REVUE DE STATISTIQUE APPLIQUÉE

R. DUVAL

Étude graphique d'une distribution lognormale

Revue de statistique appliquée, tome 7, nº 1 (1959), p. 107-115

<http://www.numdam.org/item?id=RSA_1959__7_1_107_0>

ÉTUDE GRAPHIQUE
D'UNE DISTRIBUTION LOGNORMALE

R. DUVAL

Ancien Élève de l'École Polytechnique
Ingénieur Civil des Mines
Ancien Directeur à la Société M. M. de Penarroya

La note ci-dessous se propose de faire conaître quelques constructions géométriques qui paraissent intéressantes pour faciliter l'étude des distributions obéissant à une loi lognormale ou pouvant s'y rattacher.

Ces renseignements ont été extraits d'un article de M. F. Witz paru dans "Bergbau-Archiv" fascicule 2 - 1957. Les notations adoptées dans cet article ont été modifiées pour les adapter aux usages français et les justifications des constructions, qui n'étaient pas exposées dans le dit article, ont été reconstituées.

Ainsi qu'on le verra, il s'agit de procédés graphiques permettant de déterminer rapidement les abscisses de la moyenne arithmétique et du mode d'une distribution de valeurs lognormales et d'ajuster par un décalage d'origine certaines distributions qui ne paraissent pas lognormales à une loi lognormale.

Toutes ces constructions se font sur le tracé graphique de la somme cumulée des fréquences, anamorphosé sur un réseau de coordonnées dit normalo-logarithmique (lognormal), ou gausso-logarithmique ou encore galtonien-logarithmique, car la terminologie n'en est pas très bien fixée.

RESEAU DE COORDONNEES NORMALO-LOGARITHMIQUE

C'est le réseau qui sert à anamorphoser en droite de Henry une courbe cumulée des fréquences d'une loi lognormale.

Il est nécessaire pour la compréhension des procédés exposés ci-dessous d'avoir bien présente à l'esprit la manière dont ce réseau est construit.

Nous croyons donc utile de la rappeler ci-dessous.

L'origine des ordonnées se trouve sur l'horizontale cotée 50 % du réseau. L'échelle des ordonnées est une échelle métrique des paramètres t de l'expression

$$\pi\ (t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{\frac{-z^2}{2}}\ dz$$

c'est-à-dire que les ordonnées sont proportionnelles à t mais elles sont chiffrées pour une série de valeurs rondes des π (t) correspondant aux t. On obtient donc une graduation irrégulière, il s'agit d'une échelle dite fonctionnelle.

Ce t, qui est la variable réduite de la loi normale, se mesure en écarts-types, un écart-type réduit correspond comme on le sait à la valeur 0, 3413 de π (t). Les probabilités au lieu d'être chiffrées à partir de l'origine sont chiffrées en valeurs de 0, 50 - π (t) au-dessous de l'origine, cotée elle-même 0, 50 ou 50 %, et en valeurs de 0, 50 + π (t) au-dessus de l'origine, de manière à avoir une graduation croissante de $-\infty$ à $+\infty$.

Un écart-type est donc représenté sur l'axe des ordonnées par un segment s'étendant du point coté 0, 50 - 0, 3413 = 0, 1587 ou 15, 87 % au point coté 0, 50 ou 50 %, ou bien par un segment s'étendant du point coté 0, 50 ou 50 % jusqu'au point coté 0, 50 + 0, 3413 = 0, 8413 ou 84, 13 %. Chacun de ces segments, mesuré en mm par exemple, constitue donc le module μ_1 de l'échelle des ordonnées qu'il est nécessaire de connaître pour certaines constructions que nous verrons plus loin. On se contente le plus souvent d'ailleurs des points portant les cotes 16 % et 84 %, la différence avec les valeurs exactes étant inappréciable graphiquement, à l'échelle habituelle des figures.

L'échelle des abscisses est une échelle métrique des valeurs des logarithmes ordinaires d'une suite de nombres continue. Les divisions sont chiffrées généralement pour une progression arithmétique de ces nombres, les divisions sont donc encore irrégulières, c'est encore une échelle fonctionnelle. Cette échelle n'est pas portée sur l'horizontale cotée 50 % qui passe par l'origine, car elle y génerait les constructions et ne serait que mal visible, mais elle est transportée par translation parallèlement à l'axe des ordonnées vers le bas de la feuille. Son origine qui est sur l'axe des ordonnées est cotée 1 puisque log 1 = 0. Les longueurs des segments mesurées sur cet axe des abscisses à partir de l'origine sont proportionnelles aux valeurs des logarithmes ordinaires.

On trace cette division généralement au moyen d'une table des logarithmes ordinaires, mais pour l'étude de la loi lognormale il est commode de la considérer comme faite en logarithmes népériens. Etant donné que les logarithmes népériens sont dans un rapport constant avec les logarithmes ordinaires correspondants, il s'agit simplement d'une modification de module d'échelle que nous allons indiquer.

Quand on considère l'échelle des abscisses comme représentant des logarithmes ordinaires, le module, μ_2', de l'échelle est la longueur mesurée en mm du segment s'étendant depuis l'origine, cotée 1, au point coté 10, car log 10 = 1.

Lorsqu'on désire que cette même échelle représente des logarithmes népériens et qu'on considère un point coté x de l'échelle la longueur du segment s'étendant depuis l'origine jusqu'à ce point coté x est égale à $\log x . \mu_2'$, puisque μ_2' est le module de l'échelle des logarithmes ordinaires.

Mais si l'on veut que ce même segment mesuré en mm représente un logarithme népérien, sa longueur doit représenter Lx ; en appelant μ_2 le module, encore inconnu, de l'échelle des logarithmes népériens, elle doit donc être égale en mm à $Lx . \mu_2$. Cette longueur étant égale aussi à $\log x \, \mu_2'$ on doit donc avoir :

$$\mu_2' \log x = \mu_2 \, Lx \text{ soit}$$

$$\mu_2 = \frac{\log x}{Lx} . \mu_2' = 0,43429 \, \mu_2'$$

0, 43429 = log e étant le rapport constant $\frac{\log x}{Lx}$. Cette expression permet donc de déterminer immédiatement le module μ_2 de l'échelle considérée comme représentant des logarithmes népériens lorsqu'on connait μ_2', le module de l'échelle considérée comme représentant des logarithmes ordinaires.

Cet exposé étant terminé, on peut passer à l'explication des constructions géométriques dont il a été parlé au début.

DETERMINATION DE LA MOYENNE ARITHMETIQUE ET DE L'ABSCISSE DU MODE D'UNE DISTRIBUTION LOGNORMALE SUR LA DROITE DE HENRY REPRESENTANT LA DISTRIBUTION DES FREQUENCES CUMULEES DES LOGARITHMES

Les formules donnant cette moyenne et cet abscisse du mode sont comme on le sait (1) :

$$m = \gamma \, e^{\sigma^2/2} \quad \text{ou } Lm = L\gamma + \frac{\sigma^2}{2} \qquad (1)$$

et

$$s = \gamma \, e^{-\sigma^2} \quad \text{ou } Ls = L\gamma - \sigma^2 \qquad (2)$$

Dans ces formules

m = moyenne arithmétique, de la distribution des x ;
γ = médiane et moyenne géométrique des x ;
σ = écart-type de la distribution des logarithmes népériens des x (c'est donc un logarithme népérien).
$L\gamma$ = moyenne arithmétique et médiane de la distribution des logarithmes népériens des x ;
s = abscisses du mode de la distribution des x.

A - Moyenne arithmétique m

Considérons, figure 1, la droite de Henry tracée sur le réseau normalo-logarithmique, représentant la distribution des fréquences cumulées des logarithmes népériens des x.

Traçons par les graduations de l'axe des ordonnées 16 %, 50 % et 84 % des horizontales coupant respectivement la droite de Henry aux points A, D et B et projetons orthogonalement ces points sur l'axe des abscisses en respectivement a, d et b.

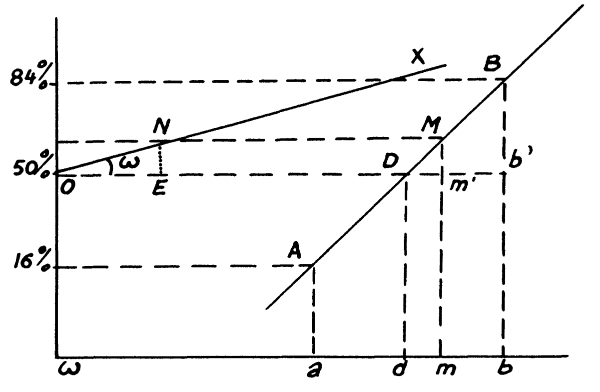

Figure 1

----

(1) Annales des Mines - Déc. 1955 - p. 18.

Nous remarquerons que les segments $\overline{ad} = \overline{db}$ représentent à l'échelle de l'axe des abscisses la valeur de , l'écart-type logarithmique. Il y a en effet 50 %- 16 % = 84 % - 50 % = 34 % de toutes les valeurs de Lx qui sont comprises entre les valeurs de Lx correspondant aux points a et d ou d et b, c'est la proportion qui correspond à la définition de l'écart-type d'une loi normale.

On peut donc admettre que $\overline{ad} = \overline{db}$ représente σ à l'échelle des abscisses.

Quant à l'abscisse de D, soit $\overline{\omega d} = \overline{oD}$, elle représente à cette même échelle $L\gamma$, moyenne arithmétique des Lx.

Supposons le problème résolu et admettons qu'on connaisse la valeur de la moyenne, m ; en portant à l'échelle des abscisses un segment représentant Lm, ce segment $\overline{\omega m}$ a son extrémité m qui se rappelle verticalement en M sur la droite de Henry. Si donc on connait M, on en déduit inversement le point m donc Lm et par suite la moyenne arithmétique.

Comme le segment $\overline{\omega d}$ représente $L\gamma$ et que $\overline{\omega m}$ représente Lm on doit avoir $\frac{\sigma^2}{2}$ représenté par le segment $\overline{dm}$ puisque

$$Lm - L\gamma = \frac{\sigma^2}{2} \quad \text{d'après l'équation (1)}$$

Considérons maintenant les 2 triangles rectangles semblables DMm' et DBb', on a évidemment :

$$\frac{\overline{Dm'}}{\overline{m'M}} = \frac{\overline{Db'}}{\overline{b'B}} \qquad (3)$$

En appelant μ_1 le module de l'échelle des ordonnées et μ_2 le module de l'échelle des abscisses, comme on constate que :

$\overline{Dm'}$ = dm doit représenter $\frac{\sigma^2}{2}$ à l'échelle des abscisses,
$\overline{m'M}$ représente à l'échelle des ordonnées une certaine valeur numérique y,
$\overline{Db'}$ = db représente σ à l'échelle des abscisses,
$\overline{b'B}$ représente ainsi qu'on l'a vu le module μ_1 des ordonnées, l'égalité (3) entre segments ci-dessus s'écrit donc :

$$y\,\mu_1 = \frac{\frac{\sigma^2}{2}\,\mu_2}{\sigma\,\mu_2}\cdot\mu_1 = \frac{\sigma}{2}\,\mu_1$$

On aboutit donc à la relation purement numérique

$$y = \frac{\sigma}{2} \qquad (4)$$

qui représente en coordonnées y et σ une droite passant par l'origine et de coefficient angulaire égal à 1/2.

On peut tracer cette droite sur les axes du réseau à partir de l'origine 0, il suffit pour cela de déterminer l'angle effectif ω qu'elle doit faire avec la droite horizontale cotée 50 %, compte tenu des modules différents μ_1 et μ_2 des échelles des deux axes.

Un point courant de cette droite aura pour coordonnées respectives, exprimées en mm, des segments $y\,\mu_1$ et $\sigma\,\mu_2$ et l'on aura :

$$y \cdot \mu_1 = tg\,\omega \,.\, \sigma \mu_2$$

Cette expression doit se réduire à l'équation (4) ci-dessus, par conséquent on doit avoir :

$$tg\,\omega = \frac{\mu_1 \cdot}{2\,\mu_2}$$

ce qui permet de tracer immédiatement la droite OX de la figure 1, qui ne dépend que de μ_1 et μ_2 , et qui est donc valable pour toutes les contructions faites sur ce réseau.

Lorsque cette droite est tracée, la détermination de M donc de m est immédiate. Il suffit de porter sur l'horizontale cotée 50 % un segment $\overline{OE} = \overline{db} = \sigma\,\mu_2$ puis par son extrémité E de tracer une verticale qui coupe la droite OX en un point N. On aura ainsi :

$$EN = \sigma\,\mu_2 \,.\, \frac{\mu_1}{2\,\mu_2} = \frac{\sigma}{2}\,\mu_1 = y\,\mu_1 \text{ représentant y à l'échelle des ordonnées}$$

Ce point N étant connu, on mène par lui une horizontale qui coupe la droite de Henry en M, d'où l'on déduit immédiatement sa projection verticale m sur l'axe des abscisses.

On a alors la valeur de la moyenne arithmétique en lisant cette valeur sur la graduation de l'échelle horizontale du réseau. Il faut souvent interpoler entre les graduations tracées. Cette interpolation se fait en mesurant en mm la longueur du segment ω m et en la divisant par la valeur du module μ_2' de l'échelle (longueur en mm du segment compris entre l'origine et la division 10 ainsi qu'il a été dit plus haut). Le quotient donne la valeur du logarithme ordinaire de la moyenne, donc la moyenne elle-même en repassant au nombre du logarithme.

B - Abscisse du mode

Ainsi que l'indique l'équation (2) vue plus haut on a :

$$Ls = L\gamma - \sigma^2$$

Reprenons (figure 2) le tracé de la droite de Henry sur le réseau normalo-logarithmique.

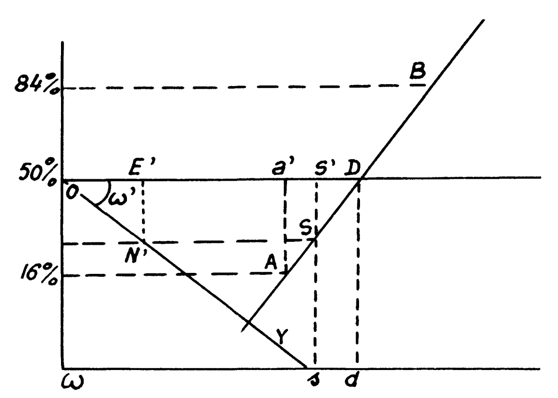

Figure 2

Le segment $\overline{\omega d}$ représente toujours à l'échelle des abscisses $L\gamma$, portons sur l'axe des abscisses, dans le sens décroissant, un segment $\overline{ds}$ représentant à l'échelle de cet axe la valeur σ^2, on voit que :

$$\overline{\omega s} = \overline{\omega d} - \overline{sd}$$

par conséquent en vertu de la relation (2), $\overline{\omega s}$ doit représenter à l'échelle de l'axe des abscisses la valeur de $Ls = L\gamma - \sigma^2 s^2$. On voit sur la figure que les triangles Da'A et Ds'S étant semblables on a :

$$\frac{\overline{Ds'}}{s'S} = \frac{\overline{Da'}}{a'A} \qquad (5)$$

D'après le même raisonnement que celui fait pour la détermination de la moyenne arithmétique cette expression peut s'écrire :

$$y'\,\mu_1 = \frac{\sigma^2\,\mu_2}{\sigma\,\mu_2}.\ \mu_1 = \sigma\,\mu_1$$

D'où

$$y' = \sigma \qquad (6)$$

expression représentant encore en coordonnées y' et σ une droite passant par l'origine de coefficient angulaire 1.

Le même raisonnement que celui fait pour la droite OX montre que l'angle ω' que doit faire la droite OY en question avec l'horizontale cotée 50 % du réseau est donné par :

$$\operatorname{tg}\omega' = \frac{\mu_1}{\mu_2}$$

On peut donc tracer immédiatement une fois pour toutes cette droite OY sur le graphique et en déduire le point S par une construction similaire à celle déjà vue pour la détermination de la moyenne arithmétique. Une fois le point s connu on en déduit l'abscisse du mode de la même façon que la moyenne arithmétique.

DECALAGE DE LA DISTRIBUTION LOGNORMALE

Il est bien rare lorsqu'on a reporté la courbe cumulée des fréquences sur un réseau normalo-logarithmique que l'on trouve un alignement rigoureusement en ligne droite, il faut, même lorsqu'on a affaire à une distribution lognormale, procéder à une compensation. Si cette compensation n'est visiblement pas possible, on est obligé de procéder à une compensation qui peut donner une courbe convexe du côté des abscisses croissantes ou inversement. Dans le premier cas on peut essayer en augmentant toutes les valeurs observées d'une quantité constante K (c'est-à-dire, en fait, en décalant l'origine de K vers la gauche) d'arriver à obtenir une courbe cumulée des fréquences se rectifiant sur le réseau normalo-logarithmique. Dans le second cas on peut faire le même essai mais en retranchant une constante K de toutes les valeurs observées. Lorsqu'on opère par tatonnements cette recherche peut être très longue et on n'a jamais une preuve que le problème cherché est impossible.

L'auteur de l'article signalé au début, M. Witz, a indiqué une méthode rapide permettant de déterminer cette valeur K et de constater lorsqu'elle ne donne pas la rectification escomptée que le problème n'a pas de solution, c'est-à-dire que la distribution étudiée n'est pas susceptible de se transformer en distribution lognormale par la transformation envisagée.

Examinons le cas de la constante K additive ; afin d'éviter les confusions dans ce qui suit il sera convenu que les notations indiquant des nombres seront soulignées, celles indiquant des vecteurs seront surlignées, les lettres indiquant simplement des points de la figure resteront telles quelles.

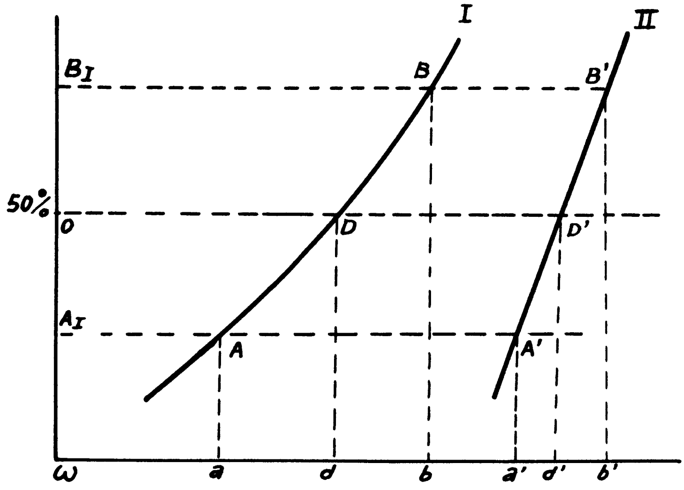

Figure 3

Supposons le problème résolu (figure 3) et admettons que la courbe cumulée observée soit la courbe I et que l'addition de la constante K aux valeurs observées la transforme en la droite II.

Les points D et D', intersections de I et II avec l'horizontale cotée 50 %, se projettent aux points d et d' de l'axe des abscisses qui correspondent respectivement aux graduations $\underline{d}$ et $\underline{d'}$ de l'échelle.

D'après la façon dont les deux courbes se déduisent l'une de l'autre on doit avoir :

$$\underline{d'} = \underline{d} + \underline{K} \qquad (1)$$

Le segment $\overline{\omega d'}$ représente à l'échelle des abscisses le logarithme népérien de la médiane de la distribution transformée qui est $\underline{d'}$, c'est donc à l'échelle des abscisses $L\underline{d'}$, quant au segment $\overline{\omega d}$, il représente à l'échelle des abscisses $L\underline{d}$.

On a donc $L\underline{d'} = L(\underline{d} + \underline{K})$

Considérons deux points A et B de la courbe I placés respectivement à égale distance de la droite 50 % mais de part et d'autre. Ces deux points correspondent aux deux points A' et B' de la droite II. Projetons ces points verticalement sur l'axe des abscisses en a et b, ils se placent sur les graduations $\underline{a}$ et $\underline{b}$.

Posons :

$$\underline{\alpha} = \underline{d} - \underline{b} \qquad \text{et } \underline{\beta} = \underline{a} - \underline{d} \qquad (2)$$

De même projetons A' et B' verticalement sur l'axe des abscisses en a' et b', ces projections se placent aux graduations $\underline{a}'$ et $\underline{b}'$ respectivement.

D'après la manière dont la droite II a été construite à partir de I, on doit avoir :

$$L\underline{a}' = L\,(\underline{a} + \underline{K}) \qquad \text{et } L\underline{b}' = L\,(\underline{b} + \underline{K})$$

Le segment $\overline{\omega a'}$ représente $L\underline{a}'$, le segment $\overline{\omega b'}$ représente $L\underline{b}'$, nous avons déjà vu que $\overline{\omega d'}$ représente $L\underline{d}'$.

Comme les deux segments $\overline{a'd'}$ et $\overline{d'b'}$ sont égaux par construction, puisque la courbe II est une droite, on doit donc avoir d'après ce qu'on a vu ci-dessus :

$$L\underline{d}' - L\,(\underline{a} + \underline{K}) = L\,(\underline{b} + \underline{K}) - L\underline{d}'$$

ou d'après (2)

$$L\underline{d}' - L\,(\underline{\beta} + \underline{d} + \underline{K}) = L\,(\underline{d} - \underline{\alpha} + \underline{K}) - L\underline{d}'$$

c'est-à-dire en remplaçant $\underline{d}$ par sa valeur tirée de (1)

$$L\underline{d}' - L\,(\underline{\beta} + \underline{d}') = L\,(\underline{d}' - \underline{\alpha}) - L\underline{d}'$$

ou

$$(\underline{\beta} + \underline{d}')\,(\underline{d}' - \underline{\alpha}) = d'^2$$

ou

$$\frac{1}{\underline{d}'} = \frac{1}{\underline{\alpha}} - \frac{1}{\underline{\beta}} \qquad (3)$$

Les valeurs $\underline{\alpha}$ et $\underline{\beta}$ se déduisent des lectures faites sur la courbe I, l'expression (3) donne donc la valeur de $\underline{d}'$ à partir de la courbe II.

Comme $\underline{d}$ est connu d'après la courbe I, l'expression (1) fournit la valeur $\underline{K}$ de la constante, dès qu'on connaît $\underline{d}'$.

Cette valeur $\underline{K}$ étant déterminée, on déduit les positions des points A' et B' à partir de celle des points A et B puisque

$L\,(\underline{a} + \underline{K}) = L\,(\underline{a}')$ est représenté par $\overline{\omega a'}$ et

$L\,(\underline{b} + \underline{K}) = L\,(\underline{b}')$ est représenté par $\overline{\omega b'}$, d'où les points A' et B'.

On peut donc placer sur le graphique les trois points A'D'B'.

Si ces trois points sont en ligne droite, on obtient la droite II qui représente la somme cumulée des fréquences de la transformée sur le réseau normalo-logarithmique. La détermination de la moyenne arithmétique et du mode de la distribution transformée par les méthodes indiquées plus haut permet de repasser facilement à la moyenne arithmétique et au mode de la distribution originelle dont la densité de probabilité est alors de la forme :

$$\frac{1}{\sigma\sqrt{2\pi}}\; e^{-\frac{L(x+K) - Ld')^2}{2\sigma^2}}$$

Si ces trois points ne sont pas en ligne droite c'est que la distribution originelle n'est pas de la forme indiquée par la densité de probabilité ci-dessus. Cependant on peut étudier si elle n'en diffère pas de trop, de la façon suivante :

En effet dans tout ce qui a été dit ci-dessus, la distance séparant les points A et B de la courbe I de la droite cotée 50 % n'ont pas été spécifiées, il suffit qu'elles soient égales. Tout couple de points A et B satisfaisant à cette condition fournit donc un jeu des points A'D'B'. En prenant une série de couples de points A et B sur I on obtiendra autant de triplets A'D'B'. En reportant ces triplets A'D'B' sur le graphique, on peut examiner si le nuage ainsi formé n'est pas large et paraît pouvoir être à peu près compensé par une droite. Cette droite pourrait alors être considérée comme représentant approximativement une transformée de la courbe II, ce qui permettrait une détermination approchée des paramètres de la distribution.

Le cas où la constante K est soustractive se traite exactement de la même façon et par les mêmes relations (2).

CONCLUSIONS

Ces constructions paraissent intéressantes car elles permettent d'essayer sur des distributions observées dissymétriques comme toutes les distributions de teneurs de gisement des ajustements à des lois mathématiques relativement simples permettant une détermination aisée notamment de la moyenne arithmétique, sans obliger à recourir à des courbes de Pearson.