# Estimation stochastique de la précision des mesures

Autor(en): **Bachmann, W.K.**

Objekttyp: **Article**

Zeitschrift: **Mensuration, photogrammétrie, génie rural**

Band (Jahr): **71-F (1973)**

Heft 4

PDF erstellt am: **22.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-226196

# Estimation stochastique de la précision des mesures

par W.K. Bachmann

*Résumé*

Ce bref aperçu des principales méthodes, utilisées pour l'estimation de la précision de mesures, a pour but essentiel d'attirer l'attention du topographe sur ces méthodes qui, hélas, semblent souvent encore être ignorées. A l'aide d'exemples numériques, l'utilisation des distributions $X^2$, t et F a été montrée et l'emploi de séries de mesures ordonnées pour l'estimation rapide esquissée.

*Zusammenfassung*

Diese kurze Darstellung der wichtigsten statistischen Methoden zur Schätzung der Genauigkeit von Messungen bezweckt vor allem den Vermessungs-Fachmann auf diese Methoden, die leider immer noch etwas stiefmütterlich behandelt werden, aufmerksam zu machen. Es wird die Verwendung der $X^2$, t und F-Verteilungen anhand numerischer Beispiele gezeigt und die Bedeutung der geordneten Messreihen zur angenäherten Schätzung der erreichten Genauigkeit kurz dargelegt.

## Chapitre 1

### Estimation par intervalles de confiance

*1.1 Généralités et distribution normale*

La statistique étant une science exacte, il est difficile, voire même impossible d'en parler sans avoir recours aux mathématiques. Dans ce qui suit, je m'efforcerai cependant d'utiliser le moins de formules possible tout en tâchant de faire ressortir quelques idées fondamentales.

De nos jours, la statistique est si vaste et si universellement appliquée que personne ne peut prétendre la connaître intégralement. Aussi n'ai-je nullement l'intention de brosser un tableau d'ensemble; je me bornerai à examiner sous un angle critique quelques différences de conception fondamentales qui séparent la *statistique* de la *théorie des erreurs classique.* Cela faisant, je supprimerai la plupart des démonstrations mathématiques, me bornant à signaler uniquement les résultats.

Lorsqu'il s'agit de déterminer expérimentalement la valeur d'une inconnue X, on la mesure une ou plusieurs fois au moyen d'un équipement approprié. En répétant la mesure, on peut poursuivre deux buts, suivant qu'il s'agit uniquement d'un contrôle ou de la recherche d'une plus grande précision.

Du point de vue de la statistique, l'ensemble des valeurs qu'une grandeur X peut prendre au cours du processus de mesure constitue une *population* et chaque valeur possible est un *individu* de cette population.

Dans ces conditions, une population n'est jamais connue complètement, et les observations servent précisément à la découvrir, à la saisir dans la mesure du possible.

Mais il y a d'autres cas, où la population peut être connue *complètement* grâce aux observations. Il en est généralement ainsi lorsque la population n'est constituée que par un nombre fini d'individus. Citons comme exemple le recensement des habitants d'une contrée.

En géodésie, la situation est autre en ce sens que la population est constituée par toutes les mesures *qu'on pourrait* effectuer sur l'inconnue si l'on était à même de les faire; mais pour des raisons d'ordre pratique et économique, le nombre de mesures est toujours très limité dans ce domaine. On voit que dans ce cas la population totale, comportant une infinité d'individus, est purement fictive; c'est une *conception de l'esprit.*

Si une série de n mesures nous fournit les valeurs $x_1, x_2, \ldots, x_n$, celles-ci constituent *un échantillon de taille n* de la population X. Mais dans tout échantillon les éléments $x_1, x_2, \ldots, x_n$ doivent être *indépendants* et choisis *„au hasard"* dans la population. Pour être précis, il faudrait naturellement définir ce que l'on entend par l'expression „au hasard". Cette définition est d'autant plus importante qu'on utilise aujourd'hui en statistique expérimentale des échantillons engendrés par un *ordinateur,* ce qui nous dispense d'avoir recours à des mesures. Aussi un certain nombre des exemples que nous donnerons ont-ils été „fabriqués" par ordinateur.

Pour des raisons d'ordre pratique, la taille d'un échantillon obtenu par un procédé de mesure sera toujours très limitée et de ce fait l'échantillon ne donnera qu'une

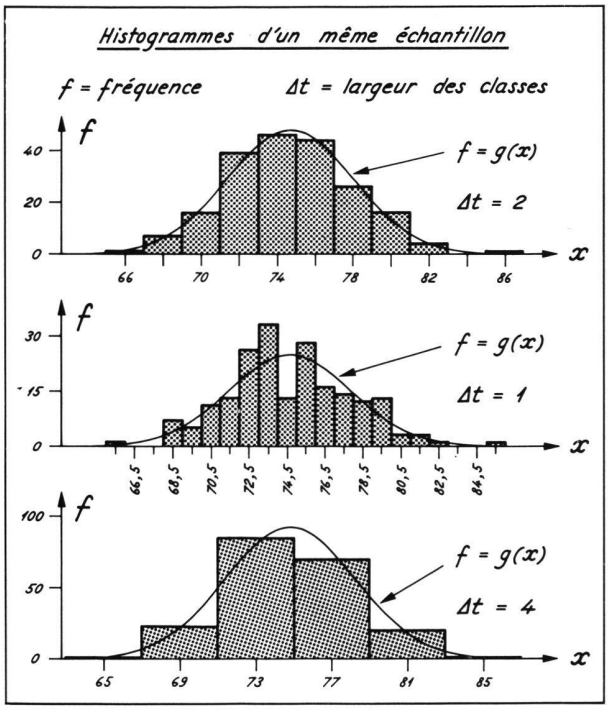



Fig. 1.1.1

vue approchée de l'ensemble de la population. Mais il est entendu que la quantité d'information augmente avec le nombre des mesures, à condition que celles-ci soient *exemptes d'erreurs systématiques.*

Les expressions „échantillon" et „échantillonnage" ont leur origine dans les *applications industrielles,* où l'on contrôle périodiquement la qualité du produit fabriqué par des sondages, c'est-à-dire par des *échantillonnages.* La notion d'échantillon intervient cependant aussi dans des domaines tout autres, tels que *la médecine, la pharmacologie, la biologie, l'économie, l'électronique, l'assurance-vie, etc.*

Ayant obtenu n mesures $x_1$, $x_2$, ..., $x_n$ pour l'inconnue X, nous pouvons construire un *histogramme* en reportant en abscisse les valeurs $x_i$ et en ordonnée les *fréquences* ou les *fréquences relatives* des mesures *groupées par classes.* Dans le dessin de l'histogramme, la largeur choisie pour la classe joue un grand rôle, car si elle est trop petite, l'histogramme prend une allure irrégulière et si elle est trop grande, on perd une partie de l'information.

Si tout se passe normalement, l'histogramme n'aura qu'*un seul maximum.* S'il en comporte deux ou plus, l'ensemble des mesures est hétérogène et nous devons admettre qu'il provient de deux populations distinctes ou plus; voir fig. 1.1.2.

Cette méthode est par exemple utilisée en biologie pour la séparation de deux espèces.

En mensuration, les histogrammes ne doivent comporter qu'un seul maximum, car sans cela les mesures n'auraient pas de sens, ou accuseraient une variation de la grandeur à déterminer au cours du processus de mesurage (exemple: réfraction verticale ou horizontale, due à un changement du milieu ambiant).

Ayant dessiné l'histogramme pour un ensemble de mesures, on est naturellement tenté de le remplacer par une fonction $f = g(x)$ en choisissant les valeurs mesurées x pour abscisses et la *fréquence* pour ordonnée f. Dans ce but, on prend pour g(x) une *forme algébrique appropriée,* qui peut comporter un ou plusieurs paramètres. En métrologie, la *courbe de Gauss* a une importance particulière vu que la majorité des mesures suivent à peu près cette loi. En fixant *la forme algébrique* de g(x), nous choisissons un *modèle mathématique* pour la présentation de la population. Si la correspondance entre ce modèle et l'ensemble des mesures est jugée suffisamment bonne, nous acceptons le modèle, dans le cas contraire nous le remplaçons par un autre, plus approprié. Cette procédure est bien connue en physique, où l'on change de modèle presque tous les jours pour tenir compte des dernières découvertes. Les géodésiens, plus conservateurs, sont restés fidèles à la loi de Gauss! Cette dernière est donnée par la *fonction de fréquence* (relative)

$$y = f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \qquad 1.1.1$$

dans laquelle f(x) désigne la *densité de probabilité* au point x, tandis que m et $\sigma$ sont deux paramètres qui s'appellent *moyenne et écart-type.*

Comme il aurait été trop compliqué de calculer des tables numériques pour chaque valeur de m et de $\sigma$, on a recours à la substitution

$$u = \frac{x-m}{\sigma} \quad \text{(variable réduite)} \qquad 1.1.2$$

qui ramène toutes les distributions normales $N(m;\sigma^2)$ à un type unique N (0;1), dit *distribution normale standard ou réduite.* C'est pour cette dernière qu'on a calculé des tables donnant notamment les valeurs de

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2} \qquad 1.1.3$$

$$P(-\infty < U \leqslant u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{1}{2}\xi^2}\, d\xi$$

$$P(-u \leqslant U \leqslant u) = \Psi(u) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1$$

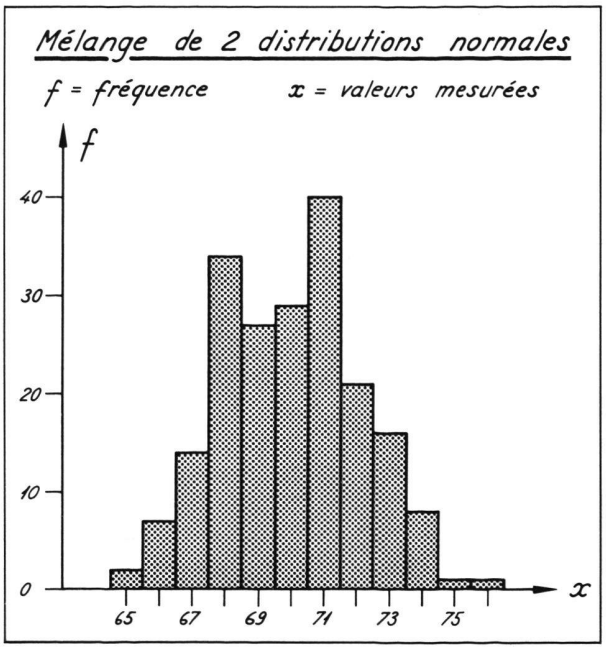



Fig. 1.1.2

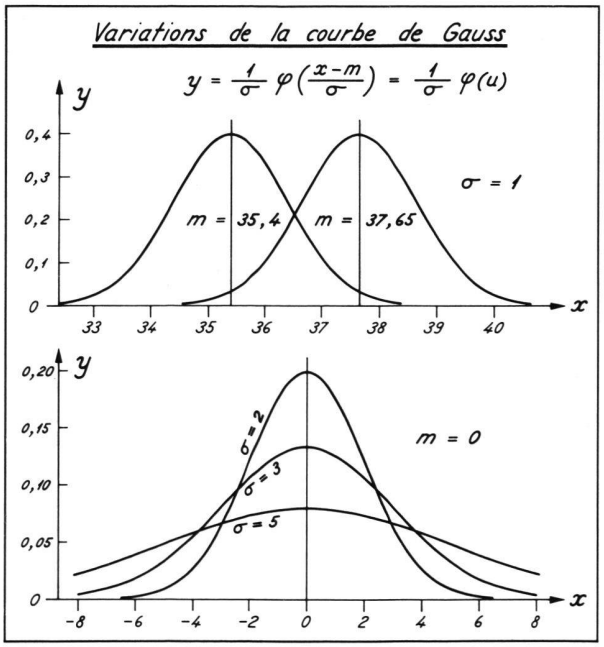



Fig. 1.1.3

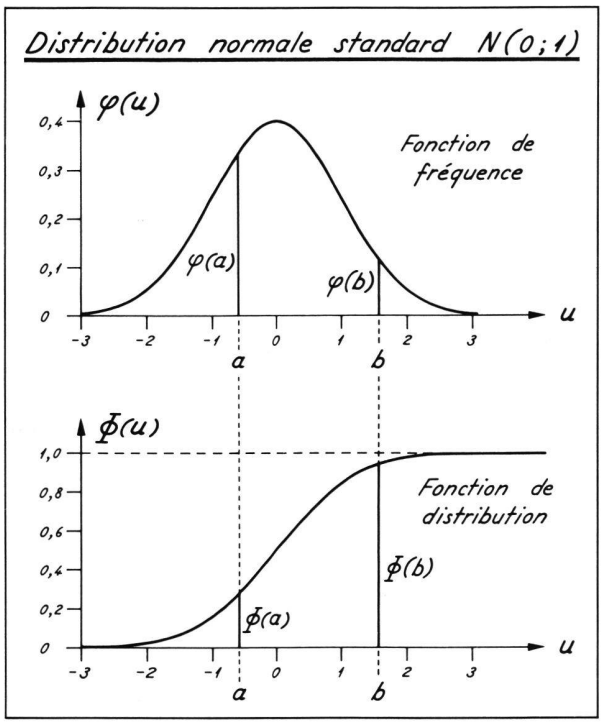
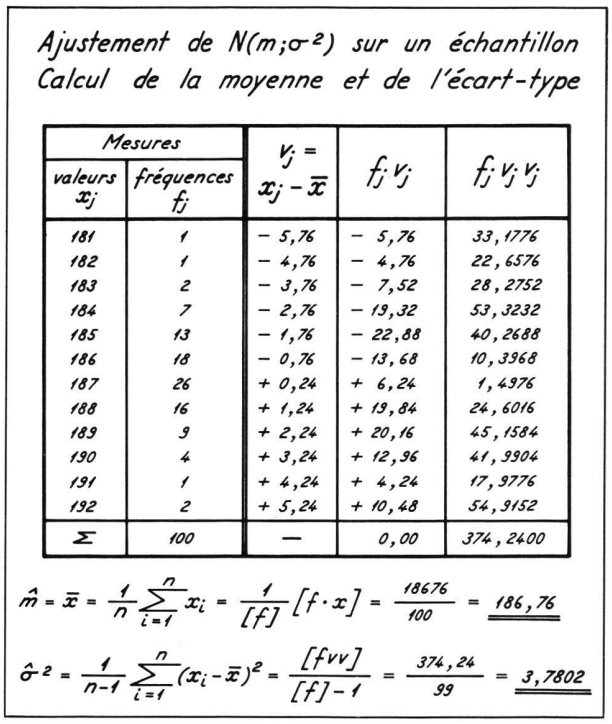



Fig. 1.1.4




| Mesures | | $v_j = x_j - \bar{x}$ | $f_j v_j$ | $f_j v_j v_j$ |
|---|---|---|---|---|
| valeurs $x_j$ | fréquences $f_j$ | | | |
| 181 | 1 | − 5,76 | − 5,76 | 33,1776 |
| 182 | 1 | − 4,76 | − 4,76 | 22,6576 |
| 183 | 2 | − 3,76 | − 7,52 | 28,2752 |
| 184 | 7 | − 2,76 | − 19,32 | 53,3232 |
| 185 | 13 | − 1,76 | − 22,88 | 40,2688 |
| 186 | 18 | − 0,76 | − 13,68 | 10,3968 |
| 187 | 26 | + 0,24 | + 6,24 | 1,4976 |
| 188 | 16 | + 1,24 | + 19,84 | 24,6016 |
| 189 | 9 | + 2,24 | + 20,16 | 45,1584 |
| 190 | 4 | + 3,24 | + 12,96 | 41,9904 |
| 191 | 1 | + 4,24 | + 4,24 | 17,9776 |
| 192 | 2 | + 5,24 | + 10,48 | 54,9152 |
| Σ | 100 | — | 0,00 | 374,2400 |

$$\hat{m} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{[f]}[f \cdot x] = \frac{18676}{100} = 186{,}76$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{[fvv]}{[f]-1} = \frac{374{,}24}{99} = 3{,}7802$$

Fig. 1.1.5

*L'ajustement* de la fonction de fréquence f(x) sur un histogramme, respectivement un échantillon

$x_1, x_2, \ldots, x_n$ échantillon 1.1.4

est un problème bien connu qu'on peut résoudre en appliquant la *méthode du maximum de vraisemblance,* qui nous fournit les *estimateurs* (ou estimations)

$$\hat{m} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad 1.1.5$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

On est ainsi amené à un jeu de formules bien connu qui ne nous offre rien de nouveau. Une application numérique est donnée sur les Fig. 1.1.5 et 1.1.6

En introduisant ces estimateurs, nous remplaçons les n mesures $x_1, x_2, \ldots, x_n$ par les deux valeurs $\hat{m}$ et $\hat{\sigma}$, dont la première représente la *valeur compensée de l'inconnue* X, tandis que la seconde caractérise la *précision des mesures.* En procédant ainsi, on perd naturellement une partie de *l'information* fournie par les mesures, ce qui est regrettable mais inévitable. Nous disons dans ce cas qu'on a fait une *estimation ponctuelle.*

Quoiqu'on se trouve dans un domaine bien connu en choisissant $\hat{m}$ et $\hat{\sigma}$, c'est ici que les *premières difficultés* commencent. En effet, nous avons calculé $\hat{m}$ et $\hat{\sigma}$ à partir d'un ensemble de mesures $x_1, x_2, \ldots, x_n$ bien déterminé, mais nous ne savons pas ce que l'on aurait obtenu si l'on avait effectué une seconde série de mesures $y_1, \ldots, y_n$. La différence entre ces deux séries aurait-elle eu une *influence significative* sur $\hat{m}$ et $\hat{\sigma}$ ou non? Autrement dit, nous ne connaissons pas la *confiance* que nous devons attribuer à $\hat{m}$ et $\hat{\sigma}$. On peut toutefois démontrer que les estimateurs $\hat{m}$ et $\hat{\sigma}$ tendent vers m et $\sigma$ si $n \to \infty$, ce qui nous montre que *la confiance que nous pouvons attribuer au résultat final augmente avec n.*

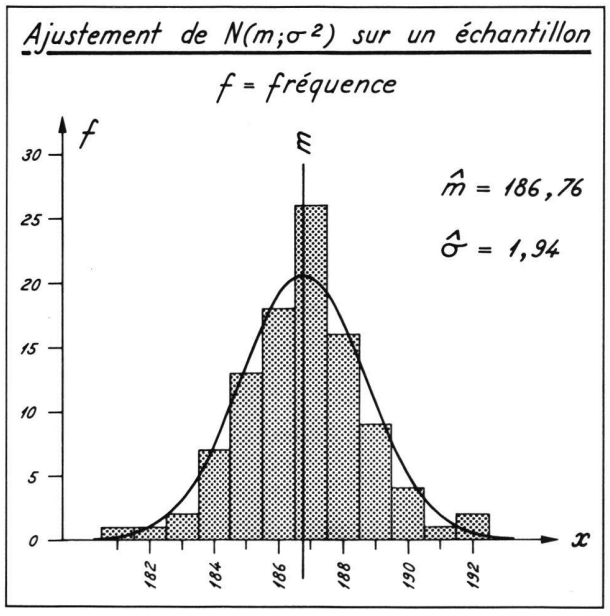



Fig. 1.1.6

Pour chiffrer cette confiance, on a recours à une *estimation par intervalle de confiance. Si l'on connaît la valeur de l'écart-type* $\sigma$, l'intervalle de confiance de m est défini par l'équation

$$P(\bar{x} - \lambda \cdot \sigma_{\bar{x}} \leqslant m \leqslant \bar{x} + \lambda \cdot \sigma_{\bar{x}}) = SE = 1 - \alpha$$
$$\text{avec } \sigma_{\bar{x}} = \sigma/\sqrt{n} \qquad 1.1.6$$

où $\alpha$ désigne le *niveau de confiance* et P la probabilité, tandis que $\lambda$ est une fonction de $\alpha$ donnée par les tables de la distribution normale standard.

L'écart-type de $\bar{x}$ est $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Le *niveau de confiance* $\alpha$ peut être choisi arbitrairement; $SE = 1 - \alpha$ désigne la *sécurité.* La relation (1.1.6) est obtenue très facilement à

l'aide de la *théorie des fonctions caractéristiques* qui nous montre que $\frac{(\bar{x} - m)}{\sigma_{\bar{x}}}$ a pour distribution N (0;1), ce qui fait qu'on a

$$P\left(a \leqslant \frac{\bar{x} - m}{\sigma_{\bar{x}}} \leqslant b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}y^2} dy = \Phi(b) - \Phi(a) \qquad 1.1.7$$

La table (1.1.8) nous donne quelques valeurs pour l'intervalle de confiance de m. 1.1.8

Intervalle de confiance de la moyenne m de la distribution N $(m;\sigma^2)$

| Intervalle | Sécurité SE = 100 - α % | Niveau de confiance α % |
|---|---|---|
| $\bar{x} \pm 1{,}645\ \sigma/\sqrt{n}$ | 90 % | 10 % |
| $\bar{x} \pm 1{,}960\ \sigma/\sqrt{n}$ | 95 % | 5 % |
| $\bar{x} \pm 2{,}576\ \sigma/\sqrt{n}$ | 99 % | 1 % |
| $\bar{x} \pm 3{,}291\ \sigma/\sqrt{n}$ | 99,9 % | 0,1 % |
| $\bar{x} \pm 3{,}891\ \sigma/\sqrt{n}$ | 99,99 % | 0,01 % |

Ce tableau nous montre que pour un niveau de confiance α donné, l'intervalle de confiance diminue lorsque n augmente, ce qui a pour conséquence une diminution de l'insécurité de $\bar{x}$, c'est-à-dire une augmentation de la précision de $\bar{x}$ avec la racine carrée de n. Ceci n'est rien d'autre qu'une conséquence de la *loi de la propagation des erreurs.*

En ne connaissant rien de la statistique, on part généralement de l'idée que la valeur de m doit être située à l'intérieur de l'intervalle

$$(\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}})$$

mais il résulte de la théorie que nous venons d'esquisser qu'il n'y a que le 68 % des résultats qui remplissent cette condition; nous avons en effet

$$P(\bar{x} - 1\sigma_{\bar{x}} \leqslant m \leqslant \bar{x} + 1\sigma_{\bar{x}}) \approx 68\% \qquad 1.1.9$$

$$P(\bar{x} - 2\sigma_{\bar{x}} \leqslant m \leqslant \bar{x} + 2\sigma_{\bar{x}}) \approx 95\%$$

$$P(\bar{x} - 3\sigma_{\bar{x}} \leqslant m \leqslant \bar{x} + 3\sigma_{\bar{x}}) \approx 99{,}7\%$$

Comme il est souhaitable d'avoir *au moins* une sécurité de 95 %, nous sommes obligés de compter avec des écarts de $\pm 2\,\sigma_{\bar{x}}$ de part et d'autre de la valeur compensée $\bar{x}$.

La situation devient encore plus défavorable si nous prenons *l'ellipse d'erreur moyenne,* qui est d'un emploi fréquent en *triangulation* et en *trilatération. Dans ce cas, la probabilité qu'un point tombe à l'intérieur de l'ellipse n'est que de 39 %.* On peut en effet démontrer que

Si nous rapportons l'ellipse d'erreur moyenne à ses axes p et q et que $m_x$, $m_y$ désignent les coordonnées vraies du point à déterminer, la probabilité $P(\lambda^2)$ qu'un point (x, y) tombe à l'intérieur de l'ellipse

$$\left(\frac{x - m_x}{p}\right)^2 + \left(\frac{y - m_y}{q}\right)^2 = \lambda^2 \qquad 1.1.10$$

est donnée par

$$P(\lambda^2) = 1 - e^{-\frac{1}{2}\lambda^2}$$

Le tableau ci-après nous donne quelques valeurs de $P(\lambda^2)$.

| λ | 0,76 | 1,18 | 1,67 | 2,15 | 2,45 | 3,04 | 1.1.11 |
|---|---|---|---|---|---|---|---|
| $P(\lambda^2)$ | 25 % | 50 % | 75 % | 90 % | 95 % | 99 % | |

Il en résulte que *si nous voulons avoir une sécurité de 95 %,* qui est pratiquement souhaitable, *nous devons agrandir l'ellipse d'erreur moyenne 2,45 fois,* nécessité dont on oublie souvent de tenir compte dans les applications pratiques.

Faisons encore une remarque au sujet des fautes (Ausreisser) qu'on peut trouver dans une série d'observations. La Probabilité qu'une mesure x s'écarte de plus de $3\sigma$ de la moyenne empirique est d'environ 3 ‰; vu que cette probabilité est très faible, on a l'habitude de rejeter ces mesures. Cette façon de faire est cependant assez arbitraire, car si l'on étudie le problème de plus près en ayant recours à la *statistique d'ordre,* qui est relativement récente, on comprend plus facilement pour quelle raison on a fréquemment des difficultés en calculant l'ajustement d'une distribution.

### *1.2 Distribution t de Student*

La formule (1.1.6), que nous venons de mentionner, a cependant un défaut vu qu'elle fait intervenir l'écart-type σ de la population, qui sera généralement inconnu. On peut remédier à cet inconvénient en ayant recours à la *distribution t de Student,* dont la *fonction de fréquence* est représentée sur la figure (1.2.1) pour différents *degrés de liberté* $f = n - 1$.

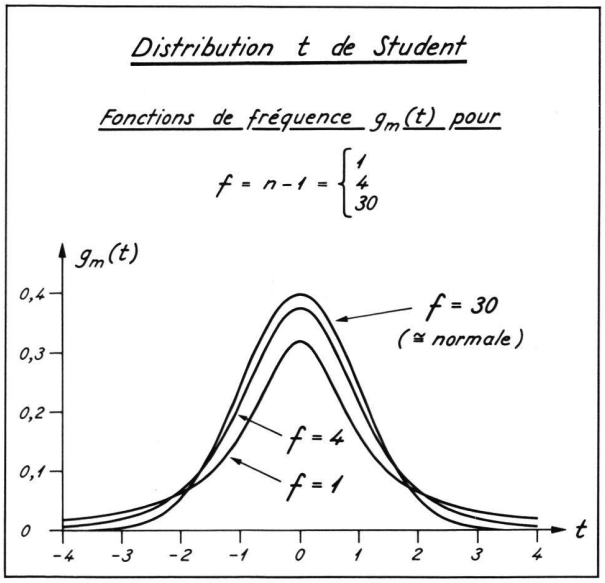



Fig. 1.2.1

Voici comment on applique cette dernière distribution:

*Règle mnémotechnique*

Pour déterminer l'intervalle de confiance de la moyenne m d'une population normale N $(m;\sigma^2)$ *à variance inconnue,* au niveau de confiance α et $f = n - 1$ *degrés de liberté,* on extrait de la population un échantillon $x_1, x_2, \ldots, x_n$ avec lequel on calcule d'abord la variance empirique $S^2$ de x, puis celle de $\bar{x}$, soit $S_{\bar{x}}^2$, en appliquant les formules

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{\bar{x}}^2 = \frac{1}{n} S^2$$

Puis on choisit un niveau de confiance $\alpha$ et on cherche dans une table de la *fonction de distribution* de t la valeur de $t_{1-\alpha/2}$ en fonction de $\alpha$ et de f = n – 1. Cela étant, on a la formule

$$P(\bar{x} - t_{1-\alpha/2}\, S_{\bar{x}} \leqslant m \leqslant \bar{x} + t_{1+\alpha/2}\, S_{\bar{x}}) = 1 - \alpha \qquad 1.2.1$$

qui nous donne le domaine de confiance de m au niveau $\alpha$. Notons que (1.2.1) nous donne un domaine de confiance symétrique par rapport à $\bar{x}$, qu'on a obtenu en coupant à chaque extrémité de la distribution t la queue de probabilité $\alpha/2$ ; voir fig. 1.2.2.

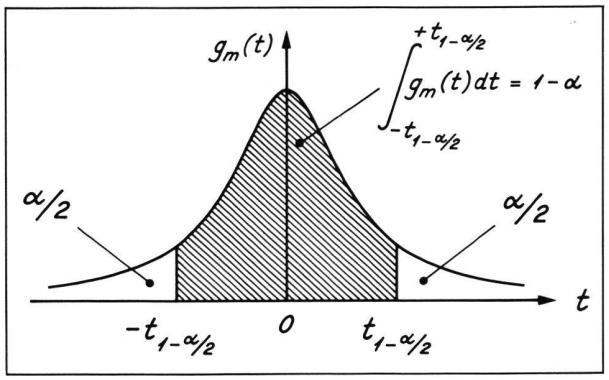



Fig. 1.2.2

Donnons quelques valeurs de cette table afin de pouvoir les comparer avec les résultats (1.1.8) précédemment obtenus.

1.2.2

Valeurs de $t_{1-\alpha/2}$

| Niveau de confiance | $\alpha$% → | 10 % | 5 % | 1 % |
|---|---|---|---|---|
| Sécurité | (1 - $\alpha$)% → | 90 % | 95 % | 99 % |
| f = n - 1 = 2 | | 2,920 | 4,303 | 9,925 |
| 5 | | 2,015 | 2,571 | 4,032 |
| 10 | | 1,812 | 2,228 | 3,169 |
| 15 | | 1,753 | 2,131 | 2,947 |
| 20 | | 1,725 | 2,086 | 2,845 |

Un examen rapide de ce tableau nous montre que

a) quelle que soit la taille n, les valeurs de $t_{1-\alpha/2}$ sont toujours supérieures à celles du tableau (1.1.8) pour un niveau de confiance $\alpha$ donné. Ceci est évident, car on possède une plus petite quantité d'informations que précédemment du moment que $\sigma$ est inconnue.

b) pour les conditions qui devraient être réalisées dans la pratique, soit $\alpha$ = 5 %, nous obtenons par exemple

avec n = 3 f = n – 1 = 2 $t_{1-\alpha/2}$ = 4,303 et
avec n = 6 f = n – 1 = 5 $t_{1-\alpha/2}$ = 2,571

### *1.3 Distribution $\chi^2$*

Si l'on a affaire à une population normale $X \sim N(m;\sigma^2)$, dont on ne connaît ni la moyenne m ni la variance $\sigma^2$, on en extrait un échantillon de taille n, soit $x_1, x_2, \ldots, x_n$. Cela étant, on calcule la moyenne empirique

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \qquad 1.3.1$$

qui est, nous l'avons vu plus haut, un estimateur sans biais de m. On calcule ensuite les grandeurs

$$v_i = \bar{x} - x_i \qquad i = 1, 2, \ldots, n \qquad 1.3.2$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2 = \frac{[vv]}{n-1} \qquad 1.3.3$$

$S^2$ étant par définition la *variance empirique;* c'est un estimateur sans biais de $\sigma^2$. *Pour connaître la précision de* $S^2$, on a besoin de la distribution $\chi^2$, dont la *fonction de fréquence* est représentée par la figure (1.3.1) pour différentes valeurs de f = n–1.

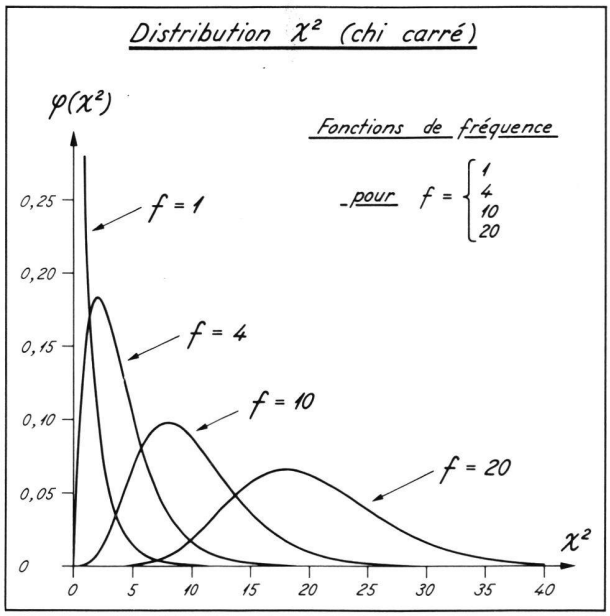



Fig. 1.3.1

L'utilisation pratique de cette distribution est facilitée par l'emploi de tables numériques qui donnent les valeurs critiques de $\chi^2$ en fonction du niveau de confiance $\alpha$ et du nombre de degrés de liberté f = n – 1. Pour calculer l'intervalle de confiance de $\sigma^2$, on coupe à chaque extrémité de la distribution une queue de probabilité $\alpha/2$; voir fig. (1.3.2).

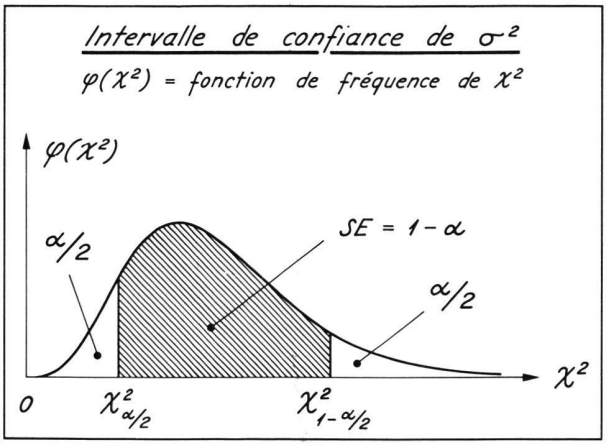



Fig. 1.3.2

*Règle mnémotechnique*

Pour trouver l'intervalle de confiance de $\sigma^2$ en fonction de $S^2$ et du niveau de confiance $\alpha$, on extrait de la table

de la distribution $X^2$ les valeurs critiques $x^2_{1-\alpha/2}$ et $x^2_{\alpha/2}$. Cela étant, l'intervalle de confiance est donné par la relation

$$\frac{n-1}{x^2_{1-\alpha/2}} S^2 \leqslant \sigma^2 \leqslant \frac{n-1}{x^2_{\alpha/2}} S^2 \qquad 1.3.4$$

*Application*

Pour $\alpha = 5\,\%$, les tables de la distribution $\chi^2$ nous donnent

| | $\alpha = 5\%$ | | | |
|---|---|---|---|---|
| $f = n-1$ | $x^2_{\alpha/2}$ | $x^2_{1-\alpha/2}$ | $\sqrt{\frac{f}{x^2_{1-\alpha/2}}}$ | $\sqrt{\frac{f}{x^2_{\alpha/2}}}$ |
| 2 | 0,051 | 7,38 | 0,52 | 6,26 |
| 3 | 0,216 | 9,35 | 0,57 | 3,73 |
| 4 | 0,484 | 11,1 | 0,60 | 2,87 |
| 5 | 0,831 | 12,8 | 0,63 | 2,45 |
| 6 | 1,24 | 14,4 | 0,65 | 2,20 |
| 7 | 1,69 | 16,0 | 0,66 | 2,04 |
| 8 | 2,18 | 17,5 | 0,68 | 1,92 |
| 9 | 2,70 | 19,0 | 0,69 | 1,83 |
| 10 | 3,25 | 20,5 | 0,70 | 1,75 |
| 11 | 3,82 | 21,9 | 0,71 | 1,70 |
| 12 | 4,40 | 23,3 | 0,72 | 1,65 |
| 13 | 5,01 | 24,7 | 0,73 | 1,61 |
| 14 | 5,63 | 26,1 | 0,73 | 1,58 |
| 15 | 6,26 | 27,5 | 0,74 | 1,55 |

d'où nous tirons par exemple

pour $n = 4$ : $f = 3$ $\quad 0{,}57\ S \leqslant |\sigma| \leqslant 3{,}73\ S$

pour $n = 12$ : $f = 11$ $\quad 0{,}71\ S \leqslant |\sigma| \leqslant 1{,}70\ S$

Vu que la taille des échantillons utilisés en géodésie est généralement très petite, nous devons compter pour $n = 4$ avec un domaine de confiance de

$0{,}6\ S \leqslant |\sigma| \leqslant 3{,}7\ S$ au niveau $\alpha = 5\,\%$.

*1.4 Distribution F de Snedecor*

Lorsqu'il s'agit de savoir si deux équipements de mesure donnent la même précision ou non, on emploie la distribution F de Snedecor, qui est également donnée par des tables numériques. Soient

| | | | |
|---|---|---|---|
| $x_1, x_2, \ldots, x_m$ | $f_1 = m - 1$ | Instrument 1 | 1.4.1 |
| $y_1, y_2, \ldots, y_n$ | $f_2 = n - 1$ | Instrument 2 | |

les valeurs obtenues en mesurant avec l'instrument 1 une inconnue X m fois et avec l'instrument 2 une inconnue Y n fois, les degrés de liberté $f_1 = m - 1$ et $f_2 = n - 1$ n'étant pas nécessairement les mêmes. On calcule ensuite les variances empiriques. 1.4.2

$$S_1^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2 \qquad S_2^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$f_1 = m - 1$ $\qquad$ $f_2 = n - 1$

ainsi que leur rapport, en plaçant au numérateur la plus grande des deux valeurs $S_1^2$ et $S_2^2$. En supposant que ce soit $S_1^2$, nous calculerons donc

$$K = \frac{S_1^2}{S_2^2} \qquad 1.4.3$$

qui a une distribution F de Snedecor avec $f_1 = m - 1$ degrés de liberté au numérateur et $f_2 = n - 1$ degrés de liberté au dénominateur. Si les deux instruments ont la même précision, on doit trouver $K \approx 1$. Si tel n'est pas le cas, il s'agit de savoir si la différence constatée est significative au niveau de confiance $\alpha$ ou non.

*Exemple*

Soient $e_1$ et $e_2$ deux côtés de polygone qu'on a mesurés plusieurs fois avec une mire horizontale en invar. Le tableau ci-après nous donne les angles parallactiques, exprimés en $^{cc}$, ainsi que les estimateurs $\bar{x}$ et $\hat{\sigma} = S$ de chacun d'eux.

| Côté $e_1$ | Côté $e_2$ |
|---|---|
| $11558^{cc}$ | $7495^{cc}$ |
| 11553 | 7491 |
| 11552 | 7493 |
| 11544 | 7495 |
| 11542 | 7492 |
| 11562 | 7488 |
| 11550 | |
| 11547 | |
| $\bar{x} = 11551{,}00^{cc}$ | $\bar{x} = 7492{,}33^{cc}$ |
| $S_1 = 6{,}78^{cc}$ | $S_2 = 2{,}66^{cc}$ |
| $n_1 = 8$ | $n_2 = 6$ |
| $f_1 = 7$ | $f_2 = 5$ |

Il s'agit de savoir si la différence entre les deux écarts-types empiriques $S_1$ et $S_2$ est significative au niveau $\alpha = 5\,\%$. Nous obtenons

$f_1 = 7 \quad S_1 = 6{,}78 \quad S_1^2 = 45{,}97$
$f_2 = 5 \quad S_2 = 2{,}66 \quad S_2^2 = 7{,}08$ $\Big|\ K = \frac{S_1^2}{S_2^2} = \underline{6{,}50}$

La table de la distribution F nous donne pour $f_1 = 7$, $f_2 = 5$, $\alpha = 5\,\%$:

$F = 4{,}88$

Comme on a $K > F$, la différence est largement significative au niveau $\alpha = 5\,\%$.

Au tableau (1.4.4), nous avons indiqué quelques valeurs de F et $\sqrt{F}$ pour les niveaux $\alpha = 5\,\%$ et $\alpha = 2{,}5\,\%$ *en supposant* $f_1 = f_2$. Il nous donne par exemple *pour* $\alpha = 5\,\%$ *et* $n = 5$*:*

$f_1 = f_2 = 4 \quad \sqrt{F} = 2{,}53$, ce qui veut dire que si l'on a $\frac{S_1}{S_2} < 2{,}53$

la différence entre $S_1$ et $S_2$ n'est pas significative. On doit donc admettre dans ce cas que les deux instruments ont la même précision.

Il en résulte que si nous voulons comparer la précision de deux instruments, nous devons avoir recours à des échantillons de taille suffisamment grande, par exemple $n \geqslant 20$.

| | $\alpha$ = 5 % | | $\alpha$ = 2,5 % | |
|---|---|---|---|---|
| $f_1 = f_2$ | F | $\sqrt{F}$ | F | $\sqrt{F}$ |
| 2 | 19,0 | 4,36 | 39,0 | 6,25 |
| 3 | 9,28 | 3,05 | 15,4 | 3,92 |
| 4 | 6,39 | 2,53 | 9,60 | 3,10 |
| 5 | 5,05 | 2,25 | 7,15 | 2,67 |
| 6 | 4,28 | 2,07 | 5,82 | 2,41 |
| 7 | 3,79 | 1,95 | 4,99 | 2,23 |
| 8 | 3,44 | 1,85 | 4,43 | 2,11 |
| 9 | 3,18 | 1,78 | 4,03 | 2,01 |
| 10 | 2,98 | 1,73 | 3,72 | 1,93 |
| 11 | 2,82 | 1,68 | 3,47 | 1,86 |
| 12 | 2,69 | 1,64 | 3,28 | 1,81 |
| 13 | 2,58 | 1,61 | 3,12 | 1,77 |
| 14 | 2,48 | 1,58 | 2,98 | 1,73 |
| 15 | 2,40 | 1,55 | 2,86 | 1,69 |

# Chapitre 2

## Statistique d'ordre

### *2.1 Introduction*

Les statistiques d'ordre sont des fonctions des observations qui tiennent compte de l'*ordre* ou de la *valeur* des observations.

Ainsi, si nous avons mesuré une inconnue X à n reprises, le résultat consiste en un *échantillon*

$x_1, x_2, ..., x_n$

de taille n. En plaçant ses éléments $x_1, x_2, ..., x_n$ dans l'ordre croissant des valeurs, nous obtenons un *échantillon ordonné*, que nous désignons par

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(i)} \leqslant \cdots \leqslant X_{(n)} \qquad 2.1.1$$

et nous dirons que $X_{(i)}$ est la *statistique d'ordre i.*

Les éléments $x_1, x_2, ..., x_n$ de l'échantillon sont *indépendants* et ont la même distribution que la population X. Les statistiques d'ordre $X_{(1)}, X_{(2)}, ..., X_{(n)}$, par contre, sont dépendantes vu que la probabilité de chacune d'elles dépend de la réalisation des précédentes.

La statistique d'ordre a pour but l'étude des variables ordonnées (2.1.1) et de leurs fonctions. Les grandeurs les plus importantes sont les *valeurs extrêmes* $X_{(1)}$, $X_{(n)}$ et le *range* $w = X_{(n)} - X_{(1)}$, qu'on appelle aussi *étendue* de l'échantillon.

Notons en passant que les valeurs extrêmes interviennent entre autres dans l'*étude de la sécheresse*, de l'*écoulement des fluides*, de la *fatigue des matériaux*, etc. Mais elles peuvent également jouer un certain rôle en topographie.

Le range $w$ permet une *estimation rapide* de l'écart-type $\sigma$. On y a recours pour le contrôle du rendement ou de la qualité de produits fabriqués en grande quantité.

Les valeurs extrêmes constituent la base pour l'étude des „outliers" (Ausreisser) en ce sens que si $X_{(n)}$ s'écarte trop de la moyenne empirique

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n)$$

de l'échantillon, il y a nécessairement des éléments dans l'échantillon qui accusent des *anomalies;* nous disons dans ce cas qu'ils ont été *contaminés.*

Pour l'*estimation des paramètres d'une population,* on se sert de *fonctions linéaires de l'échantillon ordonné* en ayant recours au *théorème de Gauss-Markov relatif à la méthode des moindres carrés.* La statistique d'ordre se révèle particulièrement efficace lorsqu'une ou plusieurs mesures doivent être rejetées, car dans une telle situation les méthodes classiques sont laborieuses et conduisent souvent à des conclusions douteuses.

Les test sur la *longévité* sont une illustration idéale de l'application de la statistique d'ordre. En effet, vu que ces observations sont toujours de très longue durée, on est souvent obligé de les interrompre après un certain laps de temps et de les reprendre ultérieurement. Dans ces conditions, la statistique d'ordre permet facilement de tenir compte après coup des observations complémentaires, ce qui n'est pas le cas de la méthode classique.

Au cours de ces dernières années, la statistique d'ordre a reçu de nouvelles impulsions dans différentes directions, grâce aux *ordinateurs*, qui ont permis d'étudier les observations sous différents angles, ce qui a conduit à une véritable *analyse des observations.*

Nous voyons donc que la statistique d'ordre tient compte, comme la statistique classique, de la *valeur* des observations, et de ce fait le résultat final dépendra de la distribution de la population. Mais il y a possibilité de s'en rendre indépendant en ayant recours à une théorie plus générale, qui est la *statistique de rang.*

En appliquant la statistique d'ordre, on perd naturellement, comme avec toute statistique du reste, une certaine quantité de l'information. Ce fait est dû à la méthode de calcul même. Mais par rapport à la méthode classique, la statistique d'ordre a l'avantage d'être *très rapide* si l'on dispose de tables numériques appropriées; en outre, elle dépend moins des conditions initiales (suppositions) que la méthode classique.

### *2.2 Distribution de la valeur la plus grande $X_{(n)}$ d'un échantillon*

La valeur la plus grande $X_{(n)}$ d'un échantillon de taille n étant d'un emploi fréquent, nous allons chercher sa distribution. Voici ce qu'il faut entendre par là: *pour chaque échantillon de taille n,* nous pouvons former une suite ordonnée $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ avec les mesures. Il est évident que la valeur la plus grande $X_{(n)}$ varie d'un échantillon à l'autre, et par conséquent $X_{(n)}$ est une variable aléatoire ayant une certaine fonction de distribution, que nous allons calculer à partir de la distribution de la population X.

Soit v une *valeur fixe quelconque* de X; voir fig. (2.2.1).

Pour que l'événement $X_{(n)} < v$ se réalise, il faut que *toutes* les valeurs $x_1, x_2, ..., x_n$ de l'échantillon soient

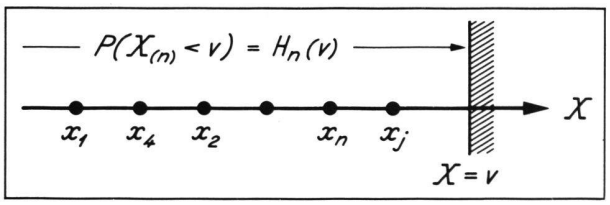



Fig. 2.2.1

inférieures à v. La probabilité P de cet événement est donnée par

$$P(X_{(n)} < v) = P(x_1 \text{ et } x_2 \text{ et } \cdots \text{ et } x_n < v) \qquad 2.2.1$$

Cette formule nous montre que la probabilité de l'événement $X_{(n)} < v$ est une fonction de v. Pour la calculer, il suffit de remarquer qu'on a

$$P(x_i < v) = F(v) \qquad i = 1,2, \ldots, n \qquad 2.2.2$$

où F(v) désigne la fonction de distribution de la population X. Mais comme les observations $x_1, x_2, \ldots, x_n$ sont indépendantes, nous avons

$$P(X_{(n)} < v) = P(x_1 < v) \cdot P(x_2 < v) \cdot \ldots \cdot P(x_n < v) \qquad 2.2.3$$

ce qui nous donne avec (2.2.2)

$$P(X_{(n)} < v) = H_n(v) = \{ F(v) \}^n \qquad 2.2.4$$

Nous constatons donc que

2.2.5

La probabilité de l'événement $X_{(n)} < v$ est complètement déterminée par la fonction de distribution F (x) de la population.

et qu'elle peut facilement être calculée à l'aide de la formule (2.2.4).
La figure (2.2.2) nous montre la représentation graphique de $H_n(v)$ pour la population normale standard N (0;1).

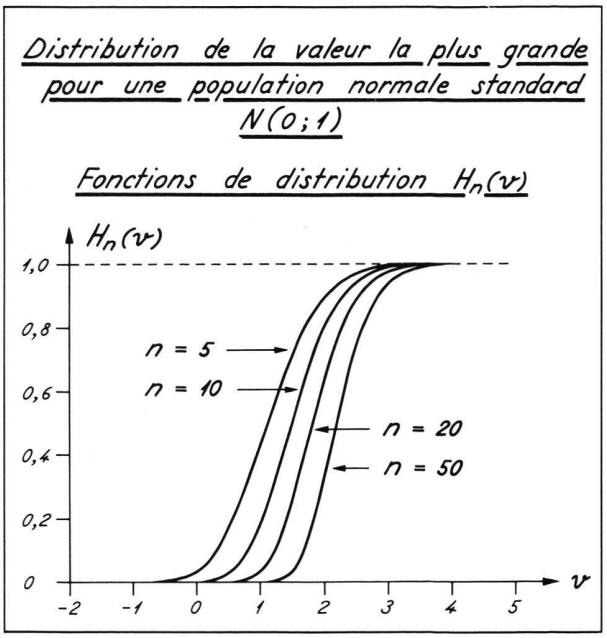


Fig. 2.2.2

Vu qu'on a pour n'importe quelle distribution $0 \leqslant F(v) \leqslant 1$, il résulte de (2.2.4) que pour une valeur fixe quelconque de v la probabilité de l'événement $X_{(n)} < v$ diminue lorsque la taille n de l'échantillon augmente, ce qui est du reste évident.

Pour obtenir la *densité de probabilité* $h_n(v)$ de $X_{(n)}$, il suffit de dériver (2.2.4) par rapport à v, ce qui nous donne

$$h_n(v) = n \cdot \{ F(v) \}^{n-1} f(v) \qquad 2.2.6$$

La figure (2.2.3) nous montre la représentation graphique de $h_n$ (v) pour la population normale standard $X \sim N(0;1)$.

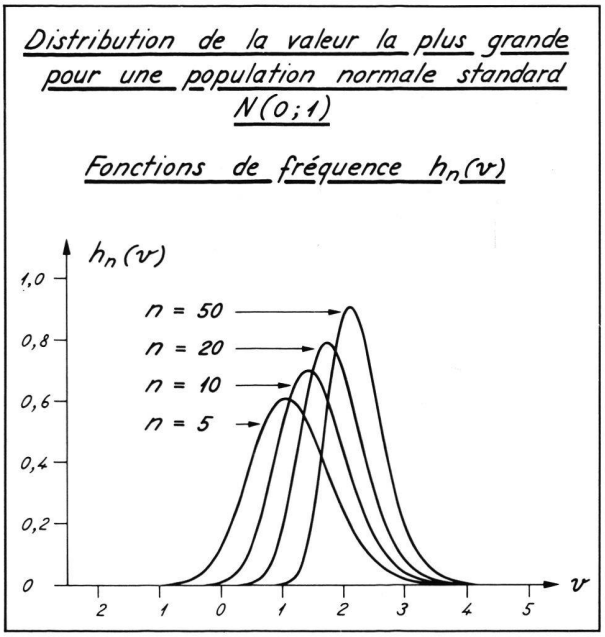


Fig. 2.2.4

Ce résultat étant acquis, nous voulons calculer l'*intervalle de confiance* de $X_{(n)}$. En vertu de (2.2.4) et (2.2.6), nous avons

$$P(v < X_{(n)} < v + dv) = h_n(v)\, dv \qquad 2.2.7$$

ou bien, en vertu de (2.2.6),

$$P(v < X_{(n)} < v + dv) = n \cdot \{ F(v) \}^{n-1} f(v)\, dv \qquad 2.2.8$$

où F(v) désigne la fonction de distribution de la population, tandis que f(v) est sa fonction de fréquence.
Afin d'intégrer l'équation (2.2.8), nous introduisons la substitution

$$t = F(v) \qquad 0 \leqslant t \leqslant 1 \qquad 2.2.9$$

F(v) étant une fonction monotone croissante, la substitution (2.2.9) est bi-univoque; elle fait correspondre à chaque valeur de v une et une seule valeur de t et vice versa pour $0 \leqslant t \leqslant 1$; voir figure (2.2.4).

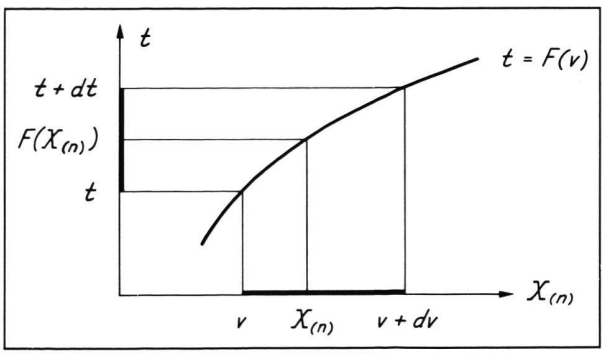


Fig. 2.2.3

En vertu de (2.2.8), la masse aléatoire dm, portée par le segment (v, v + dv), est

2.2.10

$$dm = P(v < X_{(n)} < v + dv) = n \cdot \{F(v)\}^{n-1} f(v)\, dv$$

ce qui nous donne avec (2.2.9)

$$dm = P(v < X_{(n)} < v + dv) = n \cdot t^{n-1} dt \qquad 2.2.11$$

Mais vu que la transformation (2.2.9) est bi-univoque, nous avons nécessairement

$$P(t < F(X_{(n)}) < t + dt) = P(v < X_{(n)} < v + dv) \qquad 2.2.12$$

d'où il résulte avec (2.2.11)

$$P(t < F(X_{(n)}) < t + dt) = n \cdot t^{n-1} dt \qquad 2.2.13$$

En intégrant cette équation par rapport à t entre les limites 0 et $\beta$, nous obtenons

$$P\{F(X_{(n)}) \leq \beta\} = \beta^n \qquad 2.2.14$$

ou bien, en prenant la probabilité complémentaire,

$$P\{F(X_{(n)}) > \beta\} = 1 - \beta^n \qquad 0 < \beta \leq 1 \qquad 2.2.15$$

*où F désigne comme précédemment la fonction de distribution de la population X.* Cette formule est tout à fait remarquable vu qu'elle est valable pour *toute* distribution F(X).
On obtient une interprétation stochastique de (2.2.15) en remarquant que

2.2.16

$F(X_{(n)})$ donne la fraction d'individus de la population X, dont la valeur est inférieure ou égale à l'élément le plus grand $X_{(n)}$ de l'échantillon.

Nous en concluons que

2.2.17

$F(X_{(n)}) > \beta$ signifie que la fraction des individus de la population X ayant une valeur x inférieure ou égale à l'élément le plus grand $X_{(n)}$ de l'échantillon est *au moins* égale à $\beta$.

Par conséquent $F(X_{(n)}) > \beta$ désigne un événement, dont la probabilité est donnée par (2.2.15). L'ensemble des valeurs $(-\infty < x < X_{(n)})$ est parfois appelé *„intervalle de tolérance statistique unilatéral"*.
Pour mieux faire ressortir l'importance pratique de la formule (2.2.15), nous allons considérer un

*Exemple numérique*

Lors de la vérification d'un plan topographique, on a contrôlé n = 46 points en notant les écarts planimétriques (erreurs) $f_1$, $f_2$, ..., $f_{46}$, et soit $X_{(n)}$ l'écart *maximum* qu'on a ainsi mis en évidence. Il s'agit alors de savoir dans quelle mesure cette valeur $X_{(n)}$ nous renseigne sur la précision de l'ensemble du lever. Dans ce but, nous introduisons l'hypothèse

2.2.18

H: au moins le 95 % des erreurs de tout le lever a une valeur x inférieure ou égale à $X_{(n)}$.

La probabilité que cette hypothèse soit vraie peut être calculée à l'aide de la formule (2.2.15) et nous obtenons, en prenant par exemple $\beta = 0.95$,

$$P(H = \text{vraie}) = 1 - \beta^n = 0{,}9056 \approx 91\% \qquad 2.2.19$$

ce qui nous permet de conclure que

2.2.20

Si l'on effectue 46 mesures de contrôle, il y a 91 % de chances qu'au moins le 95 % de toutes les erreurs du lever aient une valeur x inférieure ou égale à l'écart maximum $X_{(n)}$, mis en évidence par les mesures de contrôle.

Cette méthode d'estimation permet au vérificateur de se rendre facilement compte, au cours de la campagne de terrain, si le nombre de mesures qu'il a effectuées est suffisant ou non. En utilisant de plus un monogramme approprié (voir figure 2.2.5), il est possible de tirer des conclusions sans effectuer aucun calcul.

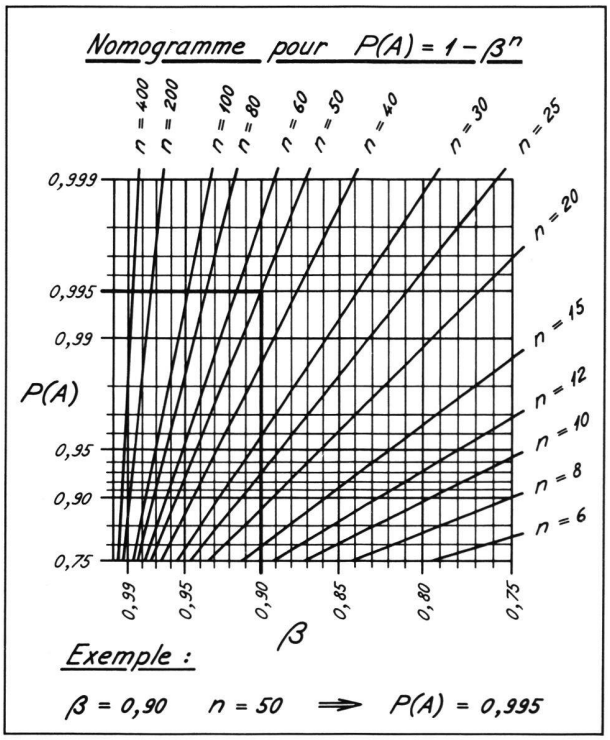



Fig. 2.2.5

La méthode que nous venons d'esquisser trouve des applications dans les domaines les plus divers et notamment aussi dans l'*essai des matériaux*.

*2.3 Estimation de l'écart-type σ. Distribution du range réduit.*

Si l'on a un échantillon ordonné $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$, provenant d'une population X ayant pour variance $\sigma^2$, on appelle $w = X_{(n)} - X_{(1)}$ *range* et

$$W = \frac{1}{\sigma} w = \frac{1}{\sigma} \{X_{(n)} - X_{(1)}\} \qquad 2.3.1$$

*range réduit* de l'échantillon. Si f(X) est la fonction de fréquence d'une *variable aléatoire réduite* X, l'intégrale de probabilité de W, pour un échantillon de taille n, s'écrit

$$P(W) = n\int_{-\infty}^{+\infty} f(X)\left[\int_{X}^{X+W} f(u)\,du\right]^{n-1} dX \qquad 2.3.2$$

L'application pratique de cette formule ne devient cependant possible que si l'on connaît l'écart-type $\sigma$ d'une mesure. Si tel n'est pas le cas, on prendra pour $\sigma$ une valeur approchée résultant d'une estimation.
Lorsqu'on a affaire à une population normale, l'application de la formule (2.3.2) devient très simple étant donné qu'on dispose de tables numériques qui nous donnent la valeur de P(W) en fonction de W et de n; voir par exemple *„Biometrika Tables for Statisticians" de Pearson & Hartley*, vol. I, table n° 23.
Afin de mieux faire ressortir l'utilité pratique de cette procédure, nous allons considérer un

*Exemple numérique*

Un angle parallactique a été mesuré 9 fois avec un théodolite ayant un écart-type $\sigma = 3^{cc}$. Les valeurs extrêmes de l'échantillon sont

$X_{(1)} = 0^{gr}\,84^{c}\,04^{cc}$

$X_{(n)} = 0^{gr}\,84^{c}\,16^{cc}$

$w = X_{(n)} - X_{(1)} = 12^{cc}$ (range)

$W = \frac{1}{\sigma}\,w = \frac{12^{cc}}{3^{cc}} = 4$ (range réduit)

$n = 9 \qquad P(W) = 0{,}8929$ (tables)

Dans ces conditions, la probabilité d'obtenir pour un échantillon de taille n = 9 un range $w \leqslant 12^{cc}$ est de 89 %.
Si l'on ne connaît ni la moyenne m ni la variance $\sigma^2$ d'une population X, on en prélève un échantillon $x_1, \ldots, x_n$, qui nous donne l'estimation ponctuelle

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad 2.3.3$$

Mais on peut aussi obtenir un estimateur de $\sigma$, soit $\hat{\hat{\sigma}}$ en n'utilisant que les valeurs extrêmes $X_{(1)}$ et $X_{(n)}$ de l'échantillon. On démontre en effet que

Si l'on a affaire à une distribution normale $(N(m;\sigma^2)$, dont on possède un échantillon ordonné

$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$

de taille n, la grandeur 2.3.4

$$\hat{\hat{\sigma}} = \frac{1}{d_n}\,w = \frac{1}{d_n}\,(X_{(n)} - X_{(1)})$$

est un estimateur sans biais de l'écart-type, les valeurs du coefficient $d_n$ étant données par la table (2.3.5).

2.3.5

$X \sim N(m;\sigma^2)$ $\qquad \hat{\hat{\sigma}} = \frac{1}{d_n}(X_{(n)} - X_{(1)})$

| n | $\frac{1}{d_n}$ | n | $\frac{1}{d_n}$ |
|---|---|---|---|
| 1 | – | 11 | 0,3152 |
| 2 | 0,8862 | 12 | 0,3069 |
| 3 | 0,5908 | 13 | 0,2998 |
| 4 | 0,4857 | 14 | 0,2935 |
| 5 | 0,4299 | 15 | 0,2880 |
| 6 | 0,3946 | 16 | 0,2831 |
| 7 | 0,3698 | 17 | 0,2787 |
| 8 | 0,3512 | 18 | 0,2747 |
| 9 | 0,3367 | 19 | 0,2711 |
| 10 | 0,3240 | 20 | 0,2677 |

Pour des échantillons de plus grande taille, voir „Biometrika Tables for Statisticians" vol. I, de Pearson & Hartley, éd. 3, tables 27.

*Exemple*

$n = 8 \qquad w = 20 \qquad \frac{1}{d_n} = 0{,}3512 \qquad \hat{\hat{\sigma}} = 7{,}02$

Nous voyons donc que l'estimation de l'écart-type à l'aide du range est excessivement simple. Elle a surtout été développée pour l'industrie, où le contrôle doit être fait par du personnel non spécialisé.

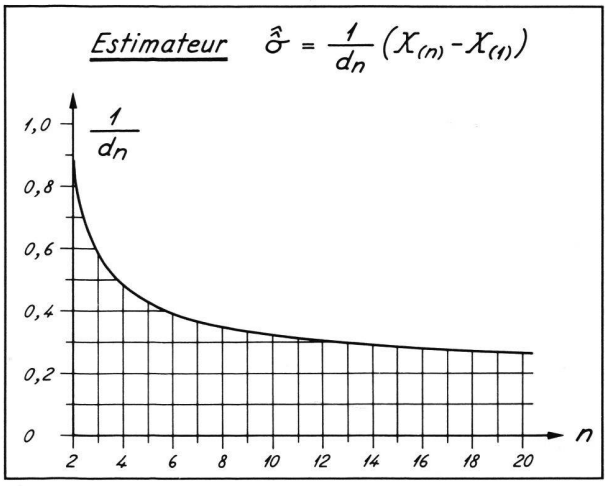



Fig. 2.3.1.

La figure (2.3.1) nous donne la représentation graphique de $1/d_n$ en fonction de la taille n de l'échantillon. On voit par exemple qu'on a

pour $n = 4 \qquad \hat{\hat{\sigma}} \approx 0{,}5\,w$

pour $n = 10 \qquad \hat{\hat{\sigma}} \approx 0{,}3\,w$

*Exemple*

On a mesuré un côté de polygone n = 8 fois avec une mire horizontale de 2 mètres. Les valeurs de l'angle parallactique X, exprimées en$^{cc}$, sont

$x_1 = 11558 \quad x_2 = 11553 \quad x_3 = 11552 \quad x_4 = 11544$

$x_5 = 11542 \quad x_6 = 11562 \quad x_7 = 11550 \quad x_8 = 11547$

L'estimation ponctuelle classique de la moyenne m et de l'écart-type $\sigma$ nous donne

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 11551{,}00^{cc}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 6{,}78^{cc}$$

En utilisant la statistique de range, nous obtenons

pour $n = 8 \qquad \frac{1}{d_n} = 0{,}3512$

$X_{(n)} = 11562$ | $w = X_{(n)} - X_{(1)} = 20$

$X_{(1)} = 11542$ |

$$\hat{\hat{\sigma}} = \frac{1}{d_n}(X_{(n)} - X_{(1)}) = 0{,}3512 \times 20 = 7{,}02^{cc}$$

Afin de comparer l'estimation classique $\hat{\sigma}$ de $\sigma$ avec celle $\hat{\hat{\sigma}}$ fournie par la statistique d'ordre, nous avons traité 40 échantillons par les deux méthodes. Il s'agit de 40 côtés de polygone, dont chacun a été mesuré entre 6 et 10 fois avec la mire horizontale en invar. $\sigma$ désigne dans ce cas l'écart-type de l'angle parallactique. Comme il dépend de la longueur du côté, les échantillons proviennent de populations distinctes. Les résultats obtenus sont indiqués au tableau (2.3.6), où nous avons

$N^o$ = numéro de l'échantillon.
n = taille de l'échantillon.
$\hat{\sigma}$ = estimation classique de $\sigma$, exprimée en $^{cc}$.
$\hat{\hat{\sigma}}$ = estimation fournie par la statistique d'ordre, en $^{cc}$.
$\frac{|\hat{\sigma} - \hat{\hat{\sigma}}|}{\hat{\sigma}} \cdot 100$ = écart en % entre les deux estimations

2.3.6

| No | n | $\frac{\lvert\hat{\sigma}-\hat{\hat{\sigma}}\rvert}{\hat{\sigma}}\cdot 100$ | No | n | $\frac{\lvert\hat{\sigma}-\hat{\hat{\sigma}}\rvert}{\hat{\sigma}}\cdot 100$ |
|---|---|---|---|---|---|
| 1 | 6 | 3,76 | 21 | 6 | 12,38 |
| 2 | 8 | 3,54 | 22 | 8 | 13,30 |
| 3 | 6 | 16,61 | 23 | 9 | 16,54 |
| 4 | 8 | *19,25* | 24 | 7 | 6,20 |
| 5 | 8 | 14,08 | 25 | 8 | 6,15 |
| 6 | 9 | 7,76 | 26 | 8 | 2,93 |
| 7 | 8 | 17,84 | 27 | 8 | 13,30 |
| 8 | 7 | 16,03 | 28 | 8 | 16,19 |
| 9 | 8 | 11,43 | 29 | 8 | 3,95 |
| 10 | 9 | *27,99* | 30 | 8 | *19,21* |
| 11 | 9 | 4,44 | 31 | 7 | 4,94 |
| 12 | 8 | 6,96 | 32 | 8 | 2,12 |
| 13 | 8 | 6,33 | 33 | 8 | 10,31 |
| 14 | 8 | 0,00 | 34 | 8 | 4,69 |
| 15 | 8 | 8,00 | 35 | 9 | 7,47 |
| 16 | 8 | 7,03 | 36 | 9 | 14,88 |
| 17 | 8 | 4,09 | 37 | 10 | 13,90 |
| 18 | 8 | 1,93 | 38 | 8 | 0,00 |
| 19 | 8 | 3,95 | 39 | 8 | 1,74 |
| 20 | 8 | 12,06 | 40 | 8 | 6,36 |

$$\left\{ \frac{|\hat{\sigma} - \hat{\hat{\sigma}}|}{\hat{\sigma}} \cdot 100 \right\}_{\text{moyenne}} = 9{,}24\,\%$$

L'écart moyen entre les deux estimations n'étant que de 9 %, nous constatons que l'estimation effectuée avec la statistique d'ordre est suffisamment précise pour pouvoir être appliquée dans la pratique. Pour l'échantillon n° 10, nous avons cependant un écart de 28 %, ce qui est inadmissible. Nous verrons par la suite que cette anomalie provient sans doute d'une ou de deux mesures qui s'écartent trop des autres, raison pour laquelle il y a lieu de les rejeter. Le rejet ne doit cependant pas s'effectuer arbitrairement; il est préférable d'avoir recours à une théorie appropriée que nous indiquerons au chapitre 3.

Vu que l'estimateur $\hat{\hat{\sigma}}$, fourni par la statistique d'ordre, ne dépend que des valeurs extrêmes $X_{(n)}$ et $X_{(1)}$, il est évident qu'une différence entre $\hat{\hat{\sigma}}$ et $\hat{\sigma}$ provient essentiellement de ces dernières valeurs. Il doit donc être possible d'établir une théorie de rejet d'observation basée sur la différence $\hat{\hat{\sigma}} - \hat{\sigma}$ respectivement sur le rapport $\hat{\hat{\sigma}}/\hat{\sigma}$.

# Chapitre 3

## Réjection d'observations

### *3.1 Généralités*

L'étude de la réjection d'observations est assez compliquée; aussi a-t-elle donné lieu à de nombreuses recherches. Elle a pour buts essentiels

a) l'élimination des „outliers" (Ausreisser).
b) l'analyse des mesures „contaminées", c'est-à-dire l'analyse des observations qui ne proviennent pas toutes de la même population; observations hétérogènes.
c) l'étude des observations s'écartant nettement d'une distribution moyenne, obliquité, etc.

Les résultats connus dans ce domaine ont surtout été obtenus grâce à la statistique d'ordre.
En mesurant une inconnue plusieurs fois, il arrive fréquemment qu'une ou deux mesures s'écartent plus ou moins des autres éléments de l'échantillon. On est alors obligé de les rejeter en admettant *qu'elles proviennent d'une autre population.* Souvent, l'analyse des mesures fait ressortir des éléments douteux dont on ne sait pas s'il y a lieu de les maintenir ou s'il est préférable de les rejeter. Mais en rejetant une mesure conforme (correcte), ou en conservant une mesure non-conforme (contaminée), on risque d'introduire un *biais* dans l'échantillon. Nous nous trouvons du reste dans la même situation lors du contrôle statistique de la qualité d'un produit fabriqué en grande quantité. Dans ce cas, on parle d'*erreurs du type I ou du type II,* notions bien connues en statistique classique.
Rappelons toutefois que le but de cette étude n'est pas la recherche d'*erreurs grossières* (fautes); il s'agit uniquement de savoir si les éléments d'un échantillon constituent un *ensemble homogène* ou si l'on doit admettre que certaines de ces mesures proviennent d'une autre population.
L'introduction d'un critère pour la réjection d'observations nous oblige à considérer au préalable les points suivants:

1) Du moment que chaque mesure nous fournit une information sur la population, il est *souhaitable* de conserver toutes les mesures. On n'éliminera donc que le strict nécessaire, et ceci uniquement lorsqu'on est quasi certain qu'on a affaire à une contamination.
2) Si l'on veut obtenir une population sans contamination, on doit ou bien rejeter les mesures contaminées si elles ont été reconnues comme telles, ou bien avoir recours à des méthodes d'estimation réduisant l'influence de la contamination au minimum.

Quoiqu'on fasse, il faut veiller à ce que la *procédure d'épuration* ne risque pas d'entraîner la réjection de mesures conformes.

### *3.2 Suppositions; conditions initiales*

En effectuant un essai, nous sommes obligés d'introduire au préalable un certain nombre de *suppositions* ou *conditions initiales.* Cela faisant, on admettra généralement que

1) la taille n de l'échantillon a été fixée.
2) une proportion $(1 - \gamma)$ des observations provient d'une distribution normale $N(m;\sigma^2)$.

3) une proportion $\gamma$ des mesures a une autre origine, soit qu'elle provienne
   a) de la population N $(m + \lambda\sigma;\sigma^2)$ ou
   b) de la population N $(m;\lambda^2\sigma^2)$

Une telle analyse nous permet de considérer une demi-douzaine de possibilités, qu'il serait cependant trop long d'examiner ici. Dans ce qui suit, nous nous bornerons à l'étude d'un seul cas, qui est celui où le range $w$ et la variance empirique S sont tirés d'un même échantillon. C'est du reste le cas qui est à la base du tableau (2.3.6).

*3.3 Distribution de $^{w}/_{S}$ si w et S proviennent du même échantillon*

Soit X une population normale N $(m;\sigma^2)$, de laquelle on possède un échantillon ordonné de taille n

$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$

Nous pouvons alors calculer le range $w = X_{(n)} - X_{(1)}$ et l'écart-type empirique

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

La distribution du rapport $^{w}/_{S}$ a été calculée en 1964 par *Pearson & Stephens,* qui ont publié une table donnant le point de pourcentage de $^{w}/_{S}$ en fonction de $w$ et n; voir „Biometrika Tables for Statisticians" vol. I, table 29 c. Nous donnons ci-après un court extrait de cette table.

3.3.1

| | Point de pourcentage supérieur du rapport $^{w}/_{S}$ | | | | | |
|---|---|---|---|---|---|---|
| n | 10,0 % | 5,0 % | 2,5 % | 1,0 % | 0,5 % | 0,0 % |
| 3 | 1,997 | 1,999 | 2,000 | 2,000 | 2,000 | 2,000 |
| 4 | 2,409 | 2,429 | 2,439 | 2,445 | 2,447 | 2,449 |
| 5 | 2,712 | 2,753 | 2,782 | 2,803 | 2,813 | 2,828 |
| 6 | 2,949 | 3,012 | 3,056 | 3,095 | 3,115 | 3,162 |
| 7 | 3,143 | 3,222 | 3,282 | 3,338 | 3,369 | 3,464 |
| 8 | 3,308 | 3,399 | 3,471 | 3,543 | 3,585 | 3,742 |
| 9 | 3,449 | 3,552 | 3,634 | 3,720 | 3,772 | 4,000 |
| 10 | 3,57 | 3,685 | 3,777 | 3,875 | 3,935 | 4,243 |

L'application de cette méthode à l'exemple considéré sous (2.3) a donné les résultats indiqués au tableau ci-après pour $\alpha = 5$ %. Dans celui-ci, $\left(^{w}/_{S}\right)_{\alpha=5\,\%}$ désigne le point de pourcentage supérieur, tiré de la table (3.3.1).

Si l'on a $^{w}/_{S} < (^{w}/_{S})_{\alpha}$, l'échantillon est conforme. Si nous optenons par contre $^{w}/_{S} \geqslant (^{w}/_{S})_{\alpha}$, l'échantillon est à examiner de plus près et l'une ou l'autre (éventuellement les deux) des valeurs extrêmes doit être rejetée. Lorsqu'on ne voit pas à priori laquelle des deux valeurs extrêmes doit être éliminée, on a la possibilité d'appliquer un autre test faisant intervenir la moyenne empirique $\bar{x}$.

| No | $w$ | S | $^{w}/_{S}$ | n | $(^{w}/_{S})$ $\alpha$=5% |
|---|---|---|---|---|---|
| 1 | 7 | 2,66 | 2,63 | 6 | 3,012 |
| 2 | 20 | 6,78 | 2,95 | 8 | 3,399 |
| 3 | 8 | 2,71 | 2,95 | 6 | 3,012 |
| 4 | 18 | 5,30 | *3,40* | 8 | 3,399 |
| 5 | 9 | 2,77 | 3,25 | 8 | 3,399 |
| 6 | 12 | 4,38 | 2,74 | 9 | 3,552 |
| 7 | 8 | 3,42 | 2,34 | 8 | 3,399 |
| 8 | 9 | 2,87 | 3,14 | 7 | 3,222 |
| 9 | 10 | 3,15 | 3,18 | 8 | 3,399 |
| 10 | 14 | 3,68 | *3,80* | 9 | 3,552 |
| 11 | 14 | 4,51 | 3,10 | 9 | 3,552 |
| 12 | 7 | 2,30 | 3,04 | 8 | 3,399 |
| 13 | 8 | 3,00 | 2,67 | 8 | 3,399 |
| 14 | 9 | 3,16 | 2,85 | 8 | 3,399 |
| 15 | 10 | 3,25 | 3,08 | 8 | 3,399 |
| 16 | 13 | 4,27 | 3,04 | 8 | 3,399 |
| 17 | 6 | 2,20 | 2,73 | 8 | 3,399 |
| 18 | 6 | 2,07 | 2,90 | 8 | 3,399 |
| 19 | 9 | 3,04 | 2,96 | 8 | 3,399 |
| 20 | 9 | 2,82 | 3,19 | 8 | 3,399 |
| 21 | 3 | 1,05 | 2,86 | 6 | 3,012 |
| 22 | 5 | 2,03 | 2,46 | 8 | 3,399 |
| 23 | 18 | 5,20 | 3,46 | 9 | 3,552 |
| 24 | 9 | 3,55 | 2,54 | 7 | 3,222 |
| 25 | 10 | 3,74 | 2,67 | 8 | 3,399 |
| 26 | 8 | 2,73 | 2,93 | 8 | 3,399 |
| 27 | 5 | 2,03 | 2,46 | 8 | 3,399 |
| 28 | 5 | 2,10 | 2,38 | 8 | 3,399 |
| 29 | 9 | 3,04 | 2,96 | 8 | 3,399 |
| 30 | 6 | 1,77 | *3,39* | 8 | 3,399 |
| 31 | 13 | 5,06 | 2,57 | 7 | 3,222 |
| 32 | 11 | 3,78 | 2,91 | 8 | 3,399 |
| 33 | 7 | 2,23 | 3,14 | 8 | 3,399 |
| 34 | 11 | 4,05 | 2,72 | 8 | 3,399 |
| 35 | 14 | 5,09 | 2,75 | 9 | 3,552 |
| 36 | 14 | 4,10 | 3,42 | 9 | 3,552 |
| 37 | 22 | 6,26 | 3,51 | 10 | 3,685 |
| 38 | 9 | 3,23 | 2,79 | 8 | 3,399 |
| 39 | 15 | 5,18 | 2,90 | 8 | 3,399 |
| 40 | 10 | 3,30 | 3,03 | 8 | 3,399 |

Dans l'exemple ci-dessus, les échantillons n° 4, 10 et 30 doivent être revus. En faisant une représentation graphique des mesures, on trouve sans autre la cause des forts écarts mis en évidence par le tableau (2.3.6).