Comment construire un diagramme de Henry avec Excel et comment l'interpréter

Kathy Chapelain et Emmanuel Grenier

emmanuel.grenier@lasalle-beauvais.fr

Relu par Henry P. Aubert, Jacques Goupy et Jacques Vaillé

Objet

Le diagramme de Henry (ou « droite de Henry ») permet d'apprécier l'adéquation d'une distribution observée à la loi de Gauss.

En abscisse, *x*, on porte les valeurs observées ou les limites supérieures des classes lorsque les valeurs sont regroupées en classes. En ordonnée, on porte le « normit » de *x* (la fonction normit sera définie à la page 2). Sur ce système d'axes, des réalisations d'une variable gaussienne donnent un nuage de points proche d'une droite.

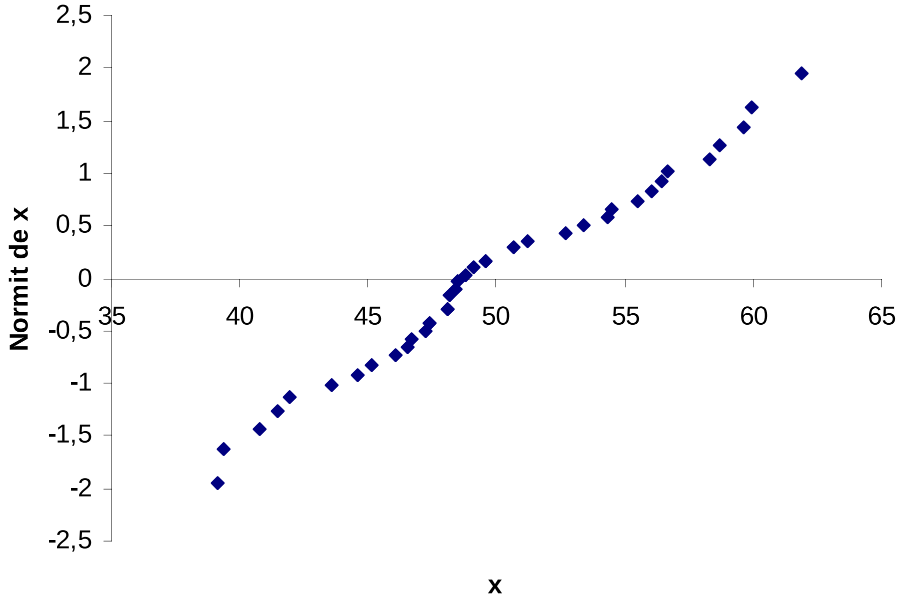


Figure 1

Pour l'exemple de la Figure 1, on peut supposer que les valeurs observées sont des réalisations d'une même variable gaussienne.

Note : le diagramme de Henry est un cas particulier de diagramme « Quantile-Quantile » (Voir les fiches « Méthodes d'ajustements graphiques » dans la page Excel'ense [2]).

La présente fiche explique comment faire un diagramme de Henry avec Excel et propose une aide à l'interprétation par des exemples. Le document joint ***Diagramme de Henry.xls*** fournit les données des exemples ainsi qu'une feuille de calcul toute faite pour vos propres données.

Principe

Nous invitons le lecteur non initié aux probabilités ou à la statistique descriptive à se reporter aux chapitres correspondants dans le manuel du groupe « Les cercles d'Excel'ense » [3].

La construction du diagramme s'opère de la manière suivante (voir la norme AFNOR « Etude de la normalité d'une distribution » [1]) :

1) On calcule l'effectif cumulé pour chaque *x* (valeur observée ou limite supérieure de la classe si les valeurs sont regroupées en classes), c'est-à-dire le nombre de valeurs inférieures ou égales à *x*.
2) On calcule la fréquence cumulée, c'est-à-dire l'effectif cumulé divisé par l'effectif total.
3) On détermine le fractile correspondant pour la loi de Gauss standard (le fractile d'ordre *p* est la valeur *u* telle que la probabilité cumulée jusqu'à *u* est égale à *p*). On obtient ainsi le « normit » de *x* (de l'anglais *<u>norm</u>al <u>unit</u>*)

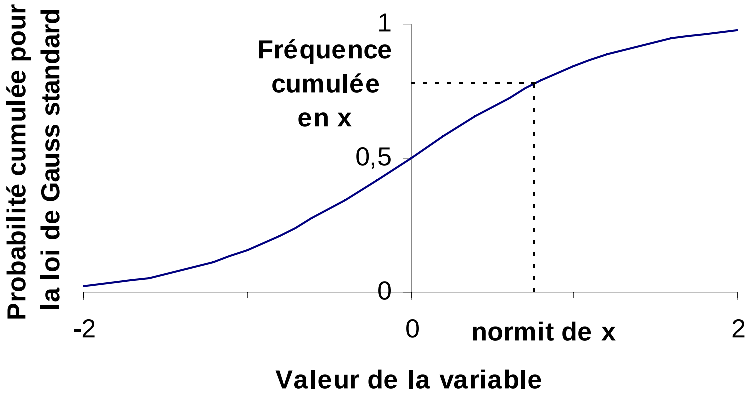


Figure 2

4) On représente les couples (*x* ; *y*), avec *y* = normit de *x*, par un nuage de points.

Prenons le cas où les valeurs observées sont des réalisations d'une variable gaussienne standard. La fréquence cumulée en *x* des valeurs observées est alors proche de la probabilité cumulée en *x* pour la loi de Gauss standard (d'autant plus proche que le nombre d'observations est important). Le normit de *x* est donc proche de *x*.

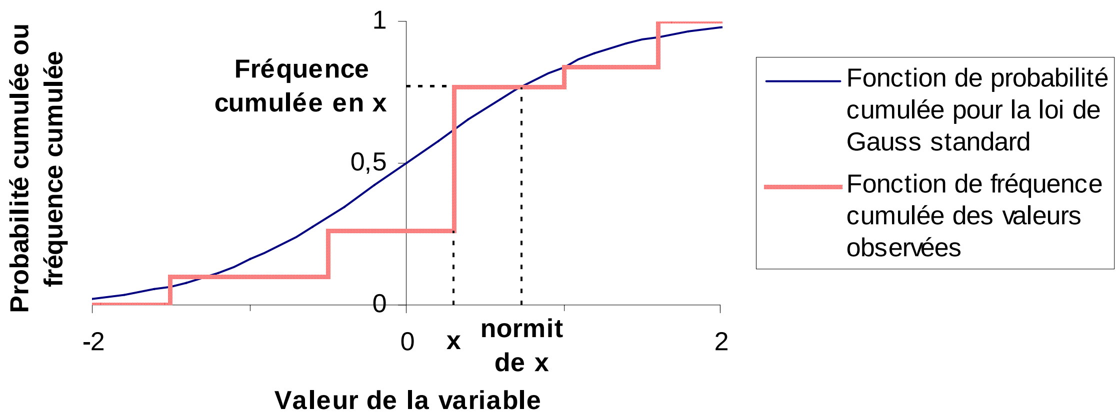


Figure 3

Le nuage des points (*x* ; *y*), avec *y* = normit de *x*, est alors proche de la droite *y* = *x*.

Prenons le cas d'une variable gaussienne de moyenne théorique μ et d'écart-type théorique σ quelconques. On revient à une variable de Gauss standard en centrant et en réduisant la variable, c'est-à-dire par la transformation $x' = (x - \mu)/\sigma$. Le nuage des points sera donc dans ce cas proche de la droite d'équation $y = (x - \mu)/\sigma$.

Construction avec Excel

Cas d'observations non regroupées en classes

Voir la feuille ***Exemple 1*** dans le document ***Diagramme de Henry.xls***

On reprend les 4 étapes décrites précédemment :

1) Calcul de l'effectif cumulé

L'effectif cumulé est le rang de classement des valeurs dans un ordre croissant : fonction **RANG**.

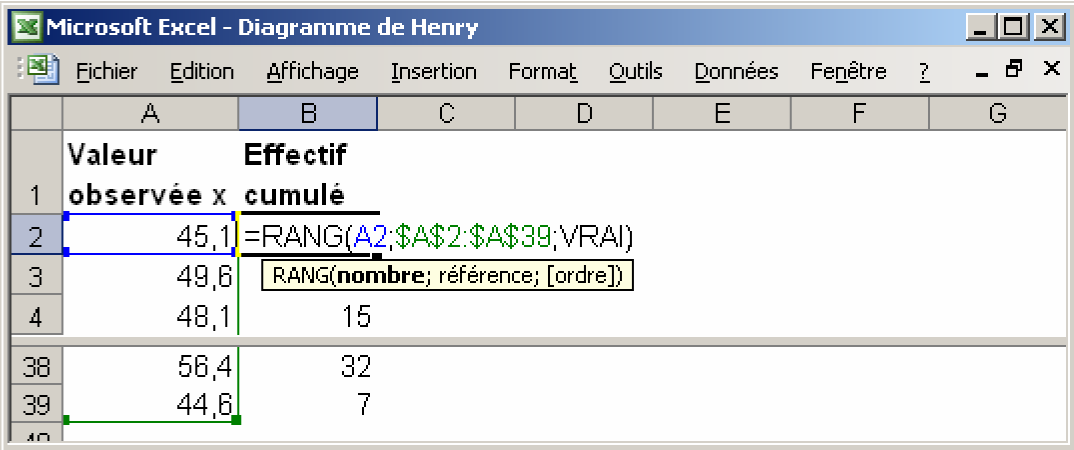


Figure 4

2) Calcul de la fréquence cumulée

On divise l'effectif cumulé par l'effectif total. L'effectif total peut être obtenu en prenant le maximum des effectifs cumulés.

En pratique, on augmente l'effectif total d'une unité pour que le normit puisse être calculé sur la valeur *x* maximale. D'autres corrections sont possibles : voir la norme AFNOR [1].

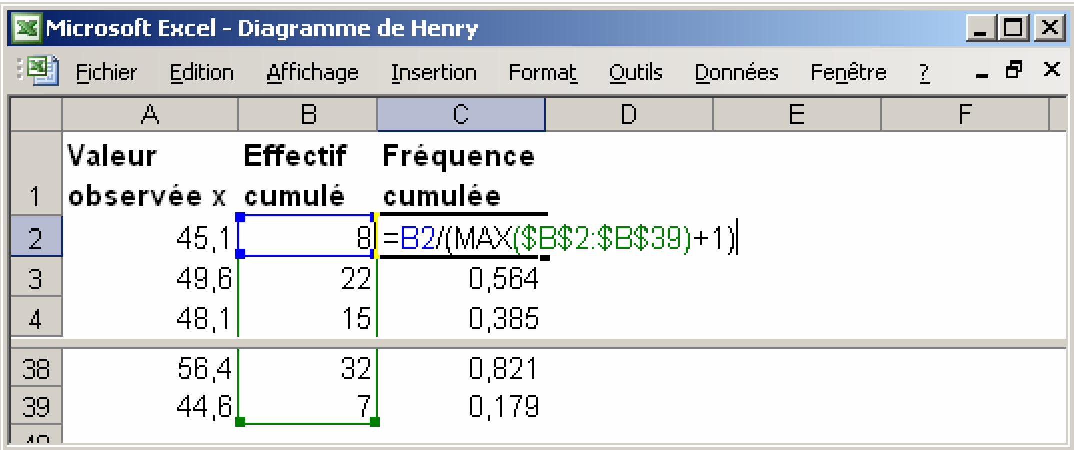
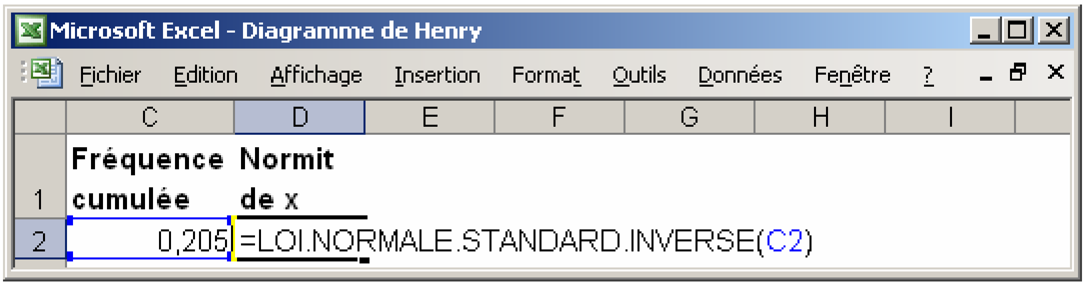


Figure 5

3) Calcul du normit

On applique aux fréquences cumulées la réciproque de la fonction de répartition de la loi de Gauss standard : fonction **LOI.NORMALE.STANDARD.INVERSE**.



	C	D	E	F	G	H	I
1	Fréquence cumulée	Normit de x					
2	0,205	=LOI.NORMALE.STANDARD.INVERSE(C2)					

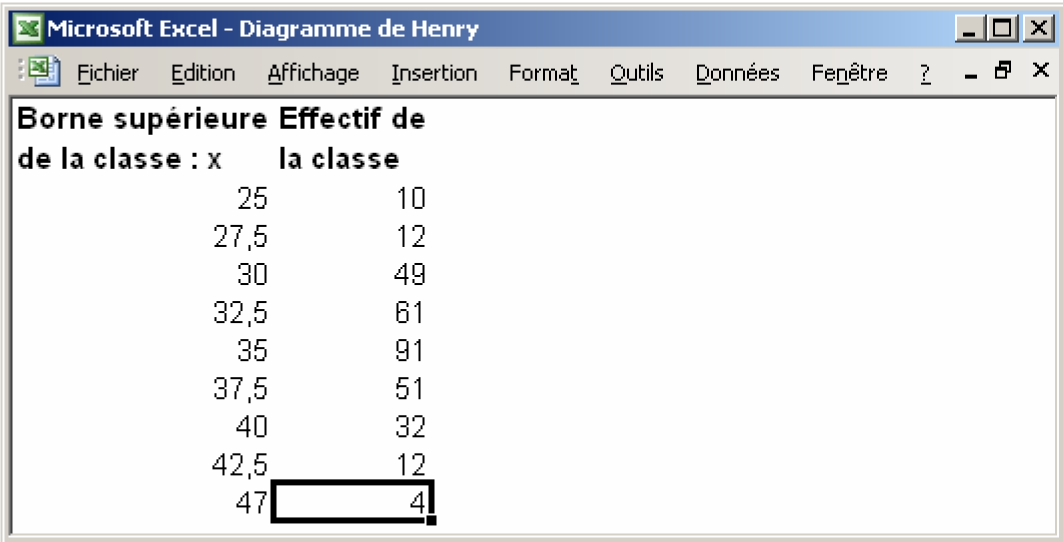
Figure 6

4) Représentation des couples (x ; y), avec y = normit de x, par un nuage de points

On sélectionne la plage des x et celle des normits (sélection de plages discontinues avec la touche **Ctrl**), on clique sur l'icône **Assistant graphique** et on choisit **Nuage de points**. Pour l'exemple, on obtient le graphique de la page 1.

Cas d'observations regroupées en classes

Les intervalles des classes doivent être fermés à droite (bornes supérieures incluses). Les valeurs x sont alors les limites supérieures des classes. On les entre par ordre croissant et on entre les effectifs des classes correspondantes. Voir la feuille ***Exemple 2***.



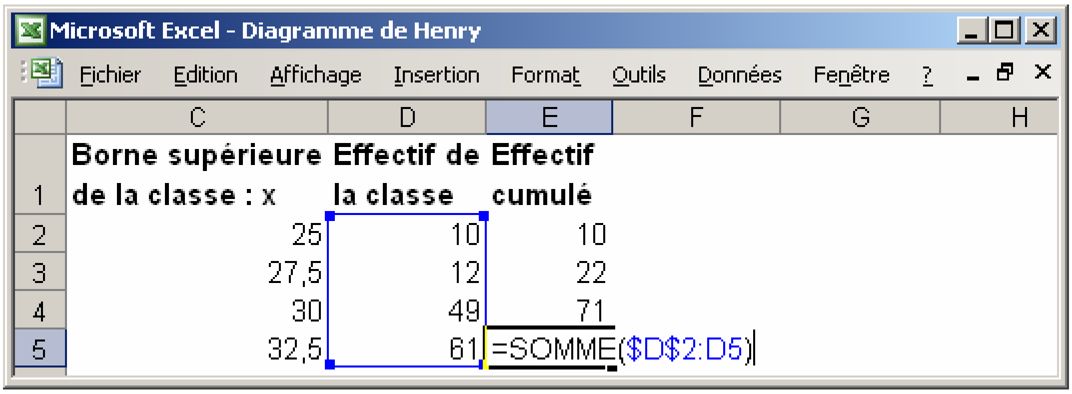
Borne supérieure de la classe : x	Effectif de la classe
25	10
27,5	12
30	49
32,5	61
35	91
37,5	51
40	32
42,5	12
47	4

Figure 7

Note :

> Dans la feuille d'exemple, les effectifs des classes n'ont pas été saisis mais calculés à partir des valeurs observées et des limites des classes. On a utilisé la fonction **FREQUENCE**. On aurait pu utiliser la fonction **NB.SI** (voir la fiche « Comment faire un histogramme » dans le manuel [3]) ou passer par l'**Utilitaire d'analyse** « **Histogramme** ».

Les effectifs cumulés se calculent en additionnant les effectifs depuis la première classe.



Microsoft Excel - Diagramme de Henry

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

	C	D	E	F	G	H
1	Borne supérieure de la classe : x	Effectif de la classe	Effectif cumulé			
2	25	10	10			
3	27,5	12	22			
4	30	49	71			
5	32,5	61	=SOMME(D2:D5)			

Figure 8

On revient ensuite à l'étape 2) de la procédure précédente.

Interprétation à partir d'exemples

Voir dans le document Excel les feuilles ***Exemple 1*** à ***Exemple 5***.

Exemple 1

L'effectif (38 observations) ne justifie pas un regroupement en classes. Représentons la distribution des valeurs observées par un nuage de points.

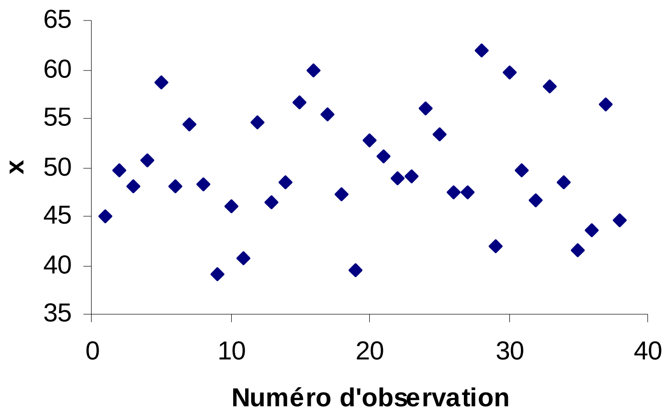


Figure 9

Peut-on faire l'hypothèse que ces valeurs sont des réalisations d'une variable gaussienne ?

On construit le diagramme de Henry. On obtient le diagramme de la Figure 1 (page 1). Le nuage étant proche d'une droite, on n'a pas de raison de rejeter l'hypothèse.

Exemple 2

On a maintenant suffisamment d'observations pour regrouper les valeurs observées en classes et représenter leur distribution par un histogramme.

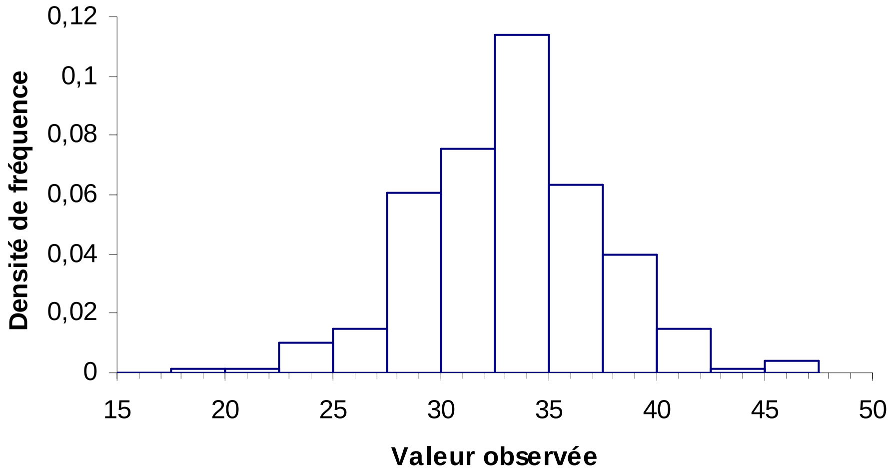


Figure 10

La distribution est symétrique, en forme de cloche. A première vue, il semblerait qu'on puisse la modéliser par une loi de Gauss. Pour le confirmer, on construit le diagramme de Henry (avec les mêmes classes mais en regroupant les classes de faible effectif).

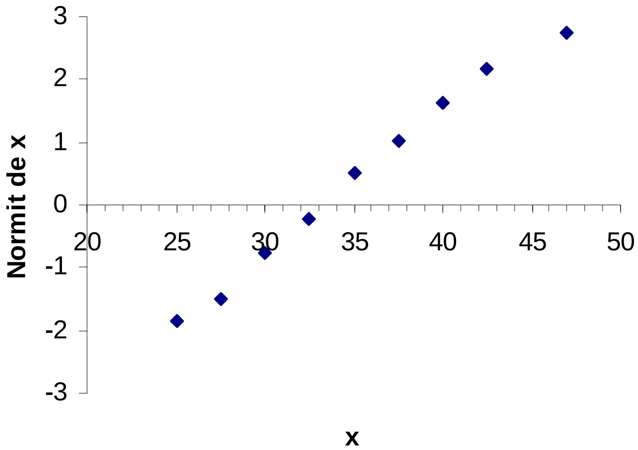


Figure 11

Le nuage de points étant proche d'une droite, on peut modéliser la distribution observée par une loi de Gauss.

Remarque :

Le diagramme de Henry permet d'approcher la moyenne et l'écart-type des réalisations d'une variable gaussienne lorsqu'on ne dispose que de la distribution des valeurs regroupées en classes. On repère la droite de tendance du nuage des points, ou « droite de Henry », et on approche la moyenne par l'intersection à l'origine, *m*, et l'écart-type par, *s*, l'inverse de la pente (voir la norme AFNOR [1]).

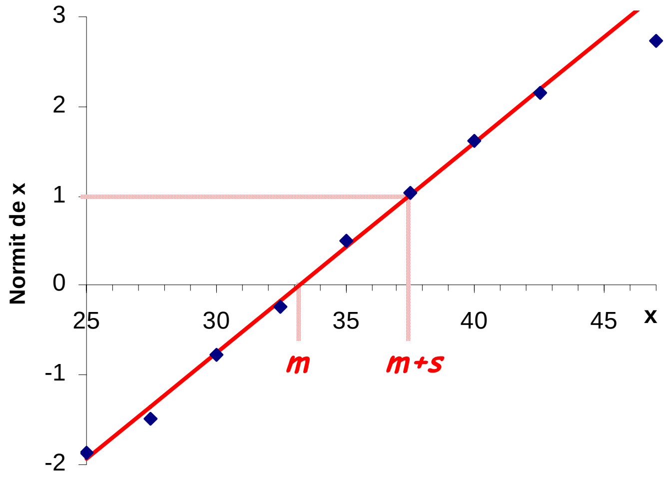


Figure 12

Pour l'exemple, on obtient m = 33,2 et s = 4,2 (valeurs très proches de la moyenne et de l'écart-type des valeurs observées).

On peut généralement déterminer la droite au jugé, les points étant quasi alignés dans le cas où les observations sont issues d'une variable gaussienne et regroupées en classes.

Attention : la fonction graphique d'Excel **Ajouter une courbe de tendance** ne donne pas la droite de tendance d'un nuage de points mais la droite de régression. La droite de tendance d'un nuage de points (x ; y) est la droite d'équation $y^* = x^*$ où y^* et x^* sont les valeurs de y et de x centrées et réduites (pour la droite de Henry, il faut pondérer la moyenne et l'écart-type par l'effectif de la classe correspondante).

Exemple 3

La distribution observée est très dissymétrique pour cet exemple.

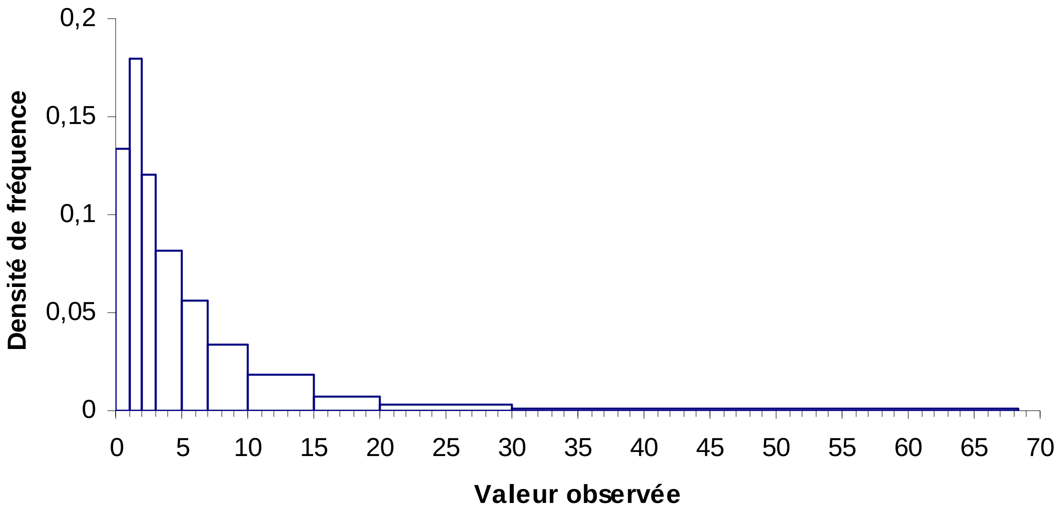


Figure 13

Cette dissymétrie se traduit sur le diagramme de Henry par une tendance non linéaire du nuage de points :

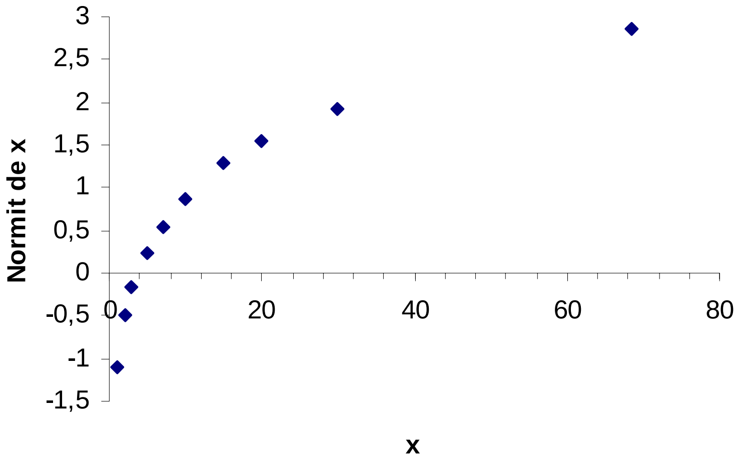


Figure 14

Une tendance dont la pente diminue avec *x* traduit une répartition dissymétrique étalée vers les valeurs importantes de *x*. La pente augmenterait dans le cas d'un étalement vers les faibles valeurs de *x*.

Remarque :

Un passage à une échelle logarithmique (Clic double sur l'axe des abscisses / **Format** / **Echelle** / Sélectionnez **Echelle logarithmique**) donne une tendance linéaire.

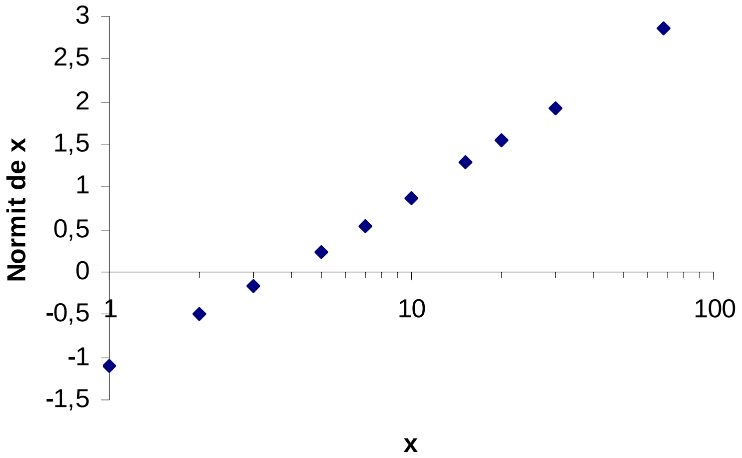


Figure 15

Pour l'exemple, on pourrait modéliser la distribution par une loi de Gauss sur les logarithmes des observations.

Exemple 4

Comme pour l'Exemple 1, on travaille sur les données non regroupées parce que l'effectif est faible.

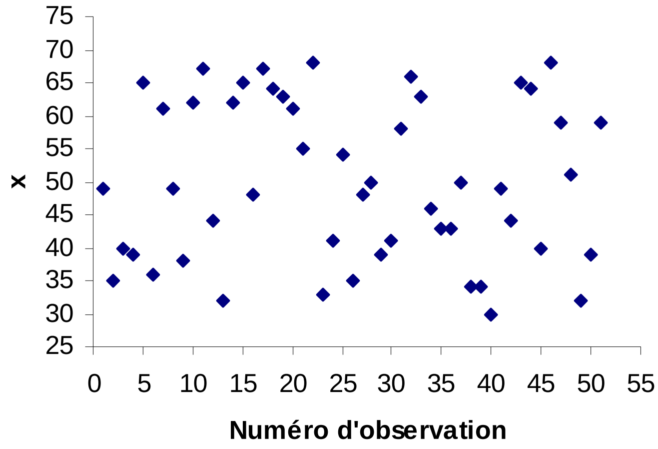


Figure 16

La distribution est symétrique. Peut-on ici aussi supposer que les valeurs observées sont des réalisations d'une variable gaussienne ? Construisons le diagramme de Henry.

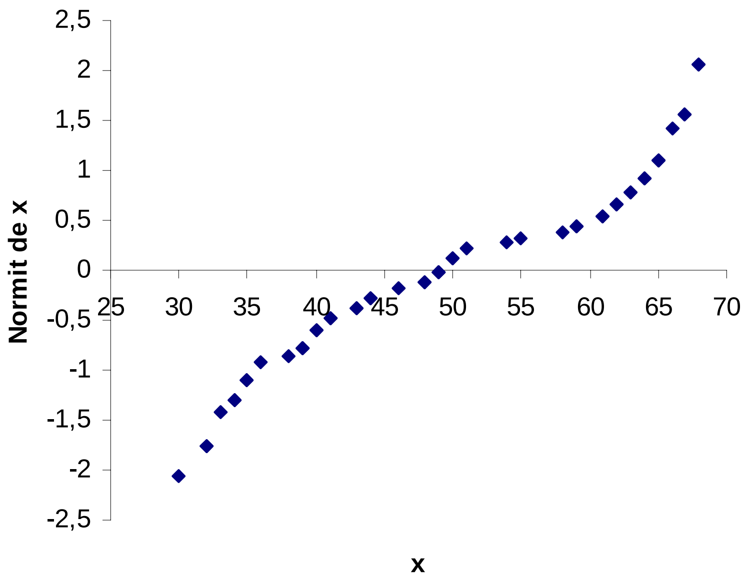


Figure 17

La forme de nuage traduit une distribution symétrique mais non gaussienne (ici, la distribution est plus proche de la loi uniforme que de la loi de Gauss).

Exemple 5

Le problème n'est plus de tester l'adéquation à une loi de Gauss. On a repéré des valeurs extrêmes (les observations n° 16, 19 et 31) et on voudrait juger si ces observations sont aberrantes sous l'hypothèse d'une distribution gaussienne.

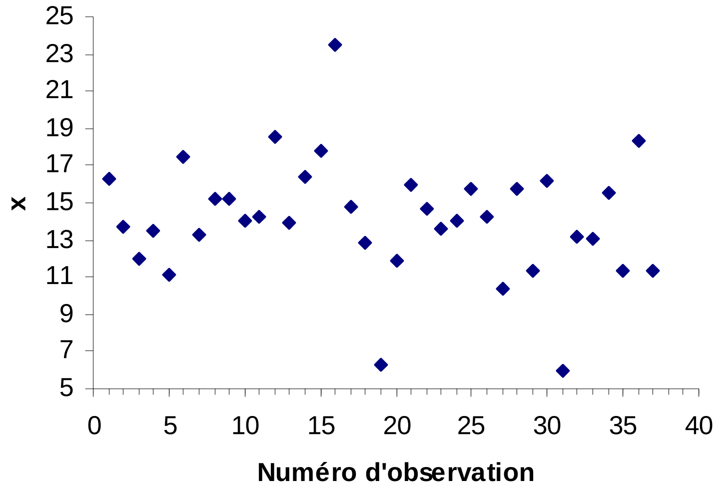


Figure 18

Ces observations s'écartent de l'alignement dans le diagramme de Henry. On peut par conséquent considérer qu'elles sont aberrantes sous l'hypothèse d'une répartition gaussienne.

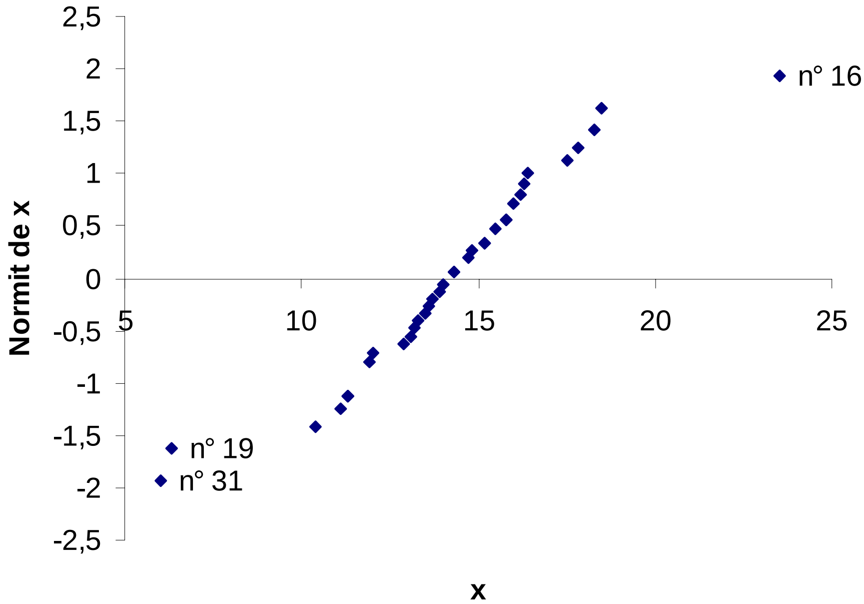


Figure 19

Remarques :

- Une observation n'est pas aberrante en soi mais sous une hypothèse de loi.
- Cette méthode est utilisée pour dépouiller les résultats des plans de criblage de facteurs. On parle alors de « diagramme de Daniel ». Voir le chapitre « Plans d'expériences » dans le manuel du groupe « Les cercles d'Excel'ense » [3].

Références

[1] AFNOR - Etude de la normalité d'une distribution. NF X 06-050, décembre 1995

[2] Goldfarb B., Pardoux C. - Méthodes d’ajustements graphiques. Excel'ense - MODULAD n°33, juillet 2005. www.modulad.fr

[3] Morineau A., Chatelin Y.-M. (Coordinateurs) – L'analyse statistique des données. Apprendre, comprendre et réaliser avec Excel. Editions Ellipses, 2005.