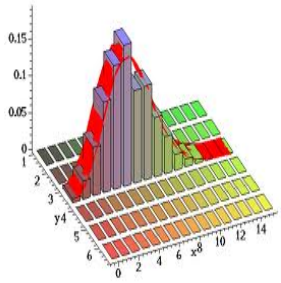
# Statistiques Inférentielles

Christian CYRILLE

28 mars 2014




# 1 Définitions

## 1.1 Statistique descriptive,statistique inférentielle

Le mot **statistique** englobe une série de techniques de traitement de l'information et de données numériques et qualitatives.
En gros, il s'agit de retirer le maximum d'informations de l'expérience. Les statistiques représentant une branche des mathématiques subdivisée en deux domaines :

- **les statistiques descriptives** qui consistent à faire une étude structurale des données afin d'en extraire des modes de représentation simples
- **les statistiques mathématiques ou inférentielles ou inductives**, basée sur la théorie des probabilités.

En statistiques inférentielles, on observe des statistiques répétées d'un phénomène . On se demande quelles sont les lois qui peuvent régir ce phénomène. Il s'agit donc de **modéliser une fraction observable du réel** comme résultant d'un phénomène aléatoire pour lequel on observe non pas une mais une famille de lois de probabilités possibles.
Dès qu'il s'agit de prendre des décisions pertinentes sur des données aléatoires(électorales, économiques, industrielles, scietifiques,...) dont on n'a qu'une représentation partielle, il faut faire appel aux statistiques inférentielles.

## 1.2 Echantillon

L'idée de décrire une population à partir d'un échantillon n'est apparue qu'à partir du milieu du 18-ème siècle, notamment par l'école *"d'arithmétique politique anglaise"*. Un échantillon de taille $n$ est une série statistique formée de $n$ résultats obtenus lorsque l'on répète $n$ fois une expérience dans les mêmes conditions. Les issues d'une expérience sont les résultats que l'on peut obtenir en réalisant cette expérience.
La distribution des fréquences associée à un échantillon est la liste des fréquences des issues de l'échantillon.
L'observation d'échantillons permet d'éviter le recensement exhaustif qui est lourd et coûteux. Mais pour avoir le maximum d'exactitude, il faut que tenir compte non seulement de la représentativité de l'échantillon mais aussi de sa fluctuation. Même tiré au sort un échantillon n'est pas l'image exacte de la population, en raison des fluctuations d'échantillonnage.

## 1.3 Fluctuation d'échantillonnage

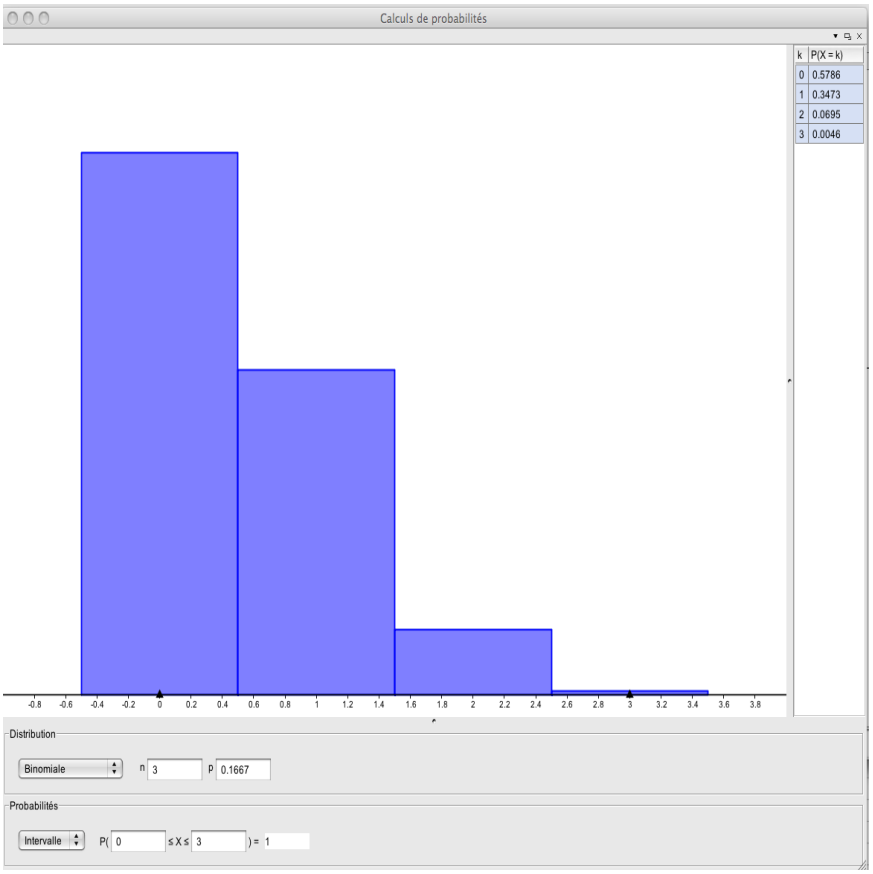
- Les distributions des fréquences varient d'un échantillon à l'autre pour une même expérience : c'est ce que l'on appelle la fluctuation d'échantillonnage.
- Pour des échantillons de même taille, les fréquences peuvent fluctuer.
- Lorsque la taille $n$ de l'échantillon augmente, l'ampleur des fluctuations des distributions de fréquence calculées sur ces échantillons de taille $n$ diminue et les fréquences tendent à se stabiliser.

## 1.4 Exemple

Lorsqu'on lance 3 fois de suite un dé non truqué. On s'intéresse au nombre de fois où l'on obtient 6. On est donc en présence d'un schéma de Bernoulli : 3 épreuves, répétées, identiques et indépendantes. Au cours d'une épreuve , on a :

- soit un succès :"sortir le numéro 6 " avec une probabilité $p = \frac{1}{6} \approx 0,16$
- soit un échec :"ne pas sortir le numéro 6 " avec une probabilité $q = 1 - p = \frac{5}{6}$

La variable aléatoire $X$ ="nombre de succès suit la loi binomiale $\mathcal{B}(3; \frac{1}{6})$.
Voici les probabilités théoriques :




| $k$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $Pr([X = k])$ | $0,5787$ | $0,3472$ | $0,0694$ | $0,0046$ |

A l'aide de l'algorithme suivant, on peut simuler $N$ fois ce schéma de Bernoulli suivant ce qui constituera un échantillon de taille $N$ :

```
1   VARIABLES
2     I EST_DU_TYPE NOMBRE
3     K EST_DU_TYPE NOMBRE
4     J EST_DU_TYPE NOMBRE
5     X EST_DU_TYPE LISTE
6     DE EST_DU_TYPE LISTE
7     F EST_DU_TYPE LISTE
8     N EST_DU_TYPE NOMBRE
9     NBSUCCES EST_DU_TYPE NOMBRE
10  DEBUT_ALGORITHME
11    LIRE N
12    POUR I ALLANT_DE 0 A 3
13      DEBUT_POUR
14      X[I] PREND_LA_VALEUR 0
15      FIN_POUR
16    POUR K ALLANT_DE 1 A N
17      DEBUT_POUR
18      AFFICHER "Essai numéro"
19      AFFICHER K
20      NBSUCCES PREND_LA_VALEUR 0
21      POUR J ALLANT_DE 1 A 3
22        DEBUT_POUR
23        DE[J] PREND_LA_VALEUR 1+floor(6*random())
24        AFFICHER "le dé marque "
25        AFFICHER DE[J]
26        SI (DE[J]==6) ALORS
27          DEBUT_SI
28          NBSUCCES PREND_LA_VALEUR NBSUCCES + 1
29          FIN_SI
30        FIN_POUR
31      X[NBSUCCES] PREND_LA_VALEUR X[NBSUCCES] +1
32      FIN_POUR
33    POUR I ALLANT_DE 0 A 3
34      DEBUT_POUR
35      AFFICHER "X = "
36      AFFICHER I
37      AFFICHER "  est apparu "
38      AFFICHER X[I]
39      AFFICHER " fois"
40      AFFICHER "avec une fréquence de "
41      F[I] PREND_LA_VALEUR X[I]/N
42      AFFICHER F[I]
43      FIN_POUR
44  FIN_ALGORITHME
```

Voici les résultats pour $N = 6$ puis pour $N = 100$

RÉSULTATS :

```
***Algorithme lancé***
Essai numéro1
le dé marque 1
le dé marque 4
le dé marque 6
Essai numéro2
le dé marque 4
le dé marque 1
le dé marque 4
Essai numéro3
le dé marque 2
le dé marque 6
le dé marque 4
Essai numéro4
le dé marque 2
le dé marque 6
le dé marque 5
Essai numéro5
le dé marque 1
le dé marque 6
le dé marque 3
Essai numéro6
le dé marque 1
le dé marque 3
le dé marque 6
X = 0  est apparu 1 fois
avec une fréquence de 0.16666667
X = 1  est apparu 5 fois
avec une fréquence de 0.83333333
X = 2  est apparu 0 fois
avec une fréquence de 0
X = 3  est apparu 0 fois
avec une fréquence de 0
***Algorithme terminé***
```

```
Essai numéro99
le dé marque 4
le dé marque 3
le dé marque 2
Essai numéro100
le dé marque 5
le dé marque 6
le dé marque 6
X = 0  est apparu 64 fois
avec une fréquence de 0.64
X = 1  est apparu 25 fois
avec une fréquence de 0.25
X = 2  est apparu 11 fois
avec une fréquence de 0.11
X = 3  est apparu 0 fois
avec une fréquence de 0
***Algorithme terminé***
```

## 1.5 Tests d'hypothèses

On considère une population dans laquelle on suppose que la proportion théorique d'un certain caractère $X$ 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 étudié $X$.
On veut donc tester l'hypothèse dite hypothèse nulle $\mathcal{H}_0$ : "$p = f$".
On rejette l'hypothèse $\mathcal{H}_0$ lorsque la fréquence observée $f$ est trop éloignée de $p$ dans un sens ou dans l'autre. On choisit de fixer le seuil à 95% de sorte que la probabilité de rejeter $\mathcal{H}_0$ (alors qu'elle serait vraie) est inférieure à 5%.
Ce seuil de 5% présente des avantages pratiques : C'est un risque faible et de plus, la détermination de la fourchette correspondante ne nécessite qu'un calcul simple, sans consultation de tables.

## 1.6 Intervalle de fluctuation

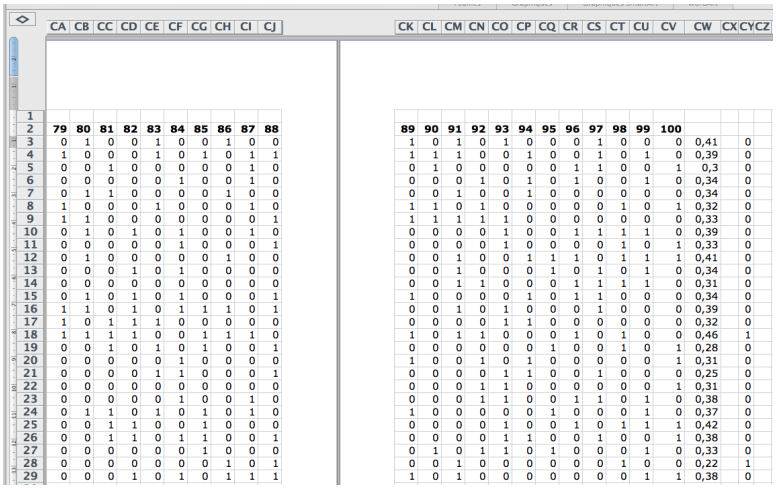
L'intervalle de fluctuation au seuil de 95% relatif aux échantillons de taille $n$ est l'intervalle centré autour de $p$ (proportion du caractère dans la population) où se situe , avec une probabilité égale à 95% , la fréquence observée dans un échantillon de taille $n$. Bien entendu, il existe une infinité de fourchettes, une pour chaque risque d'erreur adopté. On doit donc trouver un compromis entre le risque acceptable et le souci de précision.
Dans un article *"l'épidémiologiste traque le hasard"*, Le Professeur Daniel Schwartz, fondateur de l'Ecole française de Statistique Médicale, affirme :
*"A l'impossible certitude, la fourchette apporte le remède (partiel) de l'incertitude contrôlée. Le triomphe de cette solution ne fut pas aisé. Entre le moment où apparaît l'idée révolutionnaire du jugement sur échantillon et celui où on proposa le calcul d'intervalles de confiance valables s'écoulèrent deux siècles parsemés de longs et âpres combats. le premier conflit qui portait sur le principe même du jugement sur échantillon dressa les tenants du tout(exact mais onéreux) contre ceux de la partie(moins onéreuse mais moins exacte) La deuxième querelle , une fois admise la nécessité de l'échantillon , opposa les partisans du choix judicieux des sujets à ceux de leur tirage au sort. C'est seulement en* 1934 *que le célèbre statisticien Jerzy Neyman y mit un terme en montrant la supériorité du tirage au sort : son principal argument est que celui-ci engendre un échantillon comparable en moyenne à la population pour tous les caractères connus ou inconnus des sujets, ce que ne peut faire aucun choix raisonné."*

## 1.7 Exercice Indice Bordas 2de 2009 p 131

2 listes la $R(ouge)$ et la $B(leue)$ s'affrontent lors d'une élection. **On suppose connue la proportion** $p$ **d**es électeurs qui votent en général pour la liste $R$,
On constitue un échantillon de $n$ électeurs et **on calcule la proportion** $f$ des individus favorables à la liste R.
A l'aide d'un tableur , on va vérifier le théorème suivant :
Si $p \in [0,2;0,8]$ et si $n \geq 25$ alors $Pr[p - \frac{1}{\sqrt{n}} \leq f \leq p + \frac{1}{\sqrt{n}}] \approx 0,95$
En prenant par exemple $p = 35\%$ , on va créer 100 échantillons de taille 100 et l'on va dénombrer les échantillons pour lesquels la fréquence est en dehors de l'intervalle de fluctuation $[0,25;0,45]$ au seuil de confiance de 95%

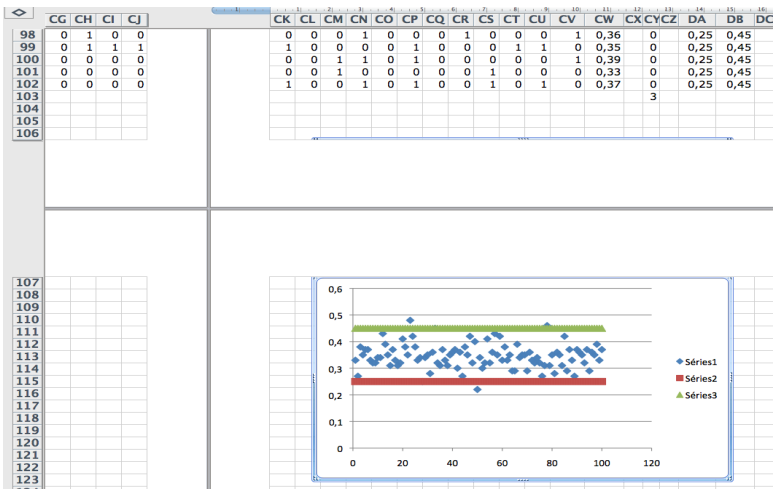
1. En $B1$ entrer la valeur de $p$
2. En $A2$ écrire 1 puis afficher les numéros de tirages de 1 à 100 à l'aide du menu Edition/ Recopie/Série
3. En $A3$ écrire $= ENT(ALEA() + 0,35)$ puis recopier cette formule jusqu'en $CV3$. Comme $0 \leq ALEA() < 1$ alors
   - ou bien avec une probabilité de 65% on a : $0 \leq ALEA() < 0,65$ donc $0,35 \leq ALEA() + 0,35 < 1$ et $= ENT(ALEA() + 0,35)$ est 0
   - ou bien avec une probabilité de 35% on a : $0,65 \leq ALEA() < 1$ donc $1 \leq ALEA() + 0,35 < 1,35$ et $= ENT(ALEA() + 0,35)$ est 1

   Tout électeur ayant voté pour la liste $R$ sera crédité d'un 1
4. Créer à l'aide de Recopie vers le bas les 100 échantillons.
5. En $CW3$ la formule $= SOMME(A3 : CV3)/100$ affichera la fréquence du nombre d'électeurs ayant voté pour la liste $R$.
   On remplit la colonne $CW$
6. En $CY3$ on écrit la formule $= SI(OU(CW3 < 0,25; CW3 > 0,45); 1; 0)$ qu'on recopie dans la colonne $CY$
7. En $CY103$ on fait la somme des nombres de la zone $CY3 : CY102$.
   On obtient ainsi le nombre de cas où la fréquence est en dehors de l'intervalle de fluctuation.
8. On représente graphiquement en nuage de points la zone $CW3 : CW102$
   On entre ensuite en colonne $DA$ la valeur $0,25$ et en colonne $DB$ la valeur $0,45$ puis l'on représente les zones $DA3 : DA102$ et $DB3 : DB102$ par un clic droit sur graphique /Données /ajouter valeurs y/



| 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |

| 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0,41 | 0 | |
| 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0,39 | 0 | |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0,3 | 0 | |
| 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0,34 | 0 | |
| 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,34 | 0 | |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0,32 | 0 | |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0,33 | 0 | |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0,39 | 0 | |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0,33 | 0 | |
| 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0,41 | 0 | |
| 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0,34 | 0 | |
| 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0,31 | 0 | |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0,34 | 0 | |
| 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0,39 | 0 | |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,32 | 0 | |
| 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0,46 | 1 | |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0,28 | 0 | |
| 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0,31 | 0 | |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0,25 | 0 | |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0,31 | 0 | |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0,38 | 0 | |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0,37 | 0 | |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0,42 | 0 | |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0,38 | 0 | |
| 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0,33 | 0 | |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0,22 | 1 | |
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0,38 | 0 | |



| 98 | 0 | 1 | 0 | 0 |
|---|---|---|---|---|
| 99 | 0 | 1 | 1 | 1 |
| 100 | 0 | 0 | 0 | 0 |
| 101 | 0 | 0 | 0 | 0 |
| 102 | 0 | 0 | 0 | 0 |

| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0,36 | 0 | | 0,25 | 0,45 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0,35 | 0 | | 0,25 | 0,45 | |
| 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0,39 | 0 | | 0,25 | 0,45 | |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0,33 | 0 | | 0,25 | 0,45 | |
| 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0,37 | 0 | | 0,25 | 0,45 | |
| | | | | | | | | | | | | | 3 | | | | |

## 2 Utilisation de la loi binomiale

Nous venons de voir sur un exemple que sous réserve de satisfaire aux conditions de validité suivantes $n > 25$ et $0,2 \leq p \leq 0,8$ on peut coincer $f$ dans l'intervalle de fluctuation suivant : $[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$ avec un risque d'erreur de 5%.
Cela supposait aussi que la fréquence suivait implicitement une loi continue.
Avec la loi binomiale, on va construire :

- construire un intervalle de fluctuation indépendant de $p$ et de $n$
- élaborer une démarche de prise de décision afin de tester l'hypothèse nulle $\mathcal{H}_0$ : "$p = f$".

La loi binomiale permet de calculer très exactement les probabilités des fréquences observables dans un échantillon de taille $n$ , à savoir les valeurs de la variable aléatoire $F = \frac{X}{n}$ qui prend les valeurs $\frac{k}{n}$ où $0 \leq k \leq n$.
La variable aléatoire $X$ correspondant au nombre de fois où le caractère est observé dans un échantillon de taille $n$ suit la loi binomiale $\mathcal{B}(n; p)$. Le diagramme de $X$ n'est pas toujours symétrique autour de $p$ et de plus la loi $X$ prend des valeurs discrètes, on ne peut donc pas déterminer précisément un intervalle où la probabilité serait exactement de 95% , on va donc construire un intervalle qui approxime l'intervalle de fluctuation de la manière suivante :

- Cet intervalle est $[\frac{k_1}{n}; \frac{k_2}{n}]$
- $k_1$ est le plus petit entier tel que $Pr([X \leq k_1]) > 2,5\%$
- $k_2$ est le plus petit entier tel que $Pr([X \leq k_2]) > 97,5\%$

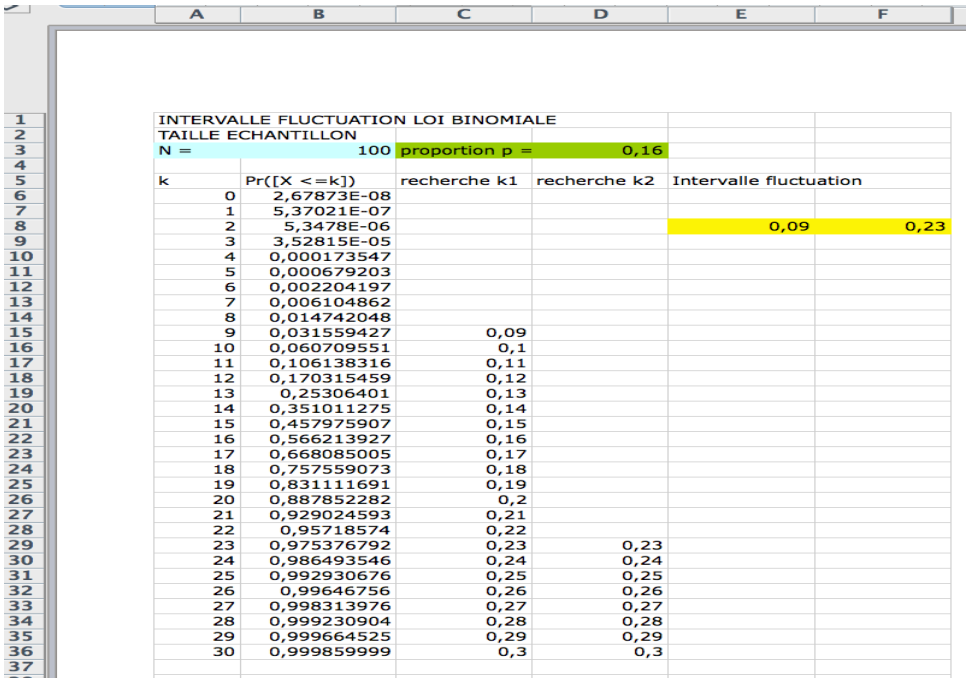
Dès que $n$ est assez grand cet intervalle $[\frac{k_1}{n}; \frac{k_2}{n}]$ est quasiment centré en $p$ et est quasiment le même que $[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$.
La règle de décision adoptée est la suivante :

- si $f \in [\frac{k_1}{n}; \frac{k_2}{n}]$ alors on accepte l'hypothèse nulle $\mathcal{H}_0$ : "$p = f$" avec un risque d'erreur de 5%
- sinon on rejette l'hypothèse nulle $\mathcal{H}_0$ : "$p = f$" avec un risque d'erreur de 5%

## 3 Exemple de prise de décision

Un médecin de santé publique veut savoir si, dans sa région, le pourcentage d'habitants atteints d'hypertension artérielle est égal à la valeur de 16% récemment publiée pour des populations semblables.
En notant $p$ la proportion d'hypertendus dans la population de sa région, le médecin formule l'hypothèse $p = 0,16$.
Pour vérifier cette hypothèse, le médecin constituera un échantillon de $n = 100$ habitants de la région ; il déterminera la fréquence $f$ d'hypertendus(l'échantillon est prélevé au hasard et la population est suffisamment importante pour considérer qu'il s'agit de tirages avec remise.)
Lorsque la proportion dans la population vaut $p = 0,16$, la variable aléatoire $X$ correspondant au nombred'hypertendus observé dans un échantillon aléatoire de taille $n = 100$ suit la loi binomiale $\mathcal{B}(n; p)$.
On cherche à partager l'intervalle $[0; 100]$ où $X$ prend ses valeurs en trois intervalles :
$[0, k_1 - 1], [k_1, k_2], [k_2 + 1, 100]$ de sorte que la variable aléatoire $X$ prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de $2,5\%$ sans dépasser cette valeur.

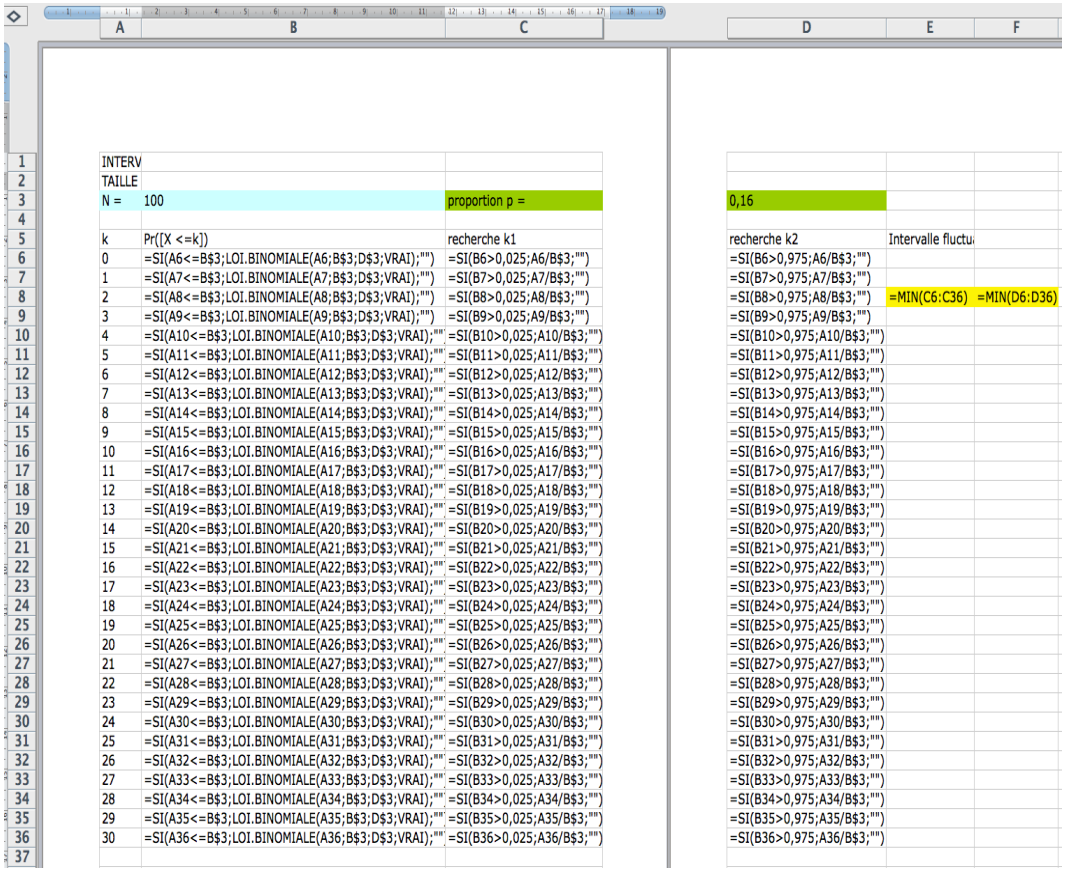
On recherche donc le plus grand entier $k_1$ tel que $P([X < k_1] \leq 2,5\%$ et le plus petit entier $k_2$ tel que $P([X < k_2] \leq 2,5\%$
La règle de décision , pour le médecin sera la suivante :



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | INTERVALLE FLUCTUATION LOI BINOMIALE | | | | | |
| 2 | TAILLE ECHANTILLON | | | | | |
| 3 | N = | 100 | proportion p = | 0,16 | | |
| 4 | | | | | | |
| 5 | k | Pr([X <=k]) | recherche k1 | recherche k2 | Intervalle fluctuation | |
| 6 | 0 | 2,67873E-08 | | | | |
| 7 | 1 | 5,37021E-07 | | | | |
| 8 | 2 | 5,3478E-06 | | | 0,09 | 0,23 |
| 9 | 3 | 3,52815E-05 | | | | |
| 10 | 4 | 0,000173547 | | | | |
| 11 | 5 | 0,000679203 | | | | |
| 12 | 6 | 0,002204197 | | | | |
| 13 | 7 | 0,006104862 | | | | |
| 14 | 8 | 0,014742048 | | | | |
| 15 | 9 | 0,031559427 | 0,09 | | | |
| 16 | 10 | 0,060709551 | 0,1 | | | |
| 17 | 11 | 0,106138316 | 0,11 | | | |
| 18 | 12 | 0,170315459 | 0,12 | | | |
| 19 | 13 | 0,25306401 | 0,13 | | | |
| 20 | 14 | 0,351011275 | 0,14 | | | |
| 21 | 15 | 0,457975907 | 0,15 | | | |
| 22 | 16 | 0,566213927 | 0,16 | | | |
| 23 | 17 | 0,668085005 | 0,17 | | | |
| 24 | 18 | 0,757559073 | 0,18 | | | |
| 25 | 19 | 0,831111691 | 0,19 | | | |
| 26 | 20 | 0,887852282 | 0,2 | | | |
| 27 | 21 | 0,929024593 | 0,21 | | | |
| 28 | 22 | 0,95718574 | 0,22 | | | |
| 29 | 23 | 0,975376792 | 0,23 | 0,23 | | |
| 30 | 24 | 0,986493546 | 0,24 | 0,24 | | |
| 31 | 25 | 0,992930676 | 0,25 | 0,25 | | |
| 32 | 26 | 0,99646756 | 0,26 | 0,26 | | |
| 33 | 27 | 0,998313976 | 0,27 | 0,27 | | |
| 34 | 28 | 0,999230904 | 0,28 | 0,28 | | |
| 35 | 29 | 0,999664525 | 0,29 | 0,29 | | |
| 36 | 30 | 0,999859999 | 0,3 | 0,3 | | |
| 37 | | | | | | |

- Si la fréquence observée $f$ appartient à l'intervalle de fluctuation $[0,09; 0,23]$ , on considère que l'hypothèse selon laquelle la proportion d'hypertendus dans la population est $p = 0,16$ n'est pas remise en question et on l'accepte.
- sinon on rejette l'hypothèse selon laquelle cette proportion vaut $p = 0,16$.

Voici la même feuille de calcul avec l'affichage des formules



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | INTERV | | | | | |
| 2 | TAILLE | | | | | |
| 3 | N = | 100 | proportion p = | 0,16 | | |
| 4 | | | | | | |
| 5 | k | Pr([X <=k]) | recherche k1 | recherche k2 | Intervalle fluctu | |
| 6 | 0 | =SI(A6<=B$3;LOI.BINOMIALE(A6;B$3;D$3;VRAI);"") | =SI(B6>0,025;A6/B$3;"") | =SI(B6>0,975;A6/B$3;"") | | |
| 7 | 1 | =SI(A7<=B$3;LOI.BINOMIALE(A7;B$3;D$3;VRAI);"") | =SI(B7>0,025;A7/B$3;"") | =SI(B7>0,975;A7/B$3;"") | | |
| 8 | 2 | =SI(A8<=B$3;LOI.BINOMIALE(A8;B$3;D$3;VRAI);"") | =SI(B8>0,025;A8/B$3;"") | =SI(B8>0,975;A8/B$3;"") | =MIN(C6:C36) | =MIN(D6:D36) |
| 9 | 3 | =SI(A9<=B$3;LOI.BINOMIALE(A9;B$3;D$3;VRAI);"") | =SI(B9>0,025;A9/B$3;"") | =SI(B9>0,975;A9/B$3;"") | | |
| 10 | 4 | =SI(A10<=B$3;LOI.BINOMIALE(A10;B$3;D$3;VRAI);"" | =SI(B10>0,025;A10/B$3;"") | =SI(B10>0,975;A10/B$3;"") | | |
| 11 | 5 | =SI(A11<=B$3;LOI.BINOMIALE(A11;B$3;D$3;VRAI);"" | =SI(B11>0,025;A11/B$3;"") | =SI(B11>0,975;A11/B$3;"") | | |
| 12 | 6 | =SI(A12<=B$3;LOI.BINOMIALE(A12;B$3;D$3;VRAI);"" | =SI(B12>0,025;A12/B$3;"") | =SI(B12>0,975;A12/B$3;"") | | |
| 13 | 7 | =SI(A13<=B$3;LOI.BINOMIALE(A13;B$3;D$3;VRAI);"" | =SI(B13>0,025;A13/B$3;"") | =SI(B13>0,975;A13/B$3;"") | | |
| 14 | 8 | =SI(A14<=B$3;LOI.BINOMIALE(A14;B$3;D$3;VRAI);"" | =SI(B14>0,025;A14/B$3;"") | =SI(B14>0,975;A14/B$3;"") | | |
| 15 | 9 | =SI(A15<=B$3;LOI.BINOMIALE(A15;B$3;D$3;VRAI);"" | =SI(B15>0,025;A15/B$3;"") | =SI(B15>0,975;A15/B$3;"") | | |
| 16 | 10 | =SI(A16<=B$3;LOI.BINOMIALE(A16;B$3;D$3;VRAI);"" | =SI(B16>0,025;A16/B$3;"") | =SI(B16>0,975;A16/B$3;"") | | |
| 17 | 11 | =SI(A17<=B$3;LOI.BINOMIALE(A17;B$3;D$3;VRAI);"" | =SI(B17>0,025;A17/B$3;"") | =SI(B17>0,975;A17/B$3;"") | | |
| 18 | 12 | =SI(A18<=B$3;LOI.BINOMIALE(A18;B$3;D$3;VRAI);"" | =SI(B18>0,025;A18/B$3;"") | =SI(B18>0,975;A18/B$3;"") | | |
| 19 | 13 | =SI(A19<=B$3;LOI.BINOMIALE(A19;B$3;D$3;VRAI);"" | =SI(B19>0,025;A19/B$3;"") | =SI(B19>0,975;A19/B$3;"") | | |
| 20 | 14 | =SI(A20<=B$3;LOI.BINOMIALE(A20;B$3;D$3;VRAI);"" | =SI(B20>0,025;A20/B$3;"") | =SI(B20>0,975;A20/B$3;"") | | |
| 21 | 15 | =SI(A21<=B$3;LOI.BINOMIALE(A21;B$3;D$3;VRAI);"" | =SI(B21>0,025;A21/B$3;"") | =SI(B21>0,975;A21/B$3;"") | | |
| 22 | 16 | =SI(A22<=B$3;LOI.BINOMIALE(A22;B$3;D$3;VRAI);"" | =SI(B22>0,025;A22/B$3;"") | =SI(B22>0,975;A22/B$3;"") | | |
| 23 | 17 | =SI(A23<=B$3;LOI.BINOMIALE(A23;B$3;D$3;VRAI);"" | =SI(B23>0,025;A23/B$3;"") | =SI(B23>0,975;A23/B$3;"") | | |
| 24 | 18 | =SI(A24<=B$3;LOI.BINOMIALE(A24;B$3;D$3;VRAI);"" | =SI(B24>0,025;A24/B$3;"") | =SI(B24>0,975;A24/B$3;"") | | |
| 25 | 19 | =SI(A25<=B$3;LOI.BINOMIALE(A25;B$3;D$3;VRAI);"" | =SI(B25>0,025;A25/B$3;"") | =SI(B25>0,975;A25/B$3;"") | | |
| 26 | 20 | =SI(A26<=B$3;LOI.BINOMIALE(A26;B$3;D$3;VRAI);"" | =SI(B26>0,025;A26/B$3;"") | =SI(B26>0,975;A26/B$3;"") | | |
| 27 | 21 | =SI(A27<=B$3;LOI.BINOMIALE(A27;B$3;D$3;VRAI);"" | =SI(B27>0,025;A27/B$3;"") | =SI(B27>0,975;A27/B$3;"") | | |
| 28 | 22 | =SI(A28<=B$3;LOI.BINOMIALE(A28;B$3;D$3;VRAI);"" | =SI(B28>0,025;A28/B$3;"") | =SI(B28>0,975;A28/B$3;"") | | |
| 29 | 23 | =SI(A29<=B$3;LOI.BINOMIALE(A29;B$3;D$3;VRAI);"" | =SI(B29>0,025;A29/B$3;"") | =SI(B29>0,975;A29/B$3;"") | | |
| 30 | 24 | =SI(A30<=B$3;LOI.BINOMIALE(A30;B$3;D$3;VRAI);"" | =SI(B30>0,025;A30/B$3;"") | =SI(B30>0,975;A30/B$3;"") | | |
| 31 | 25 | =SI(A31<=B$3;LOI.BINOMIALE(A31;B$3;D$3;VRAI);"" | =SI(B31>0,025;A31/B$3;"") | =SI(B31>0,975;A31/B$3;"") | | |
| 32 | 26 | =SI(A32<=B$3;LOI.BINOMIALE(A32;B$3;D$3;VRAI);"" | =SI(B32>0,025;A32/B$3;"") | =SI(B32>0,975;A32/B$3;"") | | |
| 33 | 27 | =SI(A33<=B$3;LOI.BINOMIALE(A33;B$3;D$3;VRAI);"" | =SI(B33>0,025;A33/B$3;"") | =SI(B33>0,975;A33/B$3;"") | | |
| 34 | 28 | =SI(A34<=B$3;LOI.BINOMIALE(A34;B$3;D$3;VRAI);"" | =SI(B34>0,025;A34/B$3;"") | =SI(B34>0,975;A34/B$3;"") | | |
| 35 | 29 | =SI(A35<=B$3;LOI.BINOMIALE(A35;B$3;D$3;VRAI);"" | =SI(B35>0,025;A35/B$3;"") | =SI(B35>0,975;A35/B$3;"") | | |
| 36 | 30 | =SI(A36<=B$3;LOI.BINOMIALE(A36;B$3;D$3;VRAI);"" | =SI(B36>0,025;A36/B$3;"") | =SI(B36>0,975;A36/B$3;"") | | |
| 37 | | | | | | |

# 4 Justification

## 4.1 Inégalité de Bienaimé-Tchebicheff

### 4.1.1 4 formes équivalentes

Soit $X$ une variable aléatoire réelle discrète de moyenne $m$ et d'écart type $\sigma$

1. Forme 1 : $\forall t > 0 \;\; P([\,|\,X - m\,| \geq \; t\,]) \leq \dfrac{\sigma^2}{t^2}$
2. Forme 2 : $\forall t > 0 \;\; P([\,|\,X - m\,| < \; t\,]) \;\geq 1 - \dfrac{\sigma^2}{t^2}$
3. Forme 3 : $\forall k > 0 \;\; P([\,|\,X - m\,| < \; k\,\sigma\,]) \;\geq 1 - \dfrac{1}{k^2}$
4. Forme 4 : $\forall k > 0 \;\; P([\,|\,X - m\,| \geq \; k\,\sigma\,]) \;\leq \dfrac{1}{k^2}$

## 4.2 Loi faible des grands nombres

Soient $n$ variables aléatoires réelles $X_1, X_2, \cdots, X_n$ :
- définies sur le même espace probabilisé $(\Omega, \mathcal{B}, P)$
- indépendantes deux à deux
- de même loi de probabilité donc
  de même espérance $E(X)$ et de même variance $Var(X)$

Alors la variable aléatoire dite moyenne ou fréquence empirique d'échantillonnage :
$F_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ vérifie les trois propriétés suivantes :

1. $E(F_n) = E(X)$
2. $Var(F_n) = \frac{1}{n} Var(X)$
3. $\forall \varepsilon > 0 \lim\limits_{n \mapsto +\infty} P(|\ (F_n - E(X)\ | > \varepsilon) = 0$
   ou encore $\forall \varepsilon > 0 \lim\limits_{n \mapsto +\infty} P(|\ (F_n - E(X)\ | \leq \varepsilon) = 1$
   On dit alors que $(F_n)$ converge donc en probabilité vers $E(X)$

### 4.2.1 Démonstration

1. En utilisant la linéarité de l'espérance, on obtient :
   $E(F_n) = E(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n} E(\sum_{i=1}^{n} X_i) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} n E(X) = E(X)$
2. On sait que la variance de variables aléatoires indépendantes est la somme de ces variances alors :
   $Var(F_n) = Var(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n Var(X) = \frac{1}{n} Var(X)$
3. D'après l'inégalité de Bienaimé-Tchebycheff on a :
   $0 \leq P(|\ (F_n - E(X)\ | > \varepsilon) \leq \frac{Var(F_n)}{\varepsilon^2} = \frac{1}{n} \frac{Var(X)}{\varepsilon^2}$.
   Or $\lim\limits_{n \mapsto +\infty} \frac{1}{n} \frac{Var(X)}{\varepsilon^2} = 0$ donc $\forall \varepsilon > 0 \lim\limits_{n \mapsto +\infty} P(|\ (F_n - E(X)\ | > \varepsilon) = 0$

## 4.3 Le théorème d'Or de Bernoulli

C'est un corollaire de la loi faible des grands nombres appliquée à la loi binomiale.

Soient $n$ variables aléatoires réelles $X_1, X_2, \cdots, X_n$ :
- définies sur le même espace probabilisé $(\Omega, \mathcal{B}, P)$
- indépendantes deux à deux
- identiquement distribuées selon la même loi de Bernoulli $\mathcal{B}(1; p)$
  donc de même espérance $p$ et de même variance $p(1-p)$

Alors la variable aléatoire dite moyenne ou fréquence empirique d'échantillonnage :
$F_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ vérifie les trois propriétés suivantes :

1. $E(F_n) = p$
2. $Var(F_n) = \frac{p(1-p)}{n}$

3. Comme $0 \leq p(1-p) \leq \dfrac{1}{4}$ alors
d'après l'inégalité de Bienaimé-Tchebycheff on a :
$1 \geq P(|\ (F_n - p\ |< \varepsilon) \geq 1 - \dfrac{Var(F_n)}{\varepsilon^2} = 1 - \dfrac{p(1-p)}{n\varepsilon^2} \geq 1 - \dfrac{1}{4n\varepsilon^2}$.
donc
$\forall \varepsilon > 0 \lim\limits_{n \mapsto +\infty} P(|\ (F_n - p\ |< \varepsilon) = 1$
On dit alors que $(F_n)$ converge donc en probabilité vers $p$ ou encore que $F_n$ est un estimateur sans biais de la moyenne de $p$.
Or dans un schéma de Bernoulli à $n$ épreuves la variable aléatoire "nombre de succès" $X = \sum\limits_{i=1}^{n} X_i$ où chaque $X_i$ suit une loi de Bernoulli $\mathcal{B}(1;p)$.
$X$ suit loi binomiale $\mathcal{B}(n;p)$ et Il est peu probable au cours d'un très grand nombre d'épreuves que la fréquence $F_n$ du succès s'écarte de beaucoup de la probabilité $p$.
On démontre de même que $\dfrac{n}{n-1}E((X_i - F_n)^2)$ converge en probabilité vers $p(1-p)$.
On dit que c'est un estimateur sans biais de la variance $pq$

## 4.4 La vérité sur l'intervalle de fluctuation

- En fait, l'intervalle de fluctuation de $f$ est non pas $[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}]$ mais en réalité $[p - t_\alpha\sqrt{\dfrac{p(1-p)}{n}}; p + t_\alpha\sqrt{\dfrac{p(1-p)}{n}}]$ où $t_\alpha = 1,96$.
C'est le nombre venant de la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0;1)$ correspondant au risque d'erreur $\alpha = 5\%$
- Mais comme lorsque $0 \leq p \leq 1$ alors en étudiant les variations de la fonction définie par $f(p) = p(1-p)$ on prouve que $0 \leq p(1-p) \leq \dfrac{1}{4}$
- Donc comme $1,96 \approx 2$ et que $0 \leq \sqrt{p(1-p)} \leq \dfrac{1}{2}$ alors
$1,96\sqrt{\dfrac{p(1-p)}{n}} \leq 2\dfrac{1}{2}\sqrt{\dfrac{1}{n}} = \sqrt{\dfrac{1}{n}}$ donc l'intervalle de fluctuation de $f$ est en fait inclus dans l'intervalle $[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}]$

## 4.5 De l'intervalle de fluctuation à l'intervalle de confiance

- On va maintenant estimer une probabilité $p$ inconnue à partir de la connaissance de $f$.
Comme $p - \dfrac{1}{\sqrt{n}} \leq f \iff p \leq f + \dfrac{1}{\sqrt{n}}$ et que $f \leq p + \dfrac{1}{\sqrt{n}} \iff f - \dfrac{1}{\sqrt{n}} \leq p$
Alors un échantillon étant pris au hasard sa fréquence $f$ étant connue, on a 95% de chances pour que $p \in [f - \dfrac{1}{\sqrt{n}}; f - \dfrac{1}{\sqrt{n}}]$.
- Ceci n'est pas tout à fait vrai car si l'on ne connaît pas $p$ l'on ne connaît pas non plus $\sqrt{p(1-p)}$. Il faudrait amener un correctif à cette fourchette mais en première analyse, on peut considérer que pour un échantillon assez grand cette fourchette est acceptable.

- Le langage est ici mal adapté : l'idée de 95% de chance induit une idée de probabilité égale à $0,95$. Or il ne peut pas s'agir d'une probabilité car la fréquence obtenue à partir d'un échantillon correspond à un événement réalisé
- Pour être rigoureux, en statistiques, on parle alors pour $p$ **d'intervalle de confiance** : $[f - \frac{1}{\sqrt{n}}; f - \frac{1}{\sqrt{n}}]$ avec un risque d'erreur de 5%

# 5 Bibliographie


- Bulletin 434 APMEP Mai-Juin 2011 : *Entre réel et virtuel, la simulation en statistiques*,Bernard Egger
- Article : *L'épistémologiste traque la hasard*- Daniel Schwartz - Pour la Science - Dossier Hors série Le Hasard - Avril 1996
- *Jeux Mathématiques - Spécial Elections* - Hors série n° 47 - La Recherche Février 2012