0,161	1	
0,163	1	

approchée par la fréquence de l'événement " |Fn - 1/6| < 0,01 " sur 10 000 expériences

Entrer la valeur de n puis cliquer sur le bouton

n = 10000

10000 exp

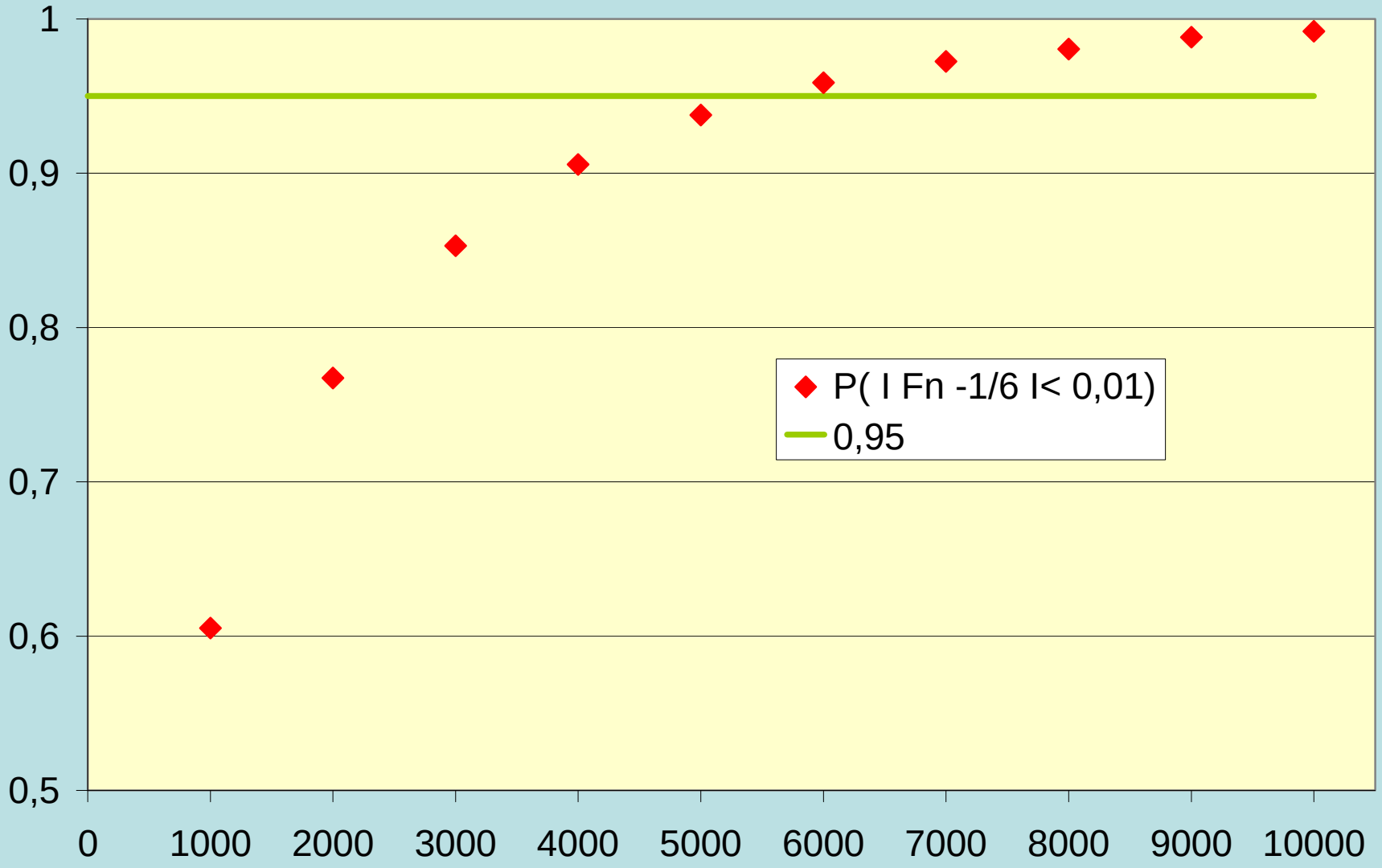
n =	P(|Fn -1/6 |< 0,01)
0	
1000	0,6053
2000	0,7673
3000	0,8529
4000	0,9056
5000	0,9377
6000	0,9587
7000	0,9725
8000	0,9805
9000	0,9881
10000	0,9920

On calcule alors la fréquence de cet événement sur 10000 expériences pour approcher la probabilité

$$P\left(\left|F_n - \frac{1}{6}\right| < 0{,}1\right).$$

Le graphique suivant représente les valeurs approchées de cette probabilité pour $n \in \{1000, 2000, 3000, ..., 10000\}$ et permet de déterminer à partir de quelle valeur de n

$$P\left(\left|F_n - \frac{1}{6}\right| < 0{,}1\right) \geq 0{,}95$$


1
0,9
0,8
0,7
0,6
0,5
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
P(I Fn -1/6 I< 0,01)
0,95

- Solution du problème du dé

➢ Méthode 1 : Inégalité de Tchebychev

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(\left|X - E(X)\right| < \varepsilon\right) \geq 1 - \frac{V(X)}{\varepsilon^2}$$

Ici : $\forall \varepsilon > 0 \quad \mathbb{P}\left(\left|F_n - p\right| < \varepsilon\right) \geq 1 - \frac{pq}{n\varepsilon^2}$

Avec $\varepsilon = 0,01$, $p = \frac{1}{6}$ et $q = \frac{5}{6}$

il suffit que $1 - \frac{5 \times 10^4}{36 \times n} \geq 0,95$ i. e. $n \geq \frac{10^6}{36}$ $\quad n \geq 27\,778$

- Méthode 2 : “Fourchette sur F_n”

$$\mathbb{P}\left(\left|F_n - \frac{1}{6}\right| < \frac{1}{\sqrt{n}}\right) \geq 0,95$$

Pour avoir $\mathbb{P}\left(\left|F_n - \frac{1}{6}\right| < 0,01\right) \geq 0,95$,

il suffit que $n \geq 10000$.

➢ Méthode 3 : approximation par loi normale

$$T_n = \frac{F_n - p}{\sqrt{\frac{pq}{n}}} = \frac{F_n - \frac{1}{6}}{\frac{1}{6}\sqrt{\frac{5}{n}}} \text{ suit } \mathcal{N}(0,1)$$

$$\left|F_n - \frac{1}{6}\right| < 0,01 \Leftrightarrow |T_n| \leq \frac{6\sqrt{n}}{100\sqrt{5}}$$

Remarques : si T suit $\mathcal{N}(0,1)$ et $\varphi(t) = \mathbb{P}(|T| \leq t)$

- $\varphi(1,96) = 0,95$
- φ est croissante

$\mathbb{P}\left(\left|F_n - \frac{1}{6}\right| < 0,01\right) \geq 0,95$ signifie $\varphi\left(\frac{6\sqrt{n}}{100\sqrt{5}}\right) \geq 0,95$

Il suffit que $\frac{6\sqrt{n}}{100\sqrt{5}} \geq 1,96$ d'où $n \geq 5336$

4- Fréquence d’échantillonnage Intervalle de confiance

4-1 Fréquence d'échantillonnage

- Données initiales

$\mathcal{P}$ population statistique (boules dans une urne)

$\mathcal{C}$ caractère à l'étude (couleur de boules (blanc par exemple))

p proportion d'individus possédant le caractère $\mathcal{C}$

➢ Échantillonnage

n entier donné

$\mathcal{E}$: échantillon de taille n issu de $\mathcal{P}$

(obtenu par tirage avec remise ;
modèle : équirépartition)

$\mathcal{E}_n$: ensemble de tous les échantillons de taille n

➢ Fréquence d'échantillonnage

- S_n est la variable aléatoire qui à chaque échantillon associe le nombre de boules blanches qu'il contient.

- F_n est la variable aléatoire qui à chaque échantillon associe la fréquence de boules blanches qu'il contient.

$$F_n = \frac{1}{n} S_n$$: fréquence d'échantillonnage

➢ Lois de probabilité - moments

- S_n suit la loi binomiale $\mathcal{B}(n, p)$

$$E(S_n) = np \qquad \sigma(S_n) = \sqrt{np(1-p)}$$

- $F_n = \frac{1}{n} S_n$

$$E(F_n) = p \qquad \sigma(F_n) = \sqrt{\frac{p(1-p)}{n}}$$

- Centrée réduite associée : $T_n = \dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$

➢ Exemple « fil rouge »

Un échantillon $\mathcal{E}$: 2000 boules.

600 boules blanches dans $\mathcal{E}$

fréquence observée : 0,3

4-2 Intervalle de confiance

- La clé de l'affaire

Le théorème de DE MOIVRE

- F_n suit approx^{t} la loi normale $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
- ou encore $T_n = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$

suit approximativement la loi normale $\mathcal{N}(0,1)$

(Conditions d'approximation ? Voir plus loin)

- Construction de l'intervalle de confiance

➢ Exemple : seuil de risque $\alpha = 0,05$

$$T_n \text{ suit } \sim \mathcal{N}(0,1)$$

$$\mathbb{P}\left(|T_n| \leq 1,96\right) = 1 - 0,05$$

$$\mathbb{P}\left(|F_n - p| \leq 1,96 \times \sqrt{\frac{p(1-p)}{n}}\right) = 0,95$$

$$r = 1,96 \times \sqrt{\frac{p(1-p)}{n}}$$

▪ $\left[F_n - r, F_n + r\right]$ est un intervalle **aléatoire**

(bornes : variables aléatoires)

La probabilité que $\left[F_n - r, F_n + r\right]$ contienne p est 0,95

➢ Seuil de risque α $(0 < \alpha < 1)$

$$[0,+\infty[\to [0,1[$$

$$t \mapsto \mathbb{P}(|T| \le t) = 2\pi(t) - 1$$

continue et strictement croissante $P(|T| < t) = 1 - \alpha$

admet une unique solution $t_\alpha > 0$: $\pi(t_\alpha) = 1 - \frac{\alpha}{2}$

$$\mathbb{P}\left(|F_n - p| \le t_\alpha \sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha$$

$$r_\alpha = t_\alpha \sqrt{\frac{p(1-p)}{n}}$$

La probabilité que l'intervalle aléatoire $[F_n - r_\alpha, F_n + r_\alpha]$ contient p est $1 - \alpha$

Le langage à connaître :

α est le seuil de risque

$1-\alpha$ est le niveau de confiance

r_α est la précision

Problème : r_α dépend de p

$$r_\alpha = t_\alpha \sqrt{\frac{p(1-p)}{n}}$$

- Sur le rayon r_α

➢ Méthode 1 : « élargissement »

$$p(1-p) \le \frac{1}{4} \Rightarrow r_\alpha \le \frac{t_\alpha}{2\sqrt{n}}$$

$\left[F_n - \frac{t_\alpha}{2\sqrt{n}}, F_n + \frac{t_\alpha}{2\sqrt{n}} \right]$ intervalle de confiance pour p de niveau $1-\alpha$ (au moins)

Exemple $\alpha = 0,05$

$\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}} \right]$ est l'intervalle de confiance pour p de niveau $0,95$ (au moins)… (déjà "vu")

Application « fil rouge »

$n = 2000$

$f = 0,3$ (valeur observée de F_n)

$$\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}} \right] \subset [0,275;0,325]$$

$[27,5\%;32,5\%]$ fourchette d'estimation de p
au niveau de confiance $0,95$

Illustration à l'aide d'une simulation

On considère une urne de Bernoulli dont la proportion p de boules rouges est fixée aléatoirement au départ. On s'intéresse à la variable aléatoire F_{100} : fréquence de boules rouges dans les échantillons de taille 100

Voici un algorithme simulant 1000 échantillons ε de taille 100, et qui calcule le pourcentage des intervalles $\left[f_\varepsilon - \frac{1}{\sqrt{100}} , f_\varepsilon + \frac{1}{\sqrt{100}} \right]$ qui contiennent la proportion p, f_ε étant la valeur observée de F_{100} dans l'échantillon ε.

M

```
1     VARIABLES
2       p EST_DU_TYPE NOMBRE
3       f EST_DU_TYPE NOMBRE
4       L EST_DU_TYPE NOMBRE
5       C EST_DU_TYPE NOMBRE
6       X EST_DU_TYPE NOMBRE
7       Y EST_DU_TYPE NOMBRE
8       FF EST_DU_TYPE NOMBRE
9     DEBUT_ALGORITHME
10      p PREND_LA_VALEUR random()
11      Y PREND_LA_VALEUR 0
12      FF PREND_LA_VALEUR 0
13      POUR C ALLANT_DE 1 A 1000
14        DEBUT_POUR
15        X PREND_LA_VALEUR 0
16        POUR L ALLANT_DE 1 A 100
17          DEBUT_POUR
18          X PREND_LA_VALEUR X+floor(p+random())
19          FIN_POUR
20        f PREND_LA_VALEUR X/100
21        FF PREND_LA_VALEUR FF+f
22        SI (abs(f-p)<=0.1) ALORS
23          DEBUT_SI
24          Y PREND_LA_VALEUR Y+1
25          FIN_SI
26        FIN_POUR
27      FF PREND_LA_VALEUR FF/1000
28      Y PREND_LA_VALEUR Y/10
```

```
29      AFFICHER "Proportion : "
30      AFFICHER p
31      AFFICHER " "
32      AFFICHER "Moyenne des fréquences : "
33      AFFICHER FF
34      AFFICHER " "
35      AFFICHER "Pourcentage de fréquences dans l'intervalle de confiance : "
36      AFFICHER Y
37    FIN_ALGORITHME
```

Résultats

```
***Algorithme lancé***
Proportion :   0.57241821
Moyenne des fréquences :  0.57301
Pourcentage d'intervalle contenant p : 95.7
***Algorithme terminé***
```

➢ Méthode 2 : « Estimation de l'écart type »

On propose comme fourchette d'estimation de p au niveau de confiance $1-\alpha$ l'intervalle :

$$\left[f - t_\alpha \sqrt{\frac{f(1-f)}{n}},\ f + t_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$$

(on remplace $\sqrt{\frac{p(1-p)}{n}} = \sigma(F_n)$ par

la valeur observée : $\sqrt{\frac{f(1-f)}{n}}$)

Application « fil rouge »

$n = 2000$; $f = 0,3$; $\alpha = 0,05$

d'où $t_\alpha = 1,96$ $(r = 0,0200084...)$

fourchette d'estimation de p
au niveau de confiance $0,95$: $[28\%, 32\%]$

À signaler $\sqrt{\dfrac{f(1-f)}{n-1}}$ pour "débiaiser"

➢ Méthode 3 : « de l'ellipse »

Pour une valeur f **observée** la fourchette d'estimation de p au niveau de confiance $1-\alpha$ s'obtient en résolvant l'inéquation $|p-f| \leq t_\alpha \sqrt{\dfrac{p(1-p)}{n}}$

Application $\alpha = 0,05, \quad t_\alpha = 1,96$

$$|p-f| \leq t_\alpha \sqrt{\frac{p(1-p)}{n}} \Leftrightarrow (p-f)^2 \leq 1,96^2 \frac{p(1-p)}{n}$$

Avec $f = 0,3$ et $n = 2000$

$1,0019\,p^2 - 0,6019\,p + 0,09 \leq 0$

$p_1 \approx 0,2804$ et $p_2 \approx 0,3203$

On retiendra $[0,28\,;\,0,32]$

Dans le cas général On pose $x = \dfrac{t_\alpha}{\sqrt{n}}$

$$|p - f| \le t_\alpha \sqrt{\frac{p(1-p)}{n}} \Leftrightarrow p^2(1+x^2) - (2f + x^2) + f^2 \le 0$$

Solutions de l'équation $\dfrac{2f + x^2 \pm x\sqrt{x^2 + 4f(1-f)}}{2(1+x^2)}$

Négligeons x^2 $f \pm x\sqrt{f(1-f)} = f \pm t_\alpha \sqrt{\dfrac{f(1-f)}{n}}$

"méthode 2"

Noter : $\alpha = 0,01$ (ambitieux)

$t_\alpha \le 2,6$ $n \ge 1000$ $\qquad x^2 = \dfrac{t_\alpha^2}{n} \le 0,007$

5 Quelques remarques pour conclure

5-1 Une question de vocabulaire

« *une fourchette d'estimation de p au niveau de confiance 0,95 est* $\left[28\% \text{ , } 32\%\right]$ »

Question : Peut on dire que la probabilité que *p* soit entre 28% et 32% est 0,95 ?

Réponse : **NON** (*p* est dans l'intervalle ou n'y est pas)

➢ *abus de langage, interprétation…*

5-2 Sur la taille de l'échantillon

Les conditions

- « grands » échantillons ($n \geq 100$ est très « bon »)
  (petits échantillons :William GAUSSET)
- fréquence observée : pas trop petite ni trop grande (par exemple : entre 0,2 et 0,8)

L'échantillon de taille 1000

Au taux de confiance 0,95 la précision est de
2% à 3% environ.
(performance habituelle des sondages médiatisés)

5-3 Et la taille de la population ?

« Alors que dans certaines branches de la Physique, les experts estiment ne pouvoir se prononcer avec quelque validité sur la fréquence d'un phénomène qu'à partir de millions d'observations, il suffirait d'un calepin et d'un questionnaire, de procéder à l'interrogation de 1000 personnes pour connaître le comportement politique d'une nation de 50 millions d'habitants » .

Et pourtant la taille de la population n'intervient pas !

5-4 La précision coûte cher

Exemple :

Niveau de confiance fixé à 95%

Echantillon de taille 1000 ; fréquence observée f ; précision 3%.

Echantillon de taille 10000 ; fréquence observée f' (on suppose $f' \approx f$) ; précision 1%

5-4 La confiance aussi

Exemple :

Précision fixée à 3%

Si l'on a un niveau de confiance de 95% avec un échantillon de taille 1000, il faut un échantillon de taille 1750 environ pour atteindre le niveau de confiance 99% (mêmes hypothèses que dans l'exemple précédent).