MÉTHODE GRAPHIQUE DE CALCULS DE LA MOYENNE ET DE L'ÉCART TYPE D'UNE DISTRIBUTION NORMALE TEST DE NORMALITÉ

PAR

R. TOMASSONE
Ingénieur des Eaux et Forêts
9e Section de la Station de Recherches

L'analyse statistique des résultats expérimentaux demande des calculs souvent longs, et lorsque les calculs sont faits à la main, à la règle à calcul ou même avec une machine à calculer de bureau, les risques d'erreur sont grands. Comme d'autre part les calculs usuels ne demandent pas une grande précision, une interprétation graphique des résultats suffit la plupart du temps ; cette interprétation possède en outre l'avantage d'être beaucoup plus claire.

Le problème que nous nous proposons de traiter ici est lié à l'étude de la distribution normale (*) :

1) Etant donné un n — échantillon (**) issu d'une loi normale, calculer sa moyenne et son écart type.

(*) Pour la définition statistique de base : distribution, fonction de répartition (courbe des fréquences cumulées), distribution normale, cf. « La méthode statistique et ses applications en matière forestière ». R.F.F., août 1953, numéro spécial et référence [5].

(Les chiffres entre crochets [ ] renvoient à la bibliographie qui se trouve à la fin de l'article).

(**) Dans la terminologie statistique un n — échantillon est un échantillon constitué par n éléments.

2) Soumettre à un test l'hypothèse « la population dont une image est donnée par le n — échantillon est une population normale » (*).

Cette seconde partie nous paraît très importante notamment en vue de l'application des méthodes statistiques à l'analyse des corrélations et à l'analyse de variance. Dans les deux cas, on suppose que les distributions des mesures sont normales; lorsqu'elles ne le sont pas, les conclusions qu'on peut tirer des résultats d'une analyse statistique sont fausses. En particulier, lorsqu'on calcule un coefficient de corrélation et lorsqu'on trouve qu'il n'est pas « significativement différent de 0 », on conclut très vite que les deux caractères étudiés sont indépendants: ceci peut être faux car l'absence de corrélation entraîne l'indépendance seulement si les distributions des deux caractères sont normales.

1. — Estimation de la moyenne et de l'écart type : Méthode de la droite de Henry

On sait que la densité de probabilité d'une distribution normale de moyenne m et d'écart type σ est donnée par la formule:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[- \frac{(x - m)^2}{2\sigma^2} \right] \quad (**)$$

La fonction de répartition normale F (x) correspond à la somme de toutes ces densités jusqu'à une valeur x: c'est cette quantité F (x) que nous allons utiliser. Si on prend un graphique dont les graduations correspondent:

— en abscisse X: les intervalles de classe,

— en ordonnée F: une graduation avec pour échelle linéaire l'écart type correspondant à la distribution $\sigma = 1$; le point de coordonnée X, F (dans le cas où la distribution parente est normale, répétons-le) décrit une droite [3].

Une table de la loi normale donne pour fréquences cumulées (F) les valeurs suivantes:

σ	— 2	— 1	0	1	2
F (en %)	2,28	15,87	50,00	84,13	97,72

(*) Ici soumettre à un test signifie: étant donné les risques d'erreur que l'on accepte de prendre pour une grandeur de l'échantillon, est-ce que les écarts observés entre la répartition de l'échantillon et la répartition normale admettant même moyenne et même écart type ne sont dus qu'à des causes non contrôlées? (ce que le statisticien appelle: le hasard).

(**) exp () = exponentielle.

Dans la pratique, on trouve des graphiques entièrement quadrillés suivant ces graduations (*).

L'intersection avec la droite F = 50 donne la moyenne estimée m. La pente de la droite est inversement proportionnelle à l'écart type σ: donc plus l'angle avec une parallèle à l'axe des abscisses est grand, plus la dispersion est faible. Pratiquement, on obtient σ comme différence entre les deux valeurs x (83,13) et x (50,00)

Exemple: Exp. de comparaison de provenances d'épicéas [M. LEMOINE 62-23] ; on a mesuré sur 100 plants (**) de la provenance AURS les hauteurs figurant au tableau I.

TABLEAU I

(1) X Classe de hauteur	(2) f fréquence relative	(3) F fréquences cumulées	(4) X — 20,6 variable centrée	(5) $\frac{X - 20,6}{5,6}$ variable centrée réduite
—	—	—	—	—
12	5	5	— 8,06	— 1,44
14	11	16	— 6,06	— 1,08
16	11	27	— 4,06	— 0,72
18	14	41	— 2,06	— 0,37
20	12	53	— 0,06	— 0,01
22	14	67	1,94	0,35
24	9	76	3,94	0,71
26	5	81	5,94	1,06
28	13	94	7,94	1,42
30	2	96	9,94	1,77
32	4	100	11,94	2,13

Nous utilisons les colonnes 1, 2 et 3 du tableau I:

— pour la classe 11 - 13 (centre de classe 12) on trace un trait à F = 5

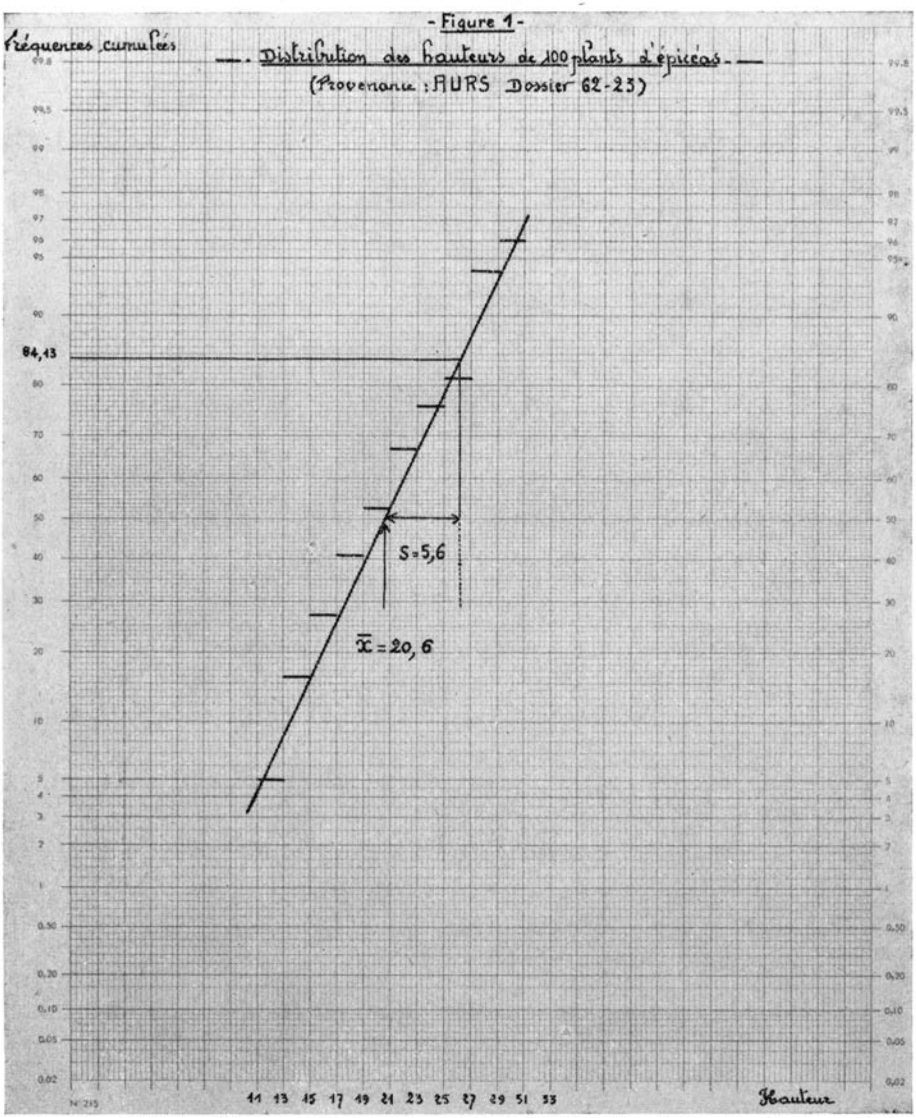
— pour la classe 13 - 15 (centre de classe 14) on trace un trait à F = 16

et ainsi de suite jusqu'à l'avant-dernière classe.

On trace la droite joignant les points les plus à droite des traits de chaque classe (voir fig. 1):

(*) On les appelle « Diagramme de la droite de Henry ou droite échantillon » (anamorphe de la courbe de probabilité totale de la loi de Gauss).

(**) donc ici fréquence relative en pourcentage est égale au nombre plants.


- Figure 1 -
Fréquences cumulées
- Distribution des hauteurs de 100 plants d'épicéas -
(Provenance : AURS Dossier 62-23)
84,13
S=5,6
x̄ = 20,6
11 13 15 17 19 21 23 25 27 29 31 33
Hauteur

— l'intersection avec F = 50,00 donne *moyenne* = *20,6*

— l'intersection avec F = 84,13 donne $x_{(84,13)}$ = 26,2 d'où

$$\sigma = 26{,}2 - 20{,}6 = 5{,}6$$

Le calcul exact donne $\bar{x} = 20{,}88 \qquad \sigma = 5{,}4$

2. — Test de normalité de la distribution

En pratique, les points ne sont pas rigoureusement alignés, un simple coup d'œil sur le graphique permet de se rendre compte si l'hypothèse de normalité est valable. Mais cette méthode est très intuitive, elle ne tient aucun compte du nombre de mesures dont on dispose. Il est normal qu'avec 40 mesures, la dispersion soit grande même si la distribution parente est normale; avec 1000 une grande dispersion est fort improbable. Des graphiques portant des intervalles de confiance ont été tracés, ce sont eux que nous allons utiliser (cf JACQUET [1]).

Pour ne pas en construire un trop grand nombre, ils correspondent à $\sigma = 1$ et au seuil de confiance couramment utilisé de 95 %.

Pour ce test on utilise les colonnes 4 et 5 du tableau I et la figure 2:

— on « centre » la variable, c'est-à-dire qu'on calcule les écarts par rapport à la moyenne estimée 20,6 (colonne 4).

— on « réduit » la variable en divisant par 5,6 (colonne 5).

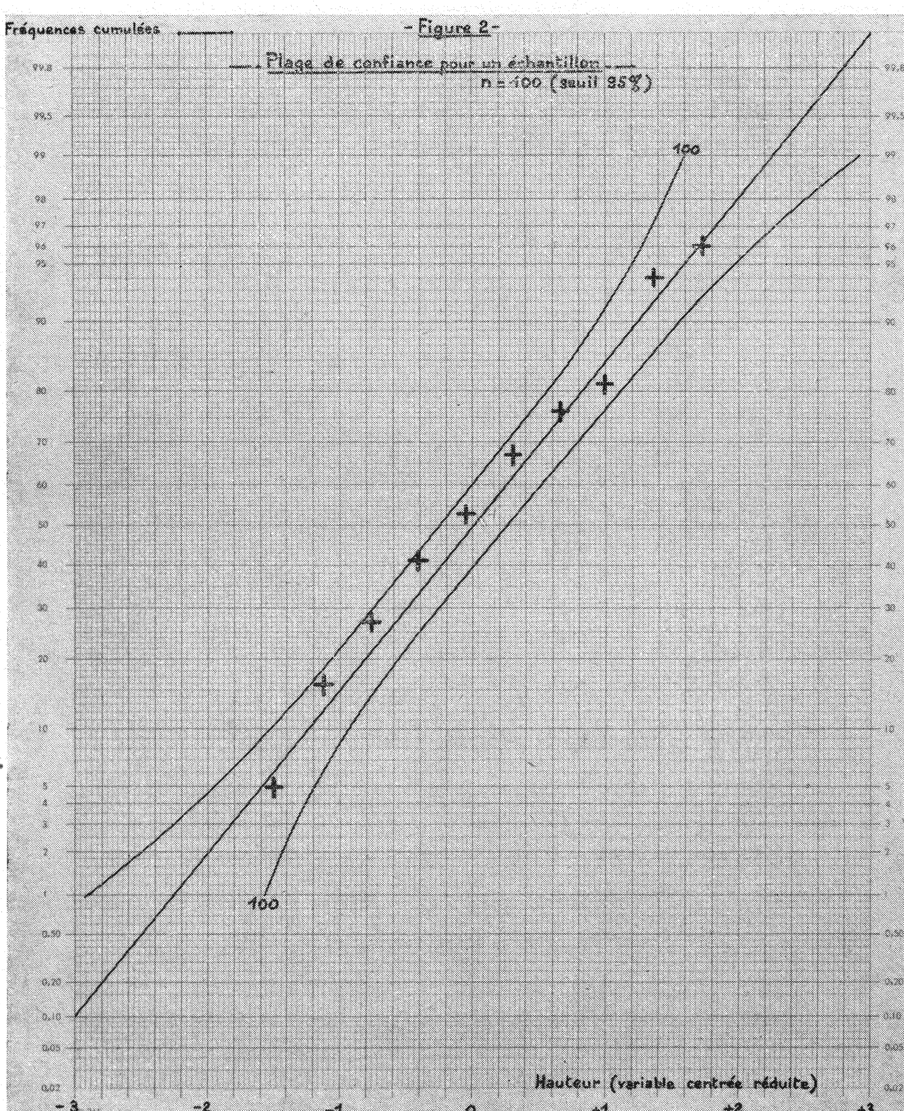
Si l'hypothèse de normalité est exacte, la nouvelle variable $\dfrac{x - 20{,}6}{5{,}6}$ suit une loi normale de moyenne 0 d'écart type 1. Pour un n — échantillon, les points, au seuil de 95 %, doivent tous être situés dans la plage de confiance correspondante. Par exemple, sur la figure 2, pour n = 100, rien ne s'oppose à admettre que la distribution est normale (*) (**).

On peut donc conclure qu'au seuil choisi et avec l'effectif dont on dispose, la distribution des hauteurs est normale et que l'estimation de la moyenne est m = 20,6, celle de l'écart type $\sigma = 5{,}6$.

Pour appliquer ce test, il suffira donc de construire un graphique correspondant à la figure 1, de centrer et de réduire la variable pour tracer un autre graphique équivalent à la figure 2, et d'utiliser le calque ci-joint pour tester l'hypothèse de normalité.

(*) Un test du χ^2 beaucoup plus long à mettre en œuvre permet d'arriver au même résultat.

(**) Pour d'autres exemples de distribution que l'on peut rencontrer dans des problèmes forestiers, cf. PARDÉ [2] et SMITH [4].


Fréquences cumulées
- Figure 2 -
Plage de confiance pour un échantillon
n = 100 (seuil 95%)
100
100
Hauteur (variable centrée réduite)
-3
-2
-1
0
+1
+2
+3

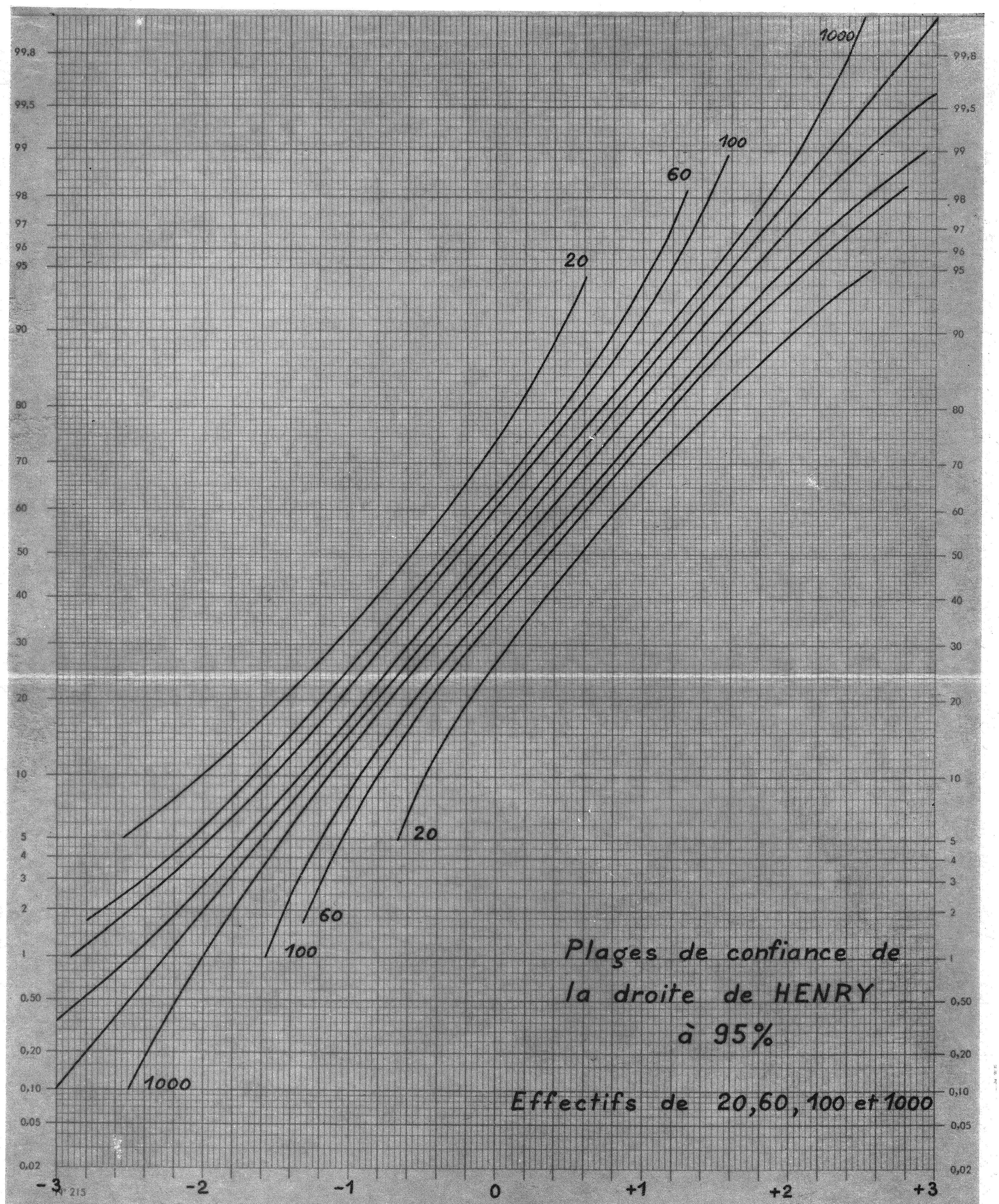


Diagramme de la droite de Henry ou droite-échantillon (anamorphe de la courbe de probabilité totale de la loi de Gauss).

BIBLIOGRAPHIE

[1] R. JACQUET (1962). — Contribution aux études de plages de confiance d'une droite de Henry. Thèse présentée à la Faculté des Sciences de l'Université de Nancy.
[2] J. PARDÉ (1961). — Dendrométrie. Louis-Jean. Gap.
[3] RISSER et TRAYNARD (1957). — Les Principes de la Statistique Mathématique. 2ᵉ édition. Gauthier-Villars. Paris.
[4] SMITH J.H.G. and KER J.W. (1957). — Some distributions encountered in sampling forest stands. Forest Science 3 (2), p. 137-144.
[5] VESSEREAU A. (1960). — Méthodes Statistiques en Biologie et en Agronomie. 2ᵉ édition. Baillère. Paris.
[6] R.F.F. (Août 1953). — Numéro spécial. La méthode statistique et ses applications en matière forestière.