Où sont les cloches de Gauss ?

Valérie Fontanieu
Statisticienne,
Institut National de
Recherche Pédagogique.

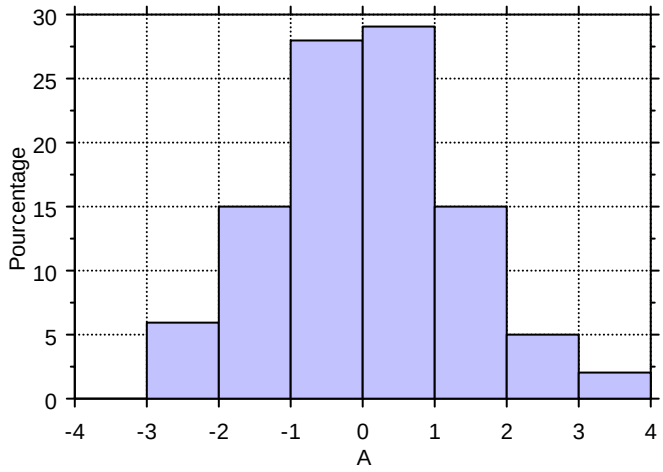
On dit que certaines mesures répétées donnent lieu à un histogramme « en forme de cloche ». C'est pourquoi, dans des manuels scolaires de physique, de biologie ou de mathématiques, on suggère parfois aux élèves de « voir une forme de cloche » émerger d'un histogramme, là où eux voient une montagne, un tas, mais vraiment, pas de cloche. L'objet cloche leur est d'ailleurs peu familier. Les cloches en fonte des cours d'école… se sont envolées il y a longtemps.

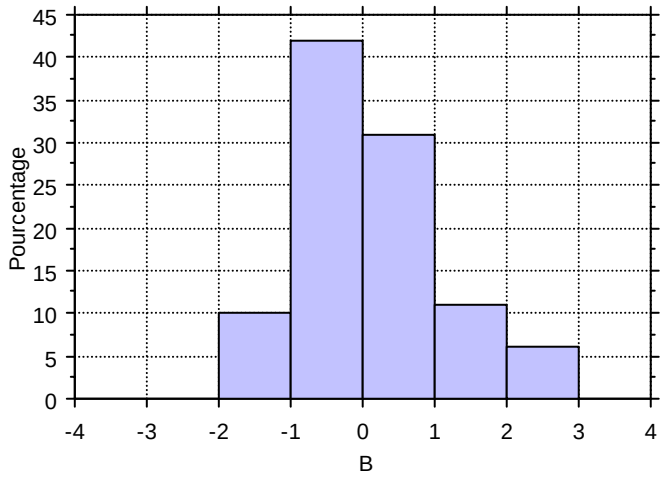
Mais que signifie « une forme de cloche » ? De quoi est-ce spécifique ?

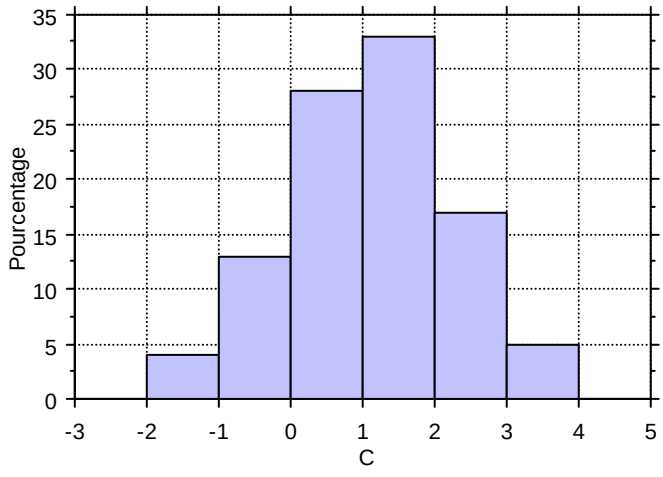
Des exemples

Voici 100 observations de 3 variables A, B et C et un histogramme de chacune de ces séries de données.

Ont-ils une forme de cloche ?



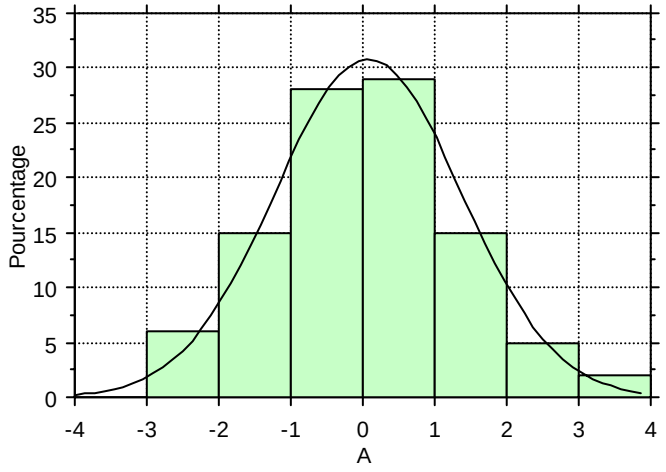


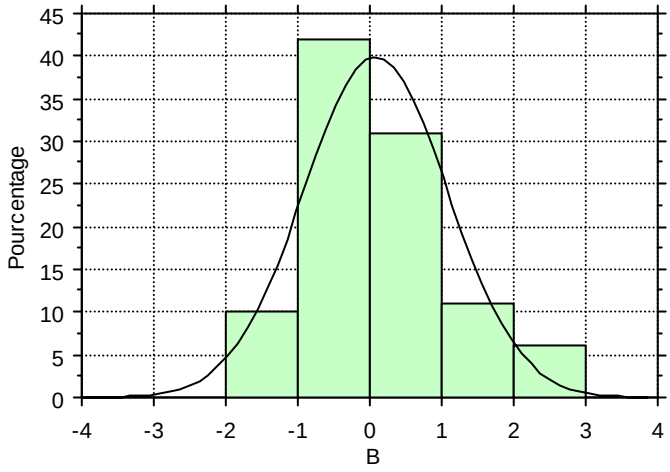


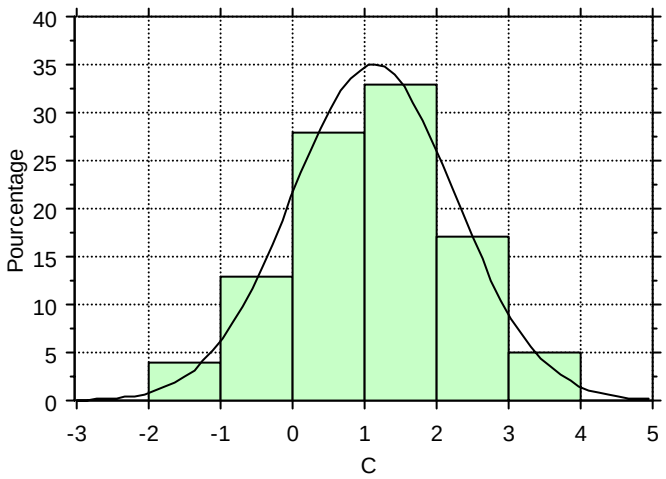
Cette question est vague : c'est quoi une forme de cloche ? De quoi parle-t-on ? Il y a toutes sortes de cloches…

En statistique, cela signifie que les données peuvent être considérées comme un échantillon d'une loi de Gauss. Pour s'en faire une meilleure idée, on superpose à l'histogramme la courbe représentative de la densité d'une loi de probabilité de Gauss dont les paramètres sont calculés, pour bien « coller » aux données (voir annexe 1) ; ces courbes matérialisent la *forme de cloche* en jeu.

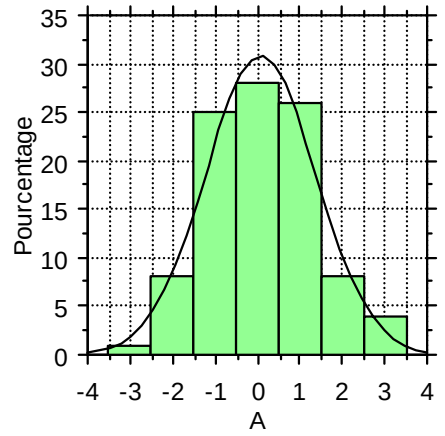
La question précédente peut être reposée ainsi : les histogrammes ci-dessous s'ajustent-ils bien à une courbe en cloche ? On a interrogé quelques personnes : les avis sont partagés, surtout pour les variables B et C, moins pour A où la forme en cloche semble acceptée.

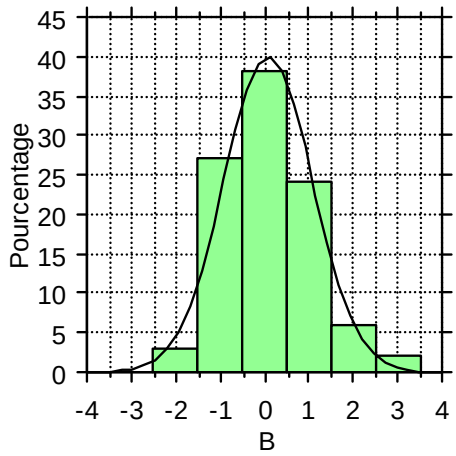


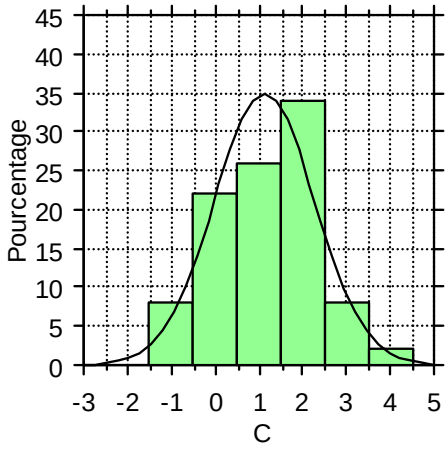




On nous a suggéré un autre découpage des classes, reproduit ci-dessous, qui a conduit à contester radicalement, au niveau visuel, le cas de C. On est toujours dans l'observation graphique…

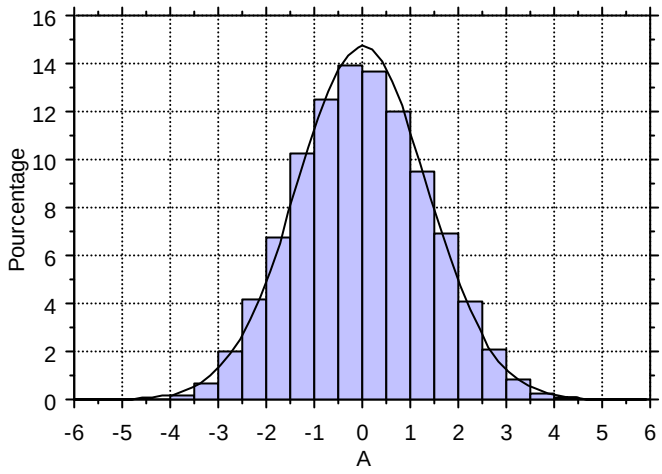


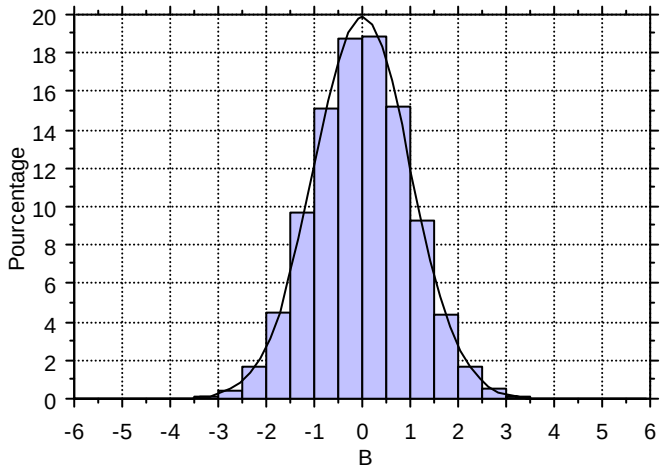


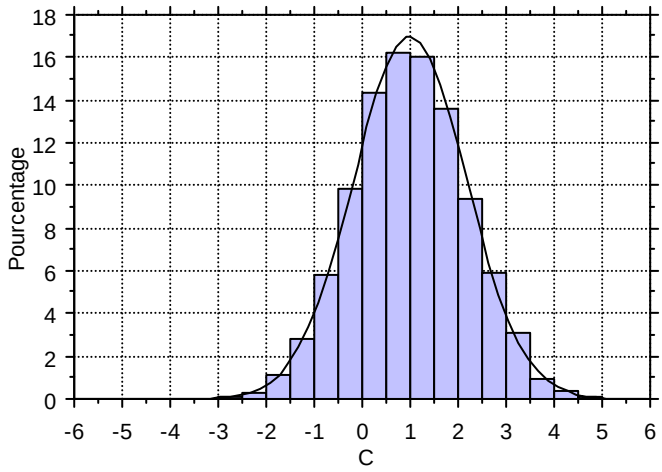


Cent observations, c'est déjà beaucoup, mais voyons ce qui se passe avec plus de données. Les séries traitées ci-dessus sont les 100 premières valeurs d'échantillons de taille 10 000, construits à partir de simulations. Les histogrammes ci-dessous correspondent aux 10 000 valeurs simulées.

Que penser de ces histogrammes ? A l'œil nu, cloche ou pas cloche ?







Là il y a consensus. Les trois histogrammes s'ajustent agréablement aux courbes représentatives des « cloches de Gauss » : à l'œil nu, les trois variables sont « gaussiennes » ; les paramètres (moyenne et écart-type) peuvent être estimés par les moyennes et écart-type des données simulées.

	Moy.	Dév. Std	Nombre	Minimum	Maximum	Médiane
A	-,003	1,355	10000	-5,044	4,511	-,021
B	-,003	1,002	10000	-3,748	3,814	-,003
C	,996	1,172	10000	-3,364	5,027	,982

Des tests

Mais allons un peu plus loin et mettons en œuvre un test statistique qui validera ou infirmera l'adéquation à ces données d'un modèle gaussien. Un test classique pour cette situation est le test de Kolmogorov Smirnov[1]. Les données sont fournies dans le fichier Excel joint ; ceux qui le souhaitent pourront mettre eux-mêmes le test en œuvre.

Sur les trois séries de 100 données, le modèle gaussien est accepté avec un test de Kolmogorov au risque de première espèce 0,1 , bien que visuellement l'ajustement soit peu convaincant. De fait, avec peu de données, l'ajustement à une courbe en cloche est rarement flagrant à l'œil nu (voir aussi l'annexe 2).

Sur les séries de 10 000 données, le modèle gaussien est accepté avec un test de Kolmogorov au risque de première espèce 0,1 pour les variables B et C. Pour la variable A, le modèle est refusé au même risque 0,1 mais il est accepté avec le même test et un risque de première espèce 0,01 : la situation est donc ambiguë[2]. On notera qu'avec 10 000 données, l'histogramme relatif à A semblait aussi bien s'ajuster à une courbe en cloche que ceux des variables B et C : l'impression à l'œil nu est ici que tout est *en cloche* : histogrammes à une seule bosse, assez bien symétriques[3].

Accepter ou non le modèle gaussien pour A dépend de l'usage que l'on a du modèle. Ici, les données sont construites, et il convient maintenant de dévoiler comment elles l'ont été.

Pour obtenir les 100, puis 10 000 valeurs de B, on a simulé la loi normale centrée réduite.
Pour obtenir les 100, puis 10 000 valeurs de C, on a simulé 10 000 valeurs de la loi uniforme sur [0,2], que l'on a ajouté à la série B.
Pour obtenir les 100, puis 10 000 valeurs de A, on a fait le calcul suivant : la i-ème valeur a_i de la série A est égale à :

$$a_i = b_i + 1{,}3\cos(i)$$

où b_i est la ième donnée de la série B et où l'unité d'angle est le radian.

[1] http://www.math-info.univ-paris5.fr/smel/cours/cadre_cours.html
http://www.math-info.univ-paris5.fr/smel/simulations/tks/tks.html

[2] Nous avions choisi ici un risque de première espèce égal à 0,1 ; il est à ce niveau plus *fort* d'accepter le modèle que si nous avions choisi un risque 0,01. En effet, dans ce dernier cas, on veut avoir une chance sur 100 de rejeter le modèle à tort, contre une chance sur 10 ici : plus on est prudent sur le risque de se tromper en cas de rejet, plus facilement on acceptera le modèle.

[3] La courbe en cloche est à *décroissance rapide*, mais ce n'est pas facile d'en juger à l'œil nu.

Pour conclure

Seule la variable B, par construction, suit une loi gaussienne.

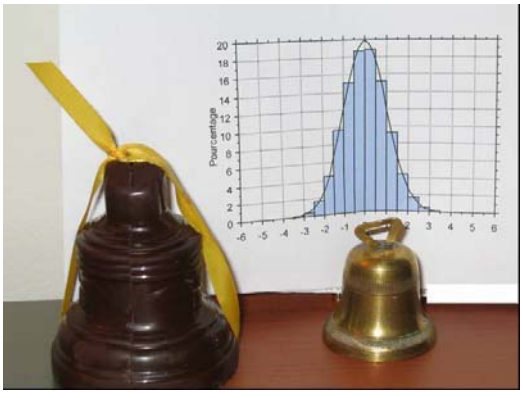
La loi de la variable C en est très proche, et même avec 10 000 données, le modèle gaussien reste acceptable[4]. La loi de la variable A est *proche* d'une loi de Gauss.

Nous avons ici construit des données dans le but de préciser l'implicite sous le terme « *histogrammes en cloche* » et pour dégager les enseignants d'une sorte d'obligation de voir des cloches même quand c'est quelque peu artificiel. Nous avons ainsi pu :

- expliciter ce qu'est ici une *forme de cloche* : c'est la représentation graphique d'une fonction (une densité de loi de probabilité) associée à un modèle des données par une loi de Gauss ;
- voir sur un exemple qu'on peut observer à l'œil nu de bons ajustements avec une loi qui n'est pas gaussienne, ou inversement, avec peu de données, considérer visuellement que l'ajustement est douteux alors que les données sont simulées avec une loi normale.

On notera qu'en pratique, on a souvent moins de 100 données et très rarement 10 000 données et la problématique est autre. L'origine des données doit toujours être prise en compte pour discuter du choix du modèle. En contrôle industriel, les recommandations classiques pour faire le test de normalité et obtenir certains labels de qualité portent sur des échantillons de 50 à 200 données. Cependant, les spécialistes savent que les tests décèlent souvent nettement moins bien les problèmes de fabrication que les vérifications du tracé chronologique des données et l'analyse précise des conditions de relevé (étude d'un procédé non maîtrisé[5], mélange de plusieurs machines, ou de plusieurs produits).

Enfin, signalons une confusion avec une autre situation où l'on parle d'*histogrammes en cloche* : il s'agit non plus d'histogrammes construits avec des données empiriques ou simulées mais d'histogrammes de lois de probabilité. En tel cas, l'ajustement est visuellement excellent et illustre graphiquement le théorème central limite (annexe 3).

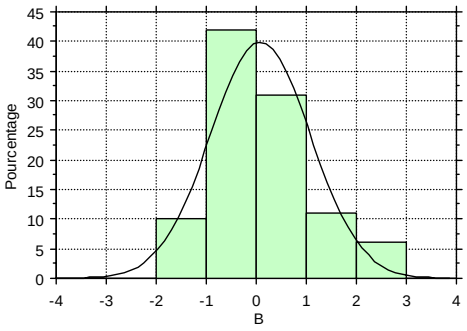


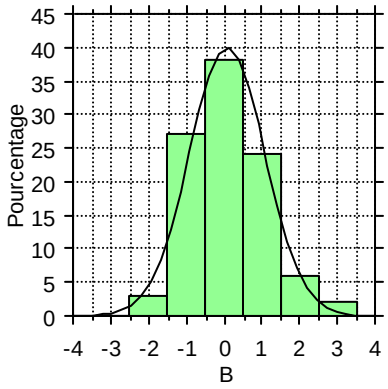
[4] N'oublions pas que face à une série de données, plusieurs modèles probabilistes sont envisageables.

[5] voir http://www.statistix.fr/spip/spip.php?article24

Annexe 1

L'équation des « courbes en cloche »





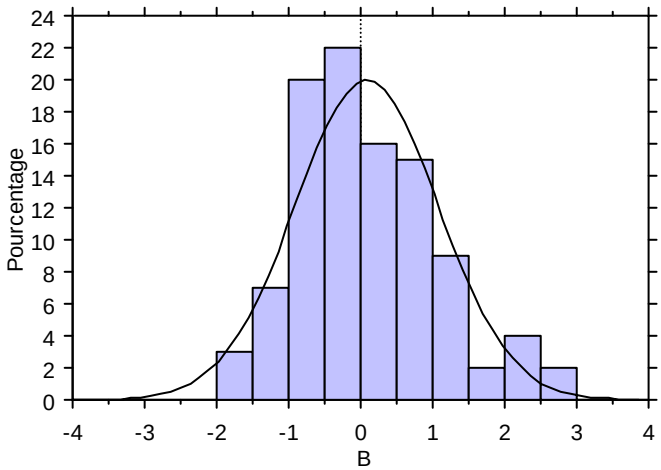
Les courbes tracées ci-dessus sont les courbes représentatives de la fonction f définie par :

$$f(x)=\frac{1}{s\sqrt{2\pi}}\exp\left(\frac{(x-m)^2}{2s^2}\right)$$, avec m=0.082 et s=0.994

(on a $f(0)\approx 0{,}40$)

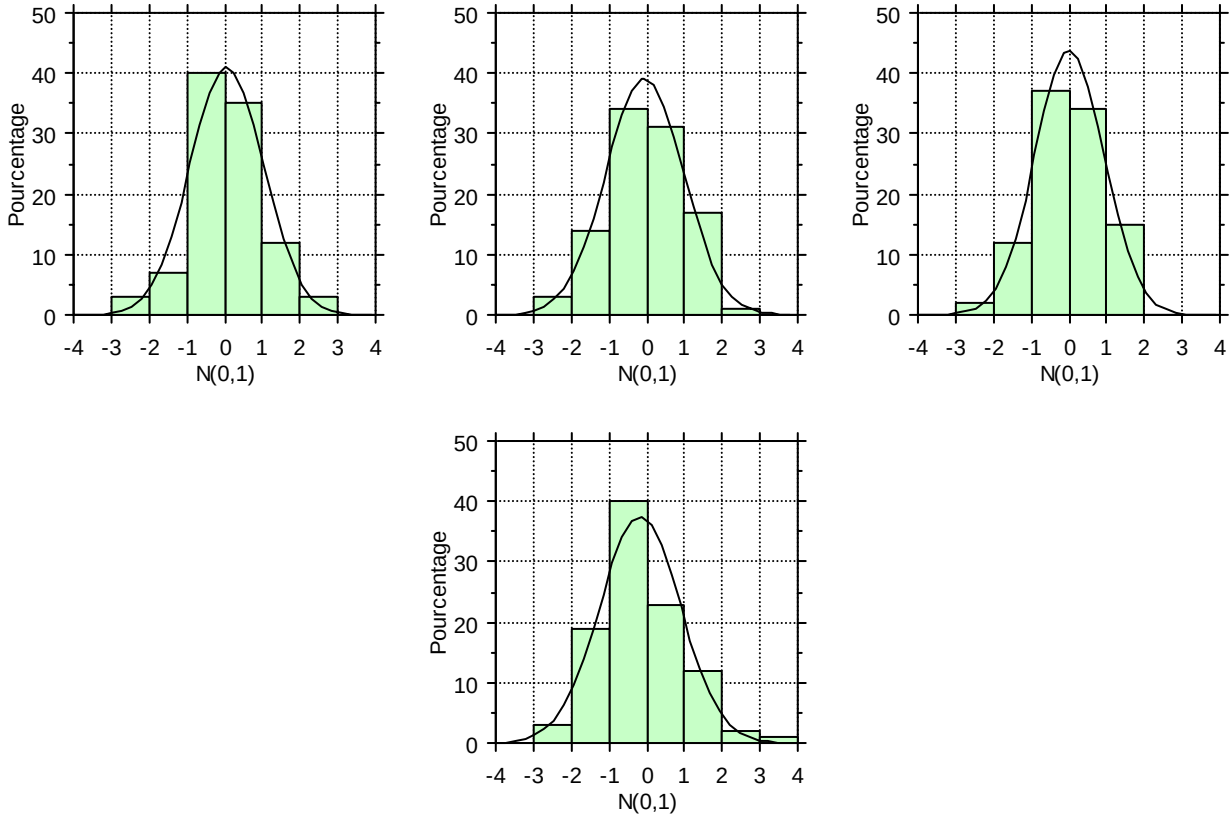
Ici, m et s sont respectivement la moyenne et l'écart type des données ayant servi à construire l'histogramme.

Si on avait tracé un histogramme avec un pas de largeur k (ce qui revient aussi à changer d'unité) on aurait tracé sur le même graphique la courbe représentative de kf. Ainsi, dans le graphique ci-dessous, on a k=0.5.



Annexe 2

Quatre autres séries de taille 100 simulées selon la loi N(0,1) et les « cloches de Gauss » associées



Annexe 3

Histogrammes théoriques et courbes en cloche

Il est difficile de « voir » des « formes de cloches » sur des histogrammes construits à partir d'un nombre *raisonnable* de données issues d'expériences réelles. Par contre, certains histogrammes théoriques *collent* bien avec une courbe en cloche.

Ci-dessous, le point d'abscisse i a pour ordonnée la valeur de la probabilité qu'une variable aléatoire de loi binomiale B(50,0.5) vaille i, $i = 0, \ldots, 50$; la courbe tracée est la courbe représentative de f, avec $f(x)=\frac{1}{s\sqrt{2\pi}}\exp\left(\frac{(x-m)^2}{2s^2}\right)$, pour $m = 25$ et $s = \sqrt{50}/2$. La raison de cet ajustement est donnée par le théorème central limite.