SCI03 - Analyse de données expérimentales

Echantillonnage

Thierry Denœux[1]

[1]Université de Technologie de Compiègne
tél : 44 96
tdenoeux@hds.utc.fr

Automne 2014

utc Université de Technologie Compiègne

Example introductif

- Dépenses en dollars US de 50 clients consécutifs dans une épicerie.

```
> x
[1] 26.04 17.15 28.76 16.55 14.52 61.57 6.90 13.67 11.63 69.49 39.28 11.34 15.01
20.58 21.13 12.95 43.97
[18] 2.32 10.26 14.55 12.66 18.22 13.72 9.45 18.30 18.71 64.30 37.52 34.76 27.07
20.89 33.26 20.91 31.99
[35] 19.55 14.35 19.54 15.33 30.54 29.15 52.36 45.58 63.85 32.82 36.22 8.04 23.85
6.61 33.80 40.80

> stem(x)
The decimal point is 1 digit(s) to the right of the |
0 | 27789
1 | 01233444555577889
2 | 00111146799
3 | 123345689
4 | 146
5 | 2
6 | 2449
```

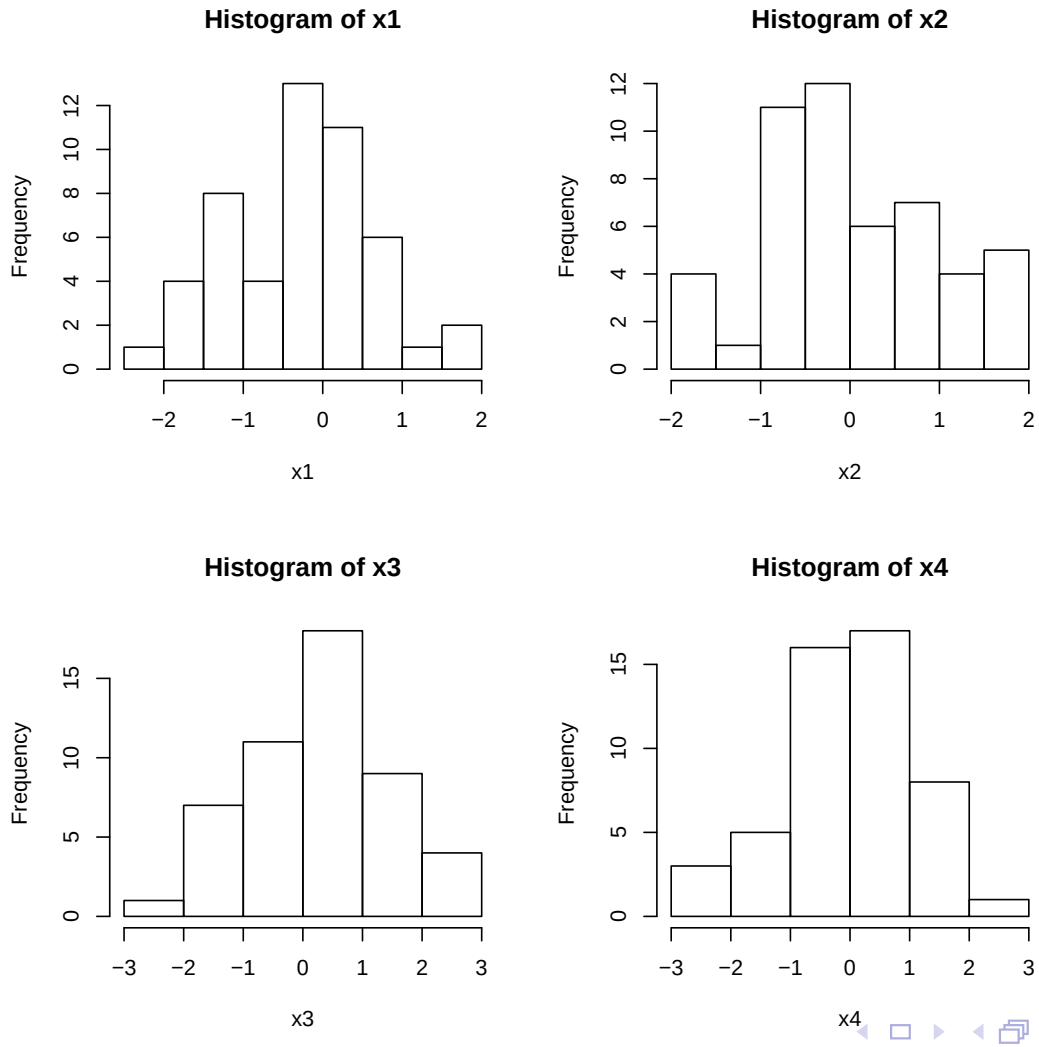
- Ces données ont été générées "au hasard".
- On dit qu'elles sont une réalisation d'un échantillon aléatoire.

Echantillon iid

- Dans l'exemple précédent, chaque observation x_i peut être considérée comme une réalisation d'une v.a. X_i.
- Les observations $x_1, \ldots, x_n$ sont une réalisation d'un vecteur aléatoire $X_1, \ldots, X_n$.
- Si on procède à un nouvel échantillonnage, on observera une autre réalisation $x'_1, \ldots, x'_n$.
- Si les v.a. X_i sont indépendantes et de même loi qu'une variable X, on dit que le vecteur $X_1, \ldots, X_n$ est un échantillon indépendant et identiquement distribué (iid), de variable parente X.
- Remarque : on distingue l'échantillon aléatoire $X_1, \ldots, X_n$ (vecteur aléatoire) et une réalisation de l'échantillon $x_1, \ldots, x_n$ (vecteur de réels).

Génération d'un échantillon iid en R

```
x1<-rnorm(50,mean=0,sd=1)
```



Statistique

- Soit $t = g(x_1, \ldots, x_n)$ un indicateur numérique calculé à partir d'une distribution empirique $x_1, \ldots, x_n$ de n observations (par exemple : $t = \overline{x}$, $t = s^2$, $t = \hat{f}_\alpha$, etc.).
- Si les observations $x_1, \ldots, x_n$ sont considérées comme des réalisations d'un échantillon $X_1, \ldots, X_n$, t est une réalisation d'une v.a. :

$$T = g(X_1, \ldots, X_n).$$

Une telle v.a. est appelée une statistique.

Moyenne empirique

Propriétés pour n fixé

- Soit $X_1, \ldots, X_n$ un échantillon iid. de variable parente X.
- Rappel : la moyenne empirique de l'échantillon est définie par

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

- Si $\mathbb{E}(X) = \mu$, alors $\mathbb{E}(\overline{X}) = \mu$.
- Si $\mathrm{Var}(X) = \sigma^2$, alors

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Convergence en probabilité

- La suite (X_n) converge en probabilité vers la constante $a \in \mathbb{R}$ (noté $(X_n) \xrightarrow{P} a$) si et seulement si, pour tout $\epsilon > 0$,

$$\lim_{n \to +\infty} \mathbb{P}\left[|X_n - a| \leq \epsilon\right] = 1.$$

- Interpretation : X_n tend à prendre des valeurs de plus en plus proches de a quand n tend vers l'infini.

Loi des grands nombres

- Si (X_n), $n = 1, \ldots, +\infty$, est une suite de v.a. iid d'espérance μ et de variance σ^2, alors la suite $(\overline{X}_n)$ définie par $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ vérifie :

$$\overline{X}_n \xrightarrow{P} \mu.$$

- Interprétation : si on extrait un nombre de plus en plus grand d'individus d'une population, la moyenne de l'échantillon devient de plus en plus proche de la moyenne de la population totale.

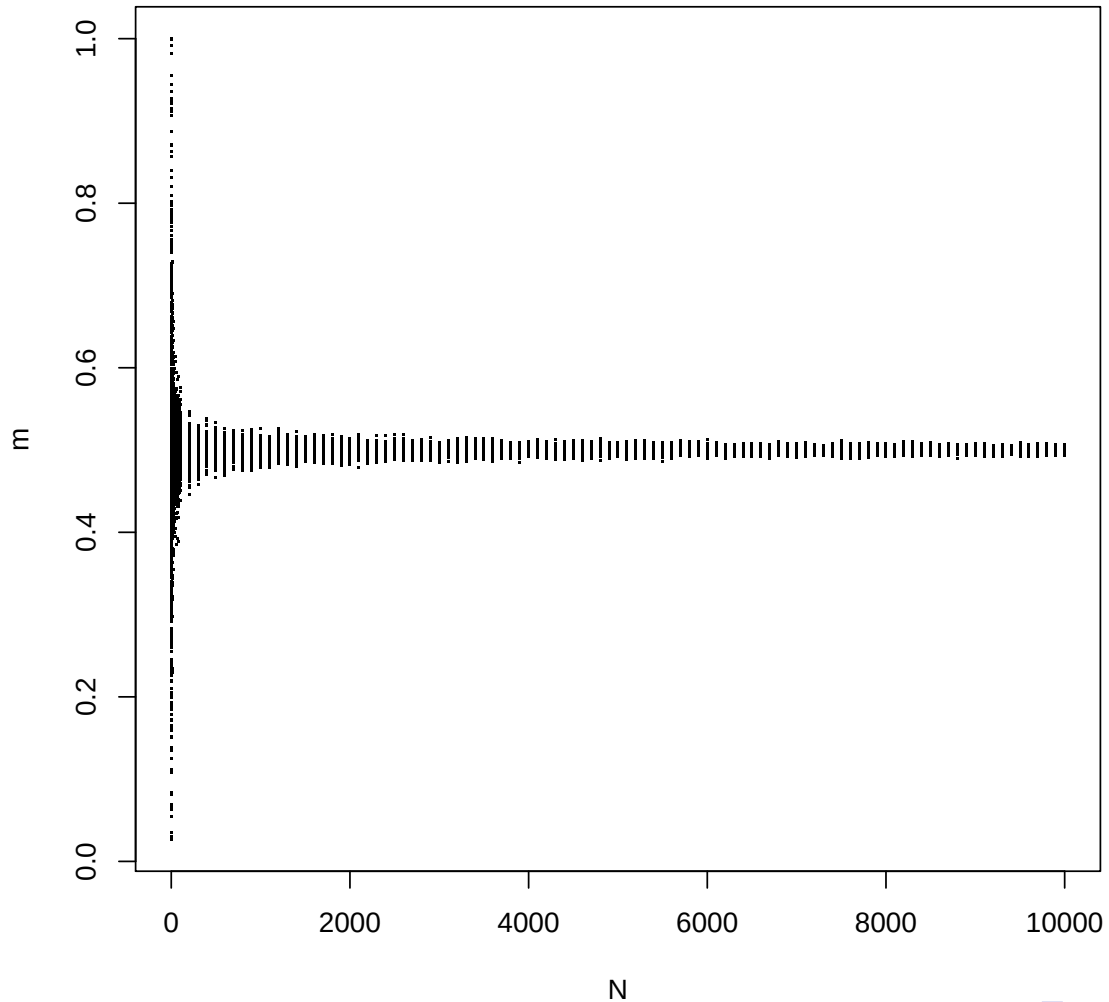
Cas particulier

- Supposons que l'on effectue n tirages avec remise dans une urne contenant une proportion p de boules blanches.
- Soit $X_i = 1$ si on a obtenu une boule blanche au i^e tirage, 0 sinon.
- On a $X_i \sim \mathcal{B}(p)$, $\mathbb{E}(X_i) = p$ et $\mathrm{Var}(X_i) = p(1-p)$.
- Dans ce cas, la moyenne $\overline{X}_n$ est la proportion de boules blanches parmi les n boules tirées.
- D'après la loi des grands nombres, on a donc $(\overline{X}_n) \xrightarrow{P} p$.
- Ce résultat justifie l'interprétation des probabilités comme limites des fréquences observées,

Example en R

```
m=0
N=0
i=0
for (n in c(1:9,seq(10,100,10),seq(200,10000,100)))
{ for (j in 1:100){
i=i+1
x=runif(n,0,1)
m[i]=mean(x)
N[i]=n }
}
plot(N,m,pch=".")
```

Exemple en R
Résultat



Théorème central limite

- Soit (X_n) une suite de v.a. iid, d'espérance μ et de variance σ^2, et $(\overline{X}_n)$ la suite de terme général $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. On a, pour tout réel x,

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) = \phi(x)$$

- Pour n grand, on a donc approximativement

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

et, de manière équivalente,

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$$

quelle que soit la loi de X, pourvu que $\mathbb{E}(X)$ et $\mathrm{Var}(X)$ existent.

Exemple d'application

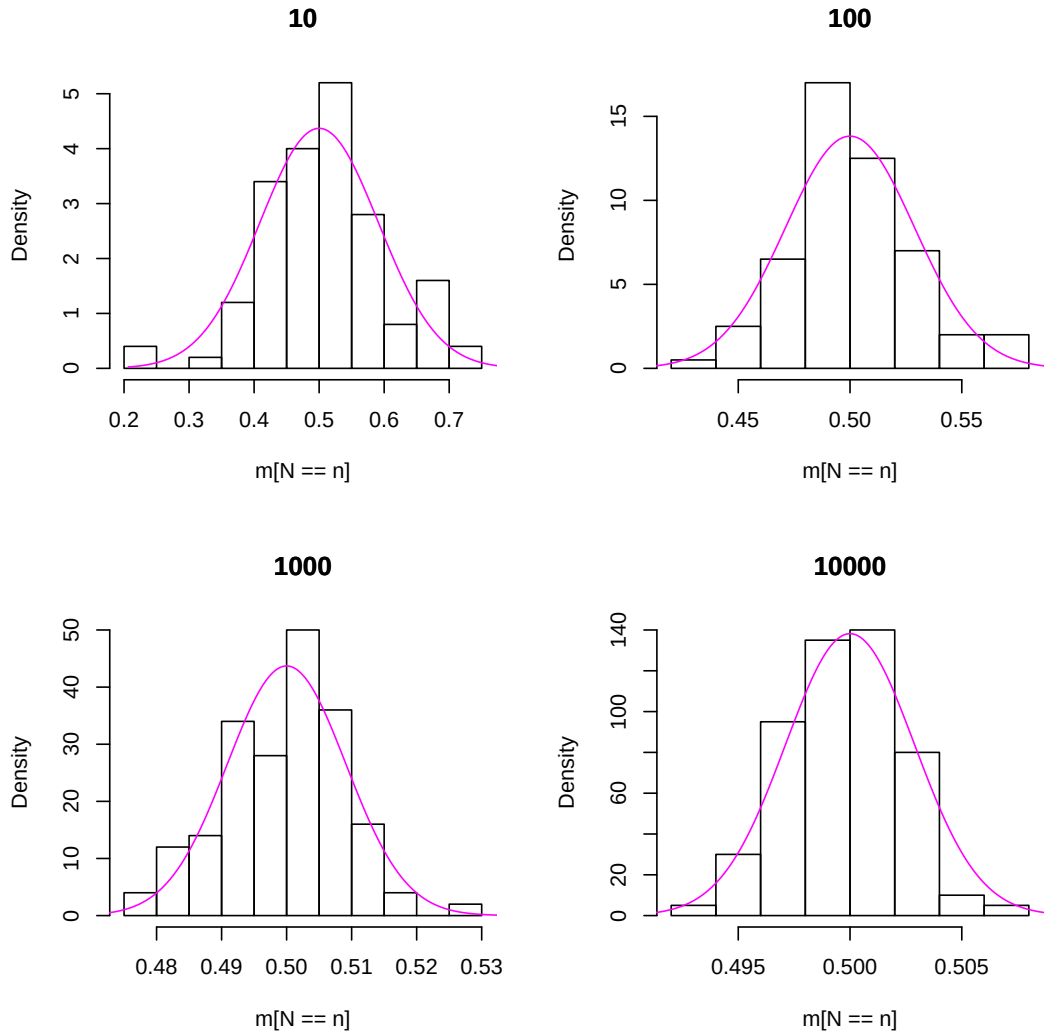
- Le poids d'un passager d'un avion avec ses bagages est une variable aléatoire X d'espérance $\mu = 70$ kg et d'écart-type $\sigma = 20$ kg. Quelle est la probabilité que le poids de 100 passagers excède 8 tonnes ?
- Soient $X_1, \ldots, X_n$ les poids de 100 passagers pris au hasard. C'est une échantillon iid.
- Soit $S = \sum_{i=1}^{n} X_i$ le poids total. D'après le TCL, S suit approximativement une loi normale $\mathcal{N}(70 \times 100, 20^2 \times 100)$.
- On obtient une approximation de la probabilité cherchée $\mathbb{P}(S \geq 8000) = 1 - \mathbb{P}(S < 8000)$ en R par

```
> 1-pnorm(8000,mean=70*100,sd=20*10)
[1] 2.866516e-07
```

Simulation en R

```
par(mfrow=c(2,2))
for (n in c(10,100,1000,10000)){
hist(m[N==n],freq=FALSE,main=n)
title(n)
a=min(m[N==n])*0.9
b=max(m[N==n])*1.1
x=seq(a,b,(b-a)/1000)
lines(x,dnorm(x,0.5,sqrt(1/(12*n))),col=550)
}
```

Simulation en R
Résultat



Approximation normale de la loi binomiale

- Soit $Y \sim \mathcal{B}(n,p)$. On a vu que Y peut s'écrire comme la somme de n v.a. indépendantes suivant une loi $\mathcal{B}(p)$: $Y = \sum_{i=1}^{n} X_i$.
- Pour n assez grand, on aura donc approximativement

$$\frac{Y - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1),$$

c'est-à-dire $Y \sim \mathcal{N}(np, np(1-p))$.

- En pratique, on admet que cette approximation est valide quand $np \geq 5$ et $n(1-p) \geq 5$. On a alors

$$\mathbb{P}(Y \leq y) \simeq \Phi\left(\frac{y - np + 0.5}{\sqrt{np(1-p)}}\right).$$

Approximation de la loi binomiale

Programme R

```
n=30
p=0.3
N=0:30
plot(N,dbinom(N,n,p))
x=seq(0,30,0.1)
lines(x,dnorm(x,n*p,sqrt(n*p*(1-p))))
```

Approximation de la loi binomiale

Résultat