ÉCOLE NORMALE SUPÉRIEURE
LETTRES SCIENCES

DÉPARTEMENT DE MATHÉMATIQUES ET APPLICATIONS
ENSEIGNEMENT

Introduction à la pensée statistique

Gilles Stoltz (chargé de recherche CNRS)

Paul Baudin (doctorant INRIA)

Table des matières

- Modèle statistique
- Notion d'estimateur
- Construction d'intervalles de confiance pour des lois de Bernoulli
- Cadre des tests d'hypothèses

- Illustrations par des simulations sous Python

- Coupures de presse et bandes dessinées

ENS | MATHÉMATIQUES | ENSEIGNEMENT | 45 rue d'Ulm 75230 Paris cedex 05 - Tél : 01 44 32 31 72 - Fax : 01 44 32 20 80 - Mél : fimfa@ens.fr

Passer de l'autre côté du miroir d'Alice :

Du calcul des probabilités au monde merveilleux de la statistique

Gilles Stoltz (chargé de recherche au CNRS)
Paul Baudin (doctorant INRIA)

En probabilités, on part de variables aléatoires $X_1, \dots X_n$ de loi connue, p.ex. de Bernoulli de paramètre $p_0 \in [0;1]$ connu, et on dit des choses sur des fonctions mesurables de $X_1 \dots X_n$. P.ex., l'inégalité de Tchebychev-Markov :

$$\forall \varepsilon > 0, \quad \mathbb{P}\left\{ |\bar{X}_n - p_0| > \varepsilon \right\} \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{p_0(1-p_0)}{n\varepsilon^2} ;$$

ie, les valeurs typiques de $\bar{X}_n$ sont proches de p_0.

En statistique, on dispose de données $x_1, \dots x_n$. Supposons (cf. ci-dessous) qu'on puisse les identifier aux réalisations de variables aléatoires $X_1 \dots X_n$ iid selon une certaine loi, dont on peut connaître la forme mais qui n'est pas entièrement déterminée. P.ex., les $X_1 \dots X_n$ sont iid selon une certaine loi de Bernoulli de paramètre $p_0 \in [0;1]$ inconnu, à déterminer.
À cause de l'inégalité de Tchebychev-Markov, on se dit que ce p_0 inconnu est sans doute proche de p_0.

Résumé :

Probabilités : loi connue $\rightsquigarrow$ comportement (d'une fonction) de variables aléatoires

Statistique : comportement de variables aléatoires $\rightsquigarrow$ loi inconnue, à déterminer
(le statisticien est le détective des nombres !)

Exemple : [Les sondages d'opinion]

On interroge au hasard dans la population française un nombre n de sondés à qui l'on pose une question fermée (oui / non). Les réponses sont notées $x_1, \dots x_n \in \{0;1\}$ (1 si oui, 0 si non).

L'opinion de l'échantillon est résumée par la fréquence $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.
Que dire de l'opinion de la population ? Une proportion notée p_0 de la population dirait oui : p_0 est inconnue, on ne la connaîtrait que si on interrogeait les 60 millions de Français. (Mais on va voir que p_0 peut être connue typiquement à $\pm 3\%$ rien qu'en interrogeant 1000 Français au hasard !)

Echantillon $\bar{x}_n$ $\longleftrightarrow$ (liens ?) Population p_0

Après le tirage : données $x_1 \dots x_n$
Avant le tirage : variables aléatoires $X_1 \dots X_n$ (dont les $x_1 \dots x_n$ sont les réalisations)
Quitte à effectuer l'approximation d'un tirage sans remise par un tirage avec remise, on peut modéliser les $X_1 \dots X_n$ comme étant iid selon la loi de Bernoulli de paramètre p_0.

I. Notion de modèle statistique, d'estimateur.

Comment bien formuler la situation statistique précédente ?
→ Notion de modèle statistique : famille de probabilités sur un espace probabilisé $(\Omega, \mathcal{F}, (\mathbb{P}_p)_{p\in[0,1]})$ tel que sous $\mathbb{P}_p$, $X_1, X_2, \dots$ sont iid $\sim Ber(p)$

Rq : une telle construction est possible avec $\Omega = \{0,1\}^{\mathbb{N}^*}$ et le théorème d'extension de Kolmogorov.

Df : Un estimateur T_n de p est toute fonction mesurable de $X_1 \dots X_n$.
Ex : $T_n = X_1$, $T_n = 1/2$, $T_n = \bar{X}_n$.
↳ Qu'est-ce qu'un bon estimateur ?

Df : T_n est sans biais si $\forall p \in]0;1[$, $E_p[T_n] = p$

Df : (T_n) est une suite d'estimateurs convergents si
$\forall p \in]0;1[$, $T_n \xrightarrow{\mathbb{P}_p\text{-prob.}} p$, i.e : $\forall p$, $\forall \varepsilon > 0$,
$$\mathbb{P}_p\{|T_n - p| > \varepsilon\} \to 0$$

Ex :
* $T_n = X_1$ et $T_n = X_n$ sont sans biais
* $(T_n) = (\bar{X}_n)$ est une suite d'estimateurs convergents, cf. loi des grands nombres.

Estimation ponctuelle ↓

II. Estimation par intervalle.

Estimation par intervalle ↙

Df : Soit U_n et V_n deux estimateurs, fondés sur les observations $X_1, \dots X_n$ (que l'on a construits tels que $U_n \leq V_n$ sur Ω).

$[U_n, V_n]$ est un intervalle de confiance pour p au niveau $1-\alpha$, où $\alpha \in]0;1[$, si
$$\forall p \in]0;1[, \quad \mathbb{P}_p\{p \in [U_n, V_n]\} \geqslant 1-\alpha.$$

Interprétations :
- Pour avoir $1-\alpha = 100\%$, il faut prendre $[U_n, V_n] = [0;1]$.
- $[U_n, V_n]$ a une probabilité au moins $1-\alpha$ de contenir le paramètre de population p.
  ↳ si on note $[u_n, v_n]$ la réalisation de $[U_n, V_n]$, alors on est confiant (au niveau $1-\alpha$) que p est dans $[u_n, v_n]$ ... mais en réalité, p y est ou n'y est pas (il n'y a plus rien d'aléatoire).
- Souvent $[U_n, V_n] = [T_n \pm E_n]$ où T_n et E_n sont des estimateurs
  ↳ E_n est la marge d'incertitude de l'estimateur T_n.

Constructions d'intervalles de confiance :

1/ Par Tchebychev - Markov :

$$\forall p,\ \forall \varepsilon > 0,\quad \mathbb{P}_p\{\,|\bar{X}_n - p| > \varepsilon\,\} \leqslant \frac{p(1-p)}{n\varepsilon^2}$$

version probabiliste

en prenant ε tq. $\frac{p(1-p)}{n\varepsilon^2} = \alpha$

$$\forall p,\quad \mathbb{P}_p\left\{\,|\bar{X}_n - p| \leqslant \sqrt{\frac{p(1-p)}{n\alpha}}\,\right\} \geqslant 1-\alpha$$

version statistique

$$= \mathbb{P}_p\left\{\,p \in \left[\bar{X}_n \pm \sqrt{\frac{p(1-p)}{n\alpha}}\right]\right\}$$

Hélas : $\left[\bar{X}_n \pm \sqrt{\frac{p(1-p)}{n\alpha}}\right]$ n'est pas un intervalle de confiance, il faudrait que la marge d'incertitude ne dépende pas de p.

On pose $\varepsilon_n = \max_{p\in[0,1]} \sqrt{\frac{p(1-p)}{n\alpha}} = \frac{1}{2\sqrt{n\alpha}}$:

$$\forall p,\quad \mathbb{P}_p\left\{\,p \in \left[\bar{X}_n \pm 1/2\sqrt{n\alpha}\right]\right\} \geqslant 1-\alpha$$

Typiquement $1-\alpha = 95\%$, $\quad 1/2\sqrt{n\alpha} = \frac{\sqrt{5}}{\sqrt{n}} \leqslant 2{,}24/\sqrt{n}$

2/ Par inégalité de Hoeffding :

$$\forall p,\ \forall \varepsilon > 0,\quad \mathbb{P}_p\{\,|\bar{X}_n - p| > \varepsilon\,\} \leqslant 2e^{-2n\varepsilon^2}$$

version probabiliste

$$\forall p,\quad \mathbb{P}_p\left\{\,|\bar{X}_n - p| \leqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}\,\right\} \geqslant 1-\alpha$$

version statistique

(en prenant ε tq. $2e^{-2n\varepsilon^2} = \alpha$)

c'est-à-dire :

$$\forall p,\quad \mathbb{P}_p\left\{\,p \in \left[\bar{X}_n \pm \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}\right]\right\} \geqslant 1-\alpha$$

Avec $1-\alpha = 95\%$:

$$\sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}} \leqslant \frac{1{,}36}{\sqrt{n}}$$

3/ Méthode la plus standard : Théorème limite central + lemme de Slutzky

On admet (cf. simulations sous Python) que

$$\forall p \quad \frac{\sqrt{n}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\left(\bar{X}_n - p\right) \xrightarrow[\text{sous } \mathbb{P}_p]{\mathcal{L}} \mathcal{N}(0,1)$$ loi normale standard

de sorte que $\forall p$, $\mathbb{P}_p\left\{-1.96 \leq \frac{\sqrt{n}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\left(\bar{X}_n - p\right) \leq 1.96\right\}$

$= \mathbb{P}_p\left\{ p \in \left[\bar{X}_n \pm 1.96\frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\right]\right\} \xrightarrow[n\to+\infty]{} \mathbb{P}\left\{-1.96 \leq \mathcal{N}(0,1) \leq 1.96\right\} = 95\%$

$\left[\bar{X}_n \pm 1.96\frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\right]$ est un intervalle de confiance dit asymptotique (de niveau 95%)

Avantage → Marge d'erreur plus faible $1.96\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq 1/\sqrt{n}$
(et parfois bien plus petit, quand $\bar{X}_n$ est éloigné de 1/2)

Inconvénient → Caractère asymptotique.

Application → Marge d'erreur des sondages en $1/\sqrt{n}$
↳ pour un échantillon de 1000 personnes, précision sur l'estimation à ± 3% (seulement)
Les journalistes commencent à intégrer ce fait.

4/ Simulation : Comparaison des trois intervalles de confiance introduits

Pour comparer des intervalles de confiance, fixer un paramètre p_0 et comparer
- les (espérances des) marges d'erreur $\mathbb{E}_{p_0}[E_n]$
- les niveaux $\mathbb{P}_{p_0}\left\{p_0 \in [\bar{X}_n \pm E_n]\right\}$ $\overset{\text{not.}}{=} 1-\alpha_{n,p_0}$

Ces quantités sont difficiles (sauf exception) à calculer explicitement : on les estime.

On fixe N, on note $Z = \mathbb{1}_{\{p_0 \in [\bar{x}_n \pm E_n]\}}$
et on simule N copies indépendantes de Z :

$Z_1 \dots Z_N$ dont on fait la moyenne $\bar{Z}_N$

C'est un estimateur de $1-\alpha_{n,p_0}$ (avec une marge d'erreur à $\pm 1/\sqrt{N}$).

III. Tests d'hypothèses : Pourquoi deux assertions (un peu) contradictoires peuvent simultanément être tenues pour vraies en statistique.

Deux hypothèses (choisies selon le contexte ... par un être humain → subjectivité) :

- Une hypothèse de départ H_0, que l'on conservera sauf si les données la contredisent gravement ;
- Une hypothèse alternative H_1, vers laquelle on se tournera dans ce dernier cas.

↳ Dissymétrie forte entre les rôles des deux hypothèses, H_0 a tendance à être conservée.

Exemple de test pour $H_0 : p_0 = p_{ref}$ contre $H_1 : p_0 \neq p_{ref}$

- Conserve H_0 si $p_{ref} \in [\bar{x}_n \pm E_n]$, intervalle de confiance à 95%
- Rejette H_0 sinon.

Application numérique : $\bar{x}_n = 51\%$ et $n = 1000$ ↳ intervalle $[51\% \pm 3\%]$

* Les hypothèses $H_0^{(1)} : p_0 = 49\%$
et $H_0^{(2)} : p_0 = 51\%$
sont toutes deux conservées ! Art de la double négation :

« Au vu de ces données, on ne peut pas affirmer que p_0 ne soit pas égal à $\begin{cases} 49\% \\ 51\% \end{cases}$ » (on énonce une certaine compatibilité, une possibilité).

Il faut faire attention à ne pas effectuer de glissement sémantique ici...

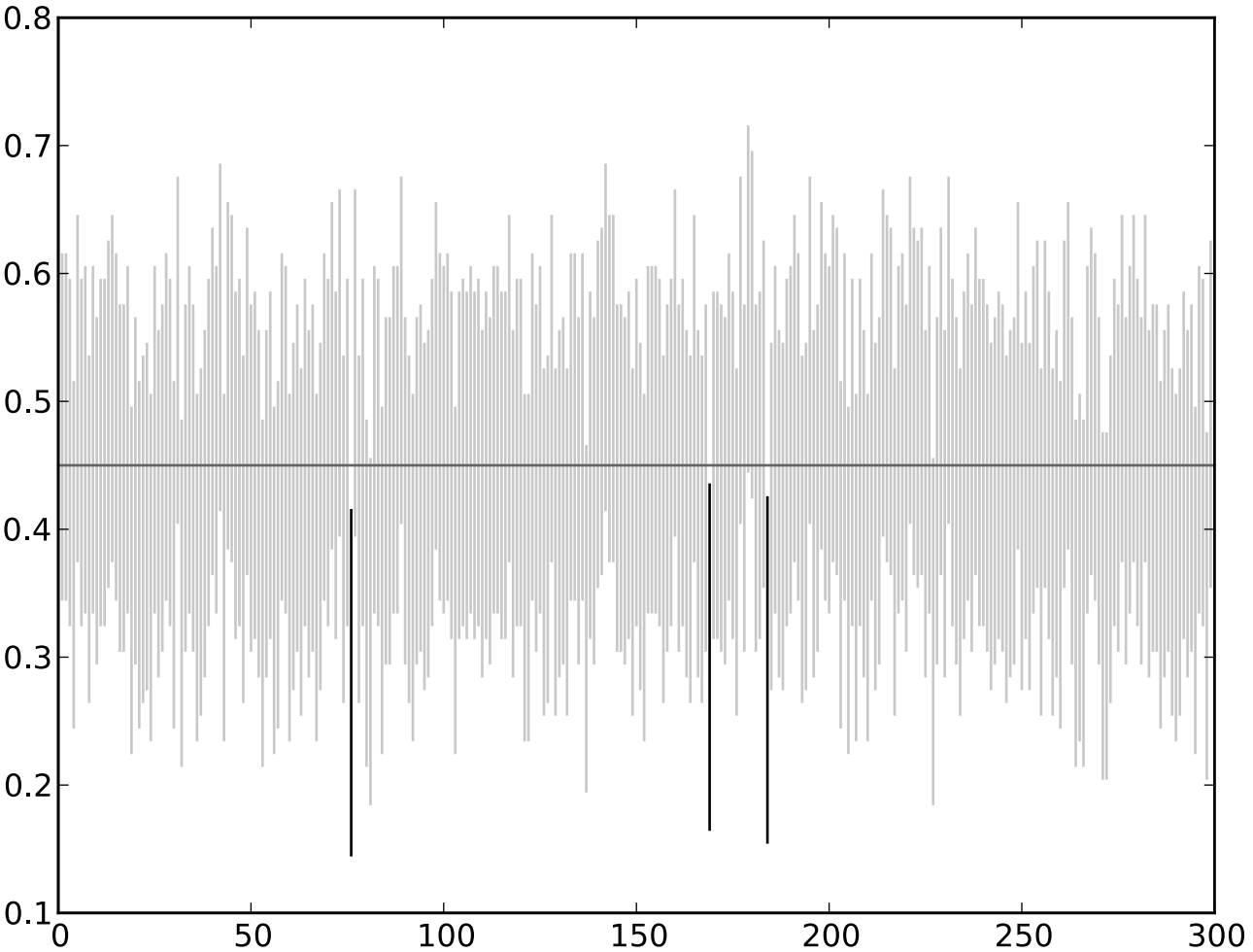
* En revanche l'hypothèse $H_0^{(3)}$: $p_0 = 47\%$ est rejetée
« Ces données montrent que p_0 est différent de 47 % »
(on énonce une impossibilité).

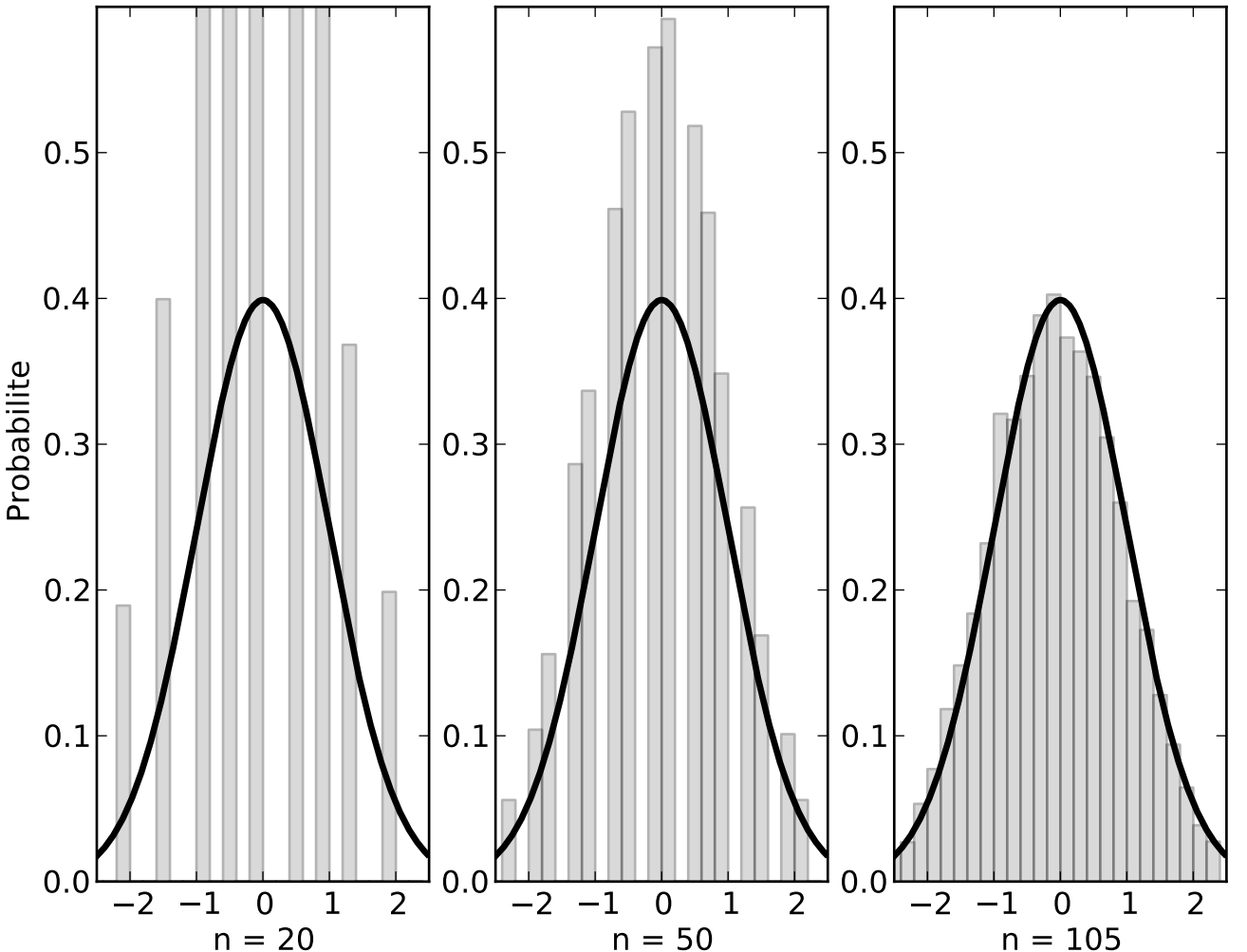
→ En statistique, H_0 est vraie jusqu'à preuve du contraire,
- soit parce qu'elle est vraiment vraie,
- soit parce l'écart du paramètre sous-jacent p_0 à la valeur de référence p_{ref} est faible et que notre taille d'échantillon n ne permet pas de la mettre en évidence.

Un test :
- lorsqu'il conserve H_0 dit simplement qu'elle n'est pas impossible ou invraisemblable
- lorsqu'il rejette H_0 fait un progrès négatif : il exclut H_0 (et passe à H_1)

Remarque culturelle :

Les débats sur l'innocuité ou la nocuité des OGM semblent contradictoires ; en fait, dans ces dernières, les hypothèses H_0' de nocuité ou H_0'' d'innocuité sont simplement conservées (parfois / souvent pour cause de tailles d'échantillon faibles).


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300


Probabilite
0.0
0.1
0.2
0.3
0.4
0.5
−2
−1
0
1
2
n = 20
n = 50
n = 105

Comparaison des performances des trois intervalles de confiance

Pour $p = 0.45$ et $n = 100$, avec $N = 10\,000$ réplications :

```
methode Hoeffding :
 marge d'erreur 0.136
 niveau de confiance 99.40

methode Plug-in :
 marge d'erreur 0.097
 niveau de confiance 94.51

methode Tchebychev :
 marge d'erreur 0.224
 niveau de confiance 100.00
```

Code pour simuler des intervalles fondés sur l'inégalité de Hoeffding

```
## MODULE IMPORTS

from numpy.random import binomial
from  numpy import *
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab

## FUNCTION AND PARAMETERS

def histogram_bernoulli(n, p, N):
    t = (random.binomial(n, p, size = (N))/n )
    return t

p = 0.50
n = 100.
N = 300
alpha = .05

## CONFIDENCE INTERVALS FOR BERNOULLI WITH HOEFFDING'S INEQUALITY

x = histogram_bernoulli(n, p, N)

half_width =  sqrt(log(2. / alpha) / (2. * n))
x_IC = [x - half_width, x + half_width]

c = np.arange(N)
list_out = c[((x_IC[0] > p) | (x_IC[1] < p))]
list_in = c[((x_IC[0] < p) & (x_IC[1] > p))]

## FIGURES

fig = plt.figure()
ax = fig.add_subplot(111)

if list_out.size >0:
    ax.vlines(list_out, x_IC[0][list_out], x_IC[1][list_out], color = 'r')

ax.vlines(list_in, x_IC[0][list_in], x_IC[1][list_in], color = 'g')
ax.axhline(y = p)

plt.savefig("Real_confidence.pdf")
```

Code pour comparer les performances des trois intervalles de confiance

```
## IMPORT

from numpy.random import binomial
from numpy import *
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab

## FUNCTION AND PARAMETERS

def histogram_bernoulli(n, p, N):
    t = (binomial(n, p, size = (N)) / n )
    return t


def half_width(n, alpha, method):

    if (method == 'Hoeffding'):
        t = sqrt(log(2. / alpha) / (2. * n))

    elif (method == 'Tchebychev'):
        t = 1. / (2. * sqrt(n * alpha))

    else:
        t = 0

    return t

# Half width of the interval confidence for the CLT and plug-in
# is empirical and would be computed outside the function.

def real_level(x_IC, p):
    N = len(x_IC[0]) * 1.
    c = arange(N)
    list_out = c[((x_IC[0] > p) | (x_IC[1] < p))]
    t = ( 1. - list_out.size / N ) * 100.
    return t

N = 10000.
n = 100.
p = 0.45
alpha = 0.05
```

```
## CONFIDENCE INTERVALS

x = histogram_bernoulli(n, p, N)

shaper = [array([-1.]), array([1.])]
results = {'Hoeffding' : [0.,0.], 'Tchebychev' : [0.,0.]}

# The Bernoulli samples are the same for the three
# interval confidence computations.

for i in results:
    half_width_spe = half_width(n, alpha, i)
    x_IC_spe = x + half_width_spe  * (shaper * ones(x.shape))
    real_level_spe = real_level(x_IC_spe, p)

    results[i] = [half_width_spe, real_level_spe]


# TCL Plug in

half_width_plugin = 1.96 * sqrt((x*(1-x)/n))
x_IC_plugin = x + (shaper * half_width_plugin)
real_level_plugin = real_level(x_IC_plugin, p)

results['Plug-in'] = [mean(half_width_plugin), real_level_plugin]

## DISPLAY RESULTS

for i in results:
    print "methode % s : " % i
    print "marge d'erreur  %.3f " % results[i][0]
    print "niveau de confiance %.2f\n" % results[i][1]
```

Code pour justifier l'approximation par une loi normale pour construire le troisième intervalle

```
## MODULE IMPORTS

from numpy.random import binomial
from numpy import *
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab


##  FUNCTIONS AND PARAMETERS
random.seed(100)

def data_bernoulli(n, p, N):
    mean_trial = binomial(n, p, size = (N))/n
    t = sqrt(n) * (mean_trial - p) / sqrt( mean_trial * (1. - mean_trial))
    return t

p = 0.45

mu = 0
sigma = 1

n = [20., 50., 105.]
N = 1e4 * ones(3)

x = [data_bernoulli(n[i], p, N[i]) for i in range(len(n)) ]


## HISTOGRAMS

num_bins = arange(-2.6, 2.6, .2)
fig = plt.figure()

for i in range(3):
    print i
    ax = fig.add_subplot(131 + i)
    ax.hist(x[i], num_bins, normed=1, facecolor='gray', alpha = 0.30)
    bincenters = arange(-2.5, 2.5, 0.01)
    y = mlab.normpdf( bincenters, mu, sigma)
    l = ax.plot(bincenters, y, 'k-', linewidth=2.5)
```

```
    if i==0:
        ax.set_ylabel('Probabilite')
        ax.set_xlabel('n = 20')
    if i==1:
        ax.set_xlabel('n = 50')
    if i==2:
        ax.set_xlabel('n = 105')
    ax.set_xlim(-2.5, 2.5)
    ax.set_ylim(0, 0.6)
    ax.grid(False)

plt.show(patches)
plt.savefig("Histogram_plug-in.pdf")
```

minimares

● **Noté**, ce titre du « Monde » (22/8) : « Niches fiscales : M. Sarkozy annonce 10 milliards d'euros de hausses d'impôts ». Même avec les niches, il faut appeler un chat un chat.

● **Ce titre**, dans « Les Echos » (20/8) : « Voté il y a trois ans, le paquet fiscal est resté un talon d'Achille pour la majorité ». Un talon menacé de rupture.

● **Le « JDD »** (22/8) consacre un quart de page à expliquer la hausse de 2 points de la cote de Sarko, selon l'Ifop. Manque de bol, un sondage Viavoice pour « Libé » (23/8), réalisé aux mêmes dates, le donne en baisse de 1 point. Il en faudrait, des tiers de page, pour expliquer les contradictions entre sondeurs !

● **Cours de fiscalité** donné par le secrétaire général de l'UMP, Xavier Bertrand (« Le Parisien », 21/8) : « Baisser la réduction d'impôts de certains, ce n'est pas augmenter les impôts pour tous. » Ni toucher

● **Ségolène Royal** a
rope 1, 21/8) les p
Sarko sur la sécurit
fie d'« agitation frén
cace ». Seulement

● **La même Ségolèn**
Figaro », 18/8) : « J'
sur moi et mon pèr
les hommes, donc
voir, donc moi fille
forcément castratri
que « je » parle de

● **De Vincent Peillon**
péen PS (« Le Par
« On a laissé se con
velle oligarchie, mé
fonctionnaires pass
nets ministériels, liés
litique, faisant des fo
dans le privé. » Pei
parler aux anciens
cabinets Fabius ou
aujourd'hui salariés
nault.

● **En marge** des journ

mbe !

omne. C'est
t à l'intéressé,

FIGURE 1 – Pan sur le bec du Canard ! Les variations à $\pm 3\%$ des cotes de popularité ne sont pas statistiquement significatives.

FIGURE 2 – Il est facile de briller à peu de frais... Un peu d'efforts (cf. formules vues ensemble) et vous saurez faire mieux que ce mauvais détective !

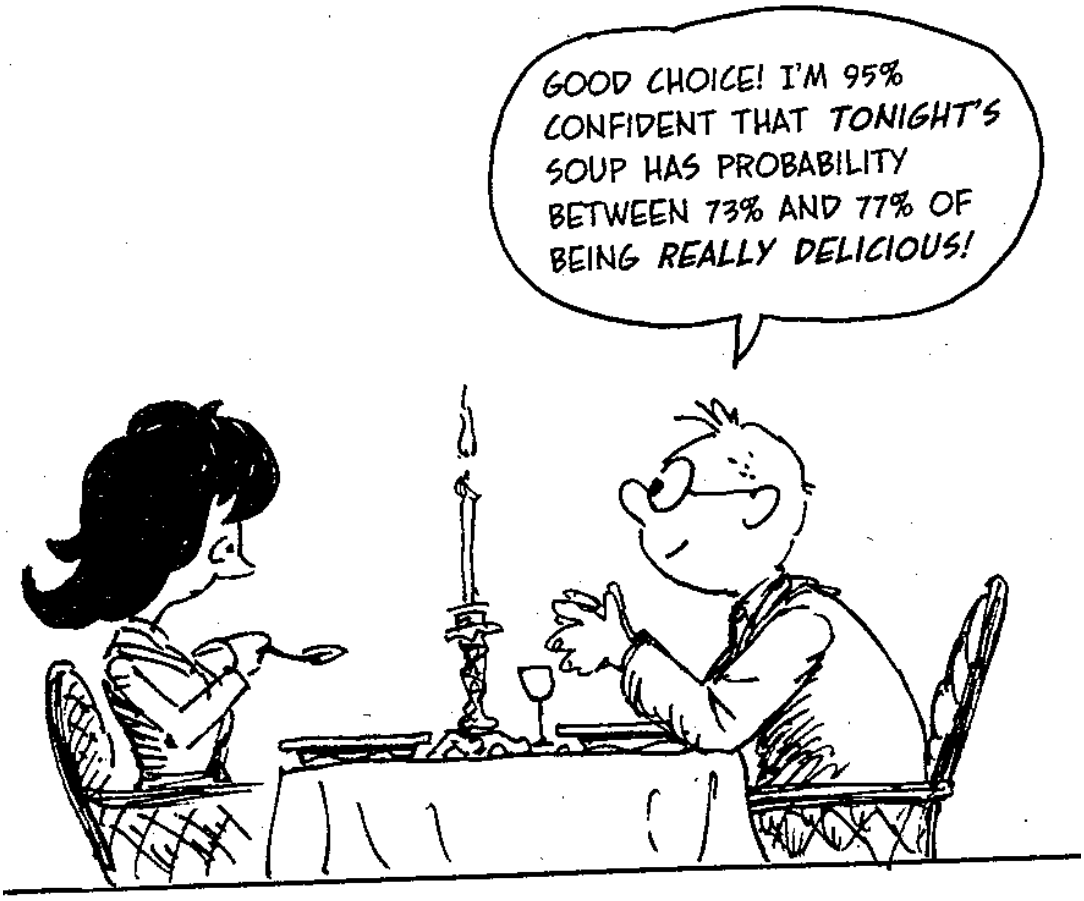


FIGURE 3 – La statistique donne des intervalles de confiance. C'est utile, mais pas nécessairement romantique.

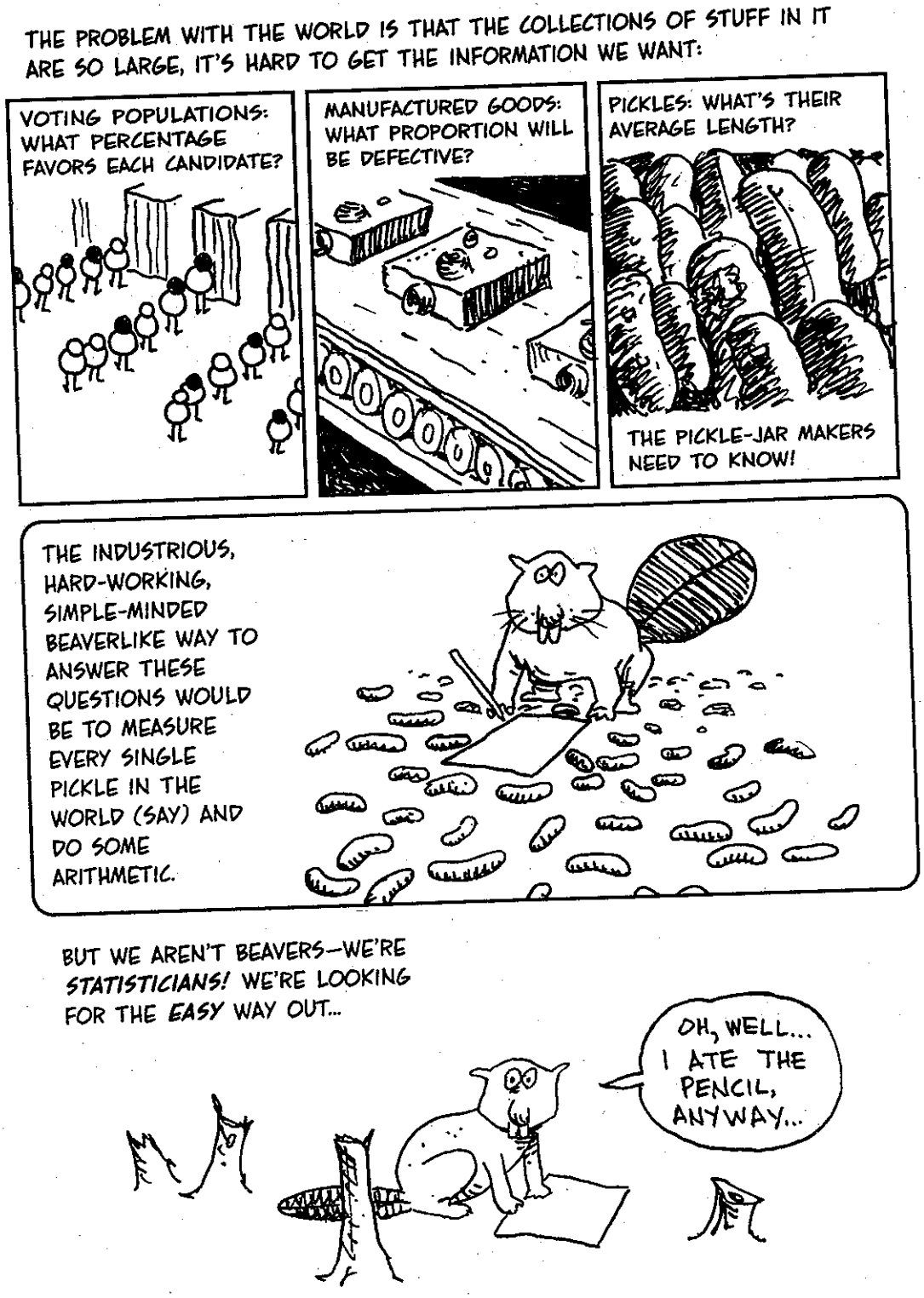
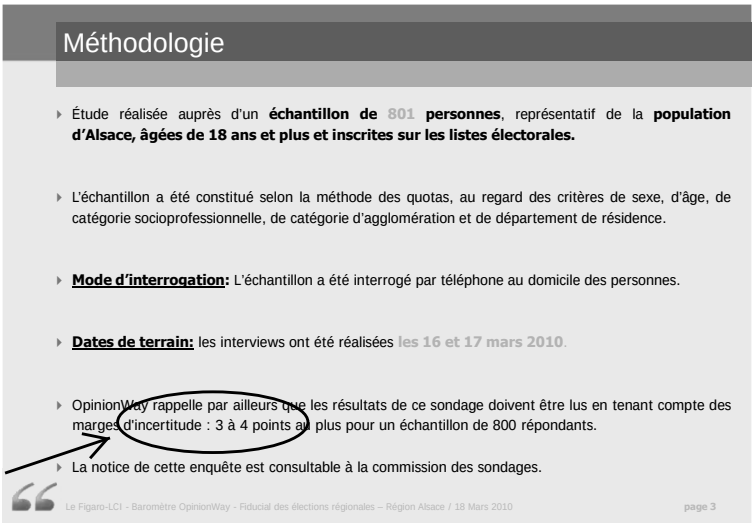


FIGURE 4 – Vous n'êtes pas des petites bêtes industrieuses : en conséquence de quoi, vous ferez désormais des sondages plutôt que des décomptes exhaustifs sur des populations immenses.



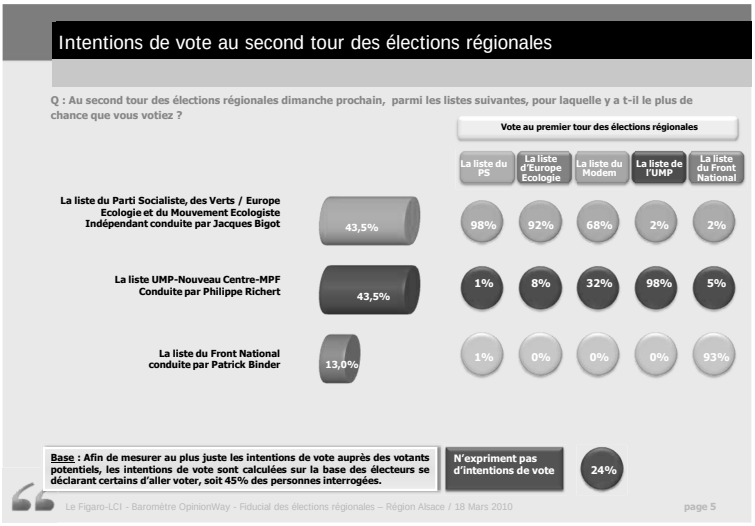


FIGURE 5 – Un exemple de résultat de sondage fourni par un institut avec indication de la marge d'erreur (d'incertitude) possible.

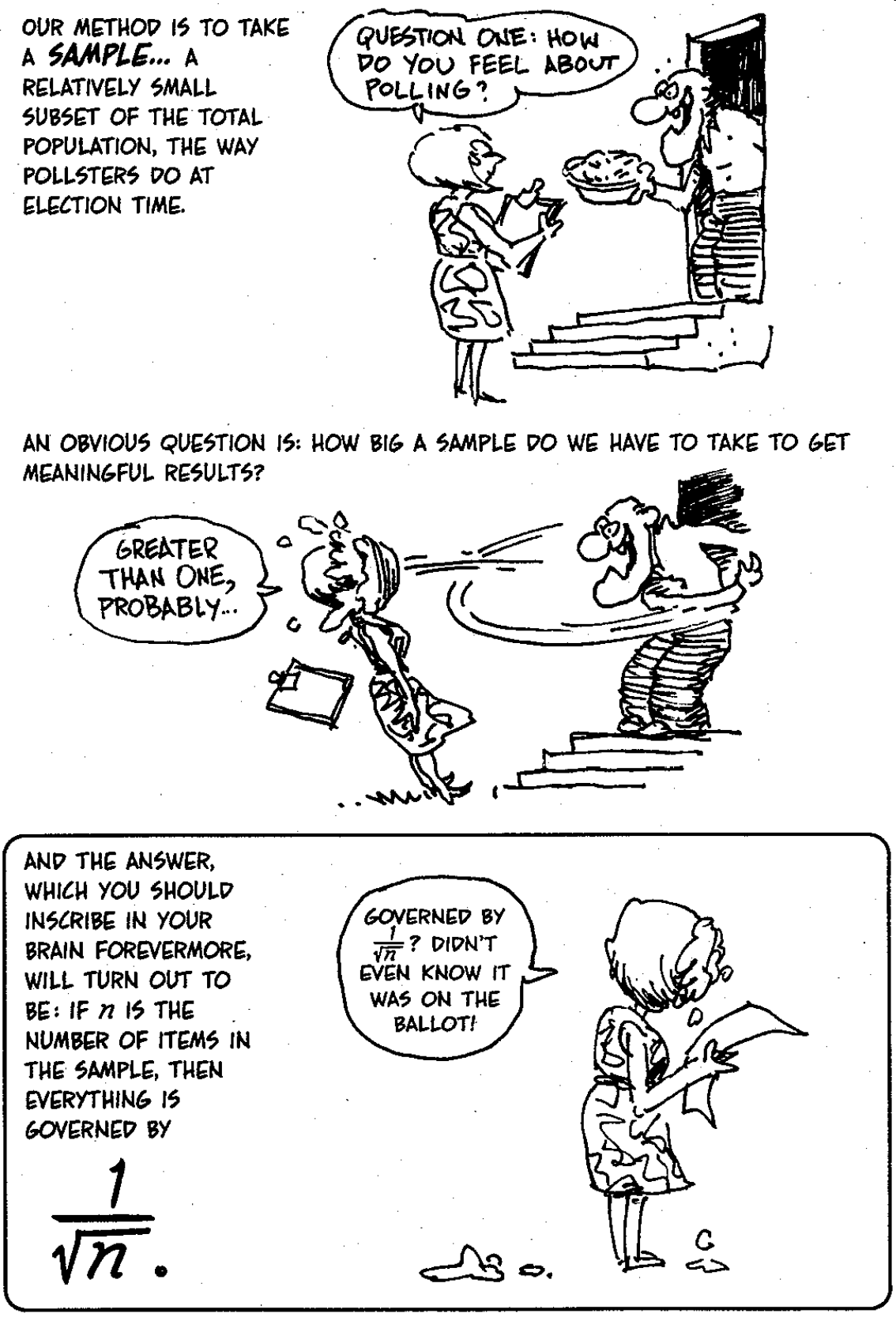


FIGURE 6 – Planification de sondages et mauvais jeux de mots associés.