REVUE DE STATISTIQUE APPLIQUÉE

B. CYFFERS
A. VESSERAU

Mesure et contrôle de la dispersion à partir de groupes de deux

Revue de statistique appliquée, tome 3, n° 4 (1955), p. 45-59
<http://www.numdam.org/item?id=RSA_1955__3_4_45_0>

MESURE ET CONTROLE DE LA DISPERSION A PARTIR DE GROUPES DE DEUX

par

B. CYFFERS

et

A. VESSERAU

Ingénieurs des Manufactures de l'Etat

Le contrôle de la dispersion est assez généralement effectué à partir de l'intervalle de variation des échantillons prélevés ; ce procédé, moins efficace que celui basé sur le calcul de l'écart-type, ayant l'avantage de réduire, de manière importante, les calculs à exécuter.

Dans le cas où l'appareil de mesure permet la comparaison qualitative immédiate des mesures d'un échantillon de deux, sans qu'il soit nécessaire d'observer individuellement celles-ci, MM. Cyffers et Vessereau envisagent une méthode supprimant tout calcul, méthode basée sur les comparaisons d'une série d'échantillons de deux.

Un exemple d'application industrielle (fabrication des ninas) est donné par les auteurs, avant une étude théorique de la méthode.

La dispersion d'un petit groupe de mesures :

$$x_1 , x_2 \ldots x_j \ldots x_K$$

est habituellement caractérisée par l'un ou l'autre des deux indices suivants :

l'écart-type empirique :

$$s = \sqrt{\frac{\sum (x_j - \bar{x})^2}{K - 1}}$$

$\bar{x}$ désignant la moyenne arithmétique des mesures,

- **l'intervalle de variation** (ou "étendue", ou "range") :

$$w = x_M - x_m$$

x_M et x_m désignant la plus grande et la plus petite des K mesures.

La variance empirique (s^2 = carré de l'écart-type) est une estimation absolument correcte de la variance σ^2 de la population totale dont le groupe de mesures est supposé constituer un échantillon. Lorsqu'on dispose de plusieurs groupes de même taille, ou échantillons indépendants, conduisant aux estimations

$$s_1^2 \qquad s_2^2 \ldots \qquad s_i^2 \quad \ldots\ldots \quad s_n^2$$

la meilleure estimation de σ^2 est :

$$\hat{\sigma}^2 = \frac{\sum_i s_i^2}{n} = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2}{n\ (K - 1)}$$

$\bar{x}_{i.}$ désignant la moyenne arithmétique de l'échantillon (i) et l'on prend pour estimation de σ :

$$\hat{\sigma}_{(s^2)} = \sqrt{\frac{\sum_i s_i^2}{n}}$$

Lorsque la population totale des x_i est normale, la loi de distribution de l'intervalle de variation w (pour des échantillons de K) est telle que la valeur moyenne (espérance mathématique) de w est proportionnelle à σ, le coefficient de proportionnalité ne dépendant que de K :

$$E(w) = d_K \sigma$$

Si l'on a calculé l'intervalle de variation sur n échantillons de K :

$$w_1 \quad w_2 \quad \ldots \quad w_n$$

de moyenne arithmétique :

$$\bar{w} = \frac{w_1 + \ldots\ldots + w_n}{n} = \frac{\sum w_i}{n},$$

on peut prendre pour estimation de σ :

$$\hat{\sigma}_{(w)} = \frac{1}{d_K} \bar{w}$$

Cette estimation est plus facile à calculer que l'estimation $\hat{\sigma}_{(s^2)}$, **mais elle est moins efficace :** elle perd une partie de l' "information" apportée par l'ensemble des n_K mesures.

Lorsque chaque échantillon est constitué de deux mesures seulement : x_{i1} et x_{i2}, on voit immédiatement que l'écart-type empirique et l'intervalle de variation sont proportionnels l'un à l'autre.

$$w_i = \left| x_{i2} - x_{i1} \right| = \sqrt{2}\ s_i$$

D'autre part (voir Chapitre II - Théorie de la Méthode) la relation entre σ et la valeur moyenne de w devient :

$$E(w) = \frac{2}{\sqrt{\pi}}\sigma \qquad \left(d_2 = \frac{2}{\sqrt{\pi}}\right)$$

de sorte que, dans ce cas, on peut estimer σ par :

$$\hat{\sigma}_{(w_2)} = \frac{\sqrt{\pi}}{2} \bar{w} = 0.8862\,\bar{w}$$

Naturellement cette estimation n'est précise que si le w moyen est calculé à partir d'un assez grand nombre d'échantillons de 2.

Dans ce qui suit, l'indice 2 attaché au symbole w (w_2) indique qu'il s'agit de l'intervalle de variation d'un échantillon de **2 mesures** (différence arithmétique des mesures).

En plus de la simplicité de calcul, l'estimation de σ à partir de w_2 apporte une autre simplification importante lorsqu'on peut facilement déterminer, sans avoir à les mesurer, laquelle des 2 quantités x_2 et x_1 est la plus grande. **C'est le cas que nous étudierons spécialement où ces quantités sont des poids.**

En effet, plaçant les deux objets sur les deux plateaux d'une balance suffisamment sensible, on détermine immédiatement lequel est le plus lourd (nous le désignerons par L) et lequel est le plus léger (ℓ). L est ensuite placé dans une boîte dite "des lourds" et ℓ dans une boîte "des légers". Ayant ainsi classé et réparti un certain nombre n de couples d'objets, il suffit de déterminer (par une seule ou au plus deux pesées) la différence de poids D entre la boîte des lourds et celle des légers. On a l'estimation de σ par :

$$\hat{\sigma}_{(w_2)} = 0.8862 \frac{D}{n}$$

Il convient de remarquer que les erreurs qui peuvent être faites dans le classement des objets ont pour effet **d'accroître de façon systématique** la valeur du coefficient de proportionnalité entre $\hat{\sigma}$ et $\frac{D}{n}$. Il y a là une difficulté qui peut être réduite si l'on connaît, de façon approchée, l'imprécision du classement; cette question sera traitée en détail au Chapitre II.

Ce préambule est suffisant pour aborder (Chapitre I) l'exposé d'une application industrielle, la théorie plus complète de la méthode, de ses avantages et de ses inconvénients, étant renvoyée à la fin de cet article (chapitre II).

I. - APPLICATION AU CONTROLE DE LA DISPERSION DES MACHINES A CIGARILLOS

Après avoir évoqué rapidement l'intérêt que présente le contrôle des poids dans les fabrications de la Régie Française des Tabacs, nous insisterons sur le cas particulier de la confection des cigarillos Ninas : nous exposerons ensuite l'organisation pratique des Contrôles et examinerons les résultats obtenus du double point de vue "statistique" et "industriel".

Intérêt du contrôle du poids dans la fabrication des tabacs

On peut diviser les productions de la Régie Française en deux grandes catégories :

- les produits vendus au poids : scaferlati pour pipe, tabacs à priser et à mâcher - pour lesquels il convient de ne léser ni le consommateur en lui offrant des paquets trop légers, ni le Monopole en mettant en vente des produits trop lourds.
- les produits vendus à l'unité - cigares, cigarillos, cigarettes - dont le remplissage et le tirage doivent être corrects.

Le poids des produits apparaît donc comme **un facteur important de qualité** qui revêt de plus un **intérêt financier** considérable, par suite du prix élevé de la matière première traitée.

Cas de la confection des cigarillos NINAS

Nous ne nous occuperons, dans ce qui suit, que du contrôle du poids des cigarillos Ninas dans la principale Manufacture fabriquant ce produit. L'atelier de confection comprend une soixantaine de machines, produisant chacune environ 1.000 Ninas par heure de travail effectif, soit au total 60 % environ des Ninas vendus en France.

Il nous paraît indiqué de donner quelques précisions sur la confection des cigarillos. Un cigarillo est formé de deux constituants; le tabac d'intérieur, et la cape. Le tabac pour intérieur est introduit dans le distributeur de la machine et celle-ci grâce à un système mécanique assez compliqué, fournit automatiquement la dose d'intérieur pour chaque cigarillo. Le travail de l'ouvrière chargée de la conduite de la machine consiste à tendre des feuilles de tabac sur un gabarit de découpe de la cape. Celle-ci est découpée, transportée automatiquement sur un tapis de roulage et l'intérieur est alors enroulé dans la cape par un dispositif analogue à celui, bien connu, de certaines boîtes à tabac permettant de rouler les cigarettes par simple fermeture du couvercle.

Le dosage des quantités d'intérieur est volumétrique. Il est possible de régler le volume et par suite le poids moyen. D'autre part, la machine possède un dispositif de régulation automatique du poids des doses. Ce dispositif très ingénieux et certainement indispensable au bon fonctionnement de la machine, présente cependant l'inconvénient d'inciter à croire qu'il n'y a à s'occuper que du poids moyen et nullement de la dispersion puisque la machine se corrige elle-même.

Or, si l'on procède à la pesée individuelle de 1.000 Ninas consécutives, (en gros la production d'une heure), on s'aperçoit que les poids sont distribués selon une loi normale, avec une dispersion qui peut varier considérablement d'une machine à l'autre. En première analyse, on peut admettre que cette dispersion est due uniquement au tabac d'intérieur, puisque la cape n'intervient que pour 15 à 20 % du poids total du cigarillo, et que son poids est assez régulier, sa surface étant sensiblement constante.

Ainsi, pour un poids moyen de l'ordre de 1,6 g, on constate des écarts-types de l'ordre de 0,06 g pour les meilleures machines, et pouvant atteindre jusqu'à 0,12 ou 0,14 g pour les machines les plus dispersées.

D'autre part, pour une même machine, la dispersion n'est pas une caractéristique stable. Si l'on n'y prend garde, elle va s'accroissant avec le temps, par suite de l'usure ou du déréglage de certaines pièces.

Il convient, pour chaque machine, de connaître sa dispersion et de chercher sans cesse à la diminuer,

- parce que la dispersion en poids est liée directement à la régularité de compacité et de tirage des cigarillos,
- parce que la dispersion en poids s'accompagne d'une dispersion des diamètres qui peut occasionner des difficultés lors du paquetage mécanique des Ninas,
- enfin, parce qu'on ne saurait effectuer un contrôle efficace du poids moyen sans connaître la dispersion.

C'est pourquoi il est indispensable d'obtenir, pour chacune des machines, des informations sur leur dispersion qui soient suffisamment rapprochées dans le temps.

Organisation pratique du contrôle de la dispersion

Avant de mesurer la dispersion d'une machine, il convient de définir le nombre de cigarillos sur lesquels portera la mesure, et leur mode de prélèvement.

En effet, si l'on désigne par σ_n la valeur de l'écart-type mesurée sur n cigarillos consécutifs, on constate que σ_n croît avec n : ceci provient de ce que le poids moyen fluctue dans le temps. De la même manière, l'écart type mesuré sur n cigarillos non consécutifs sera d'autant plus grand que les n cigarillos auront été prélevés dans une population couvrant une plus grande période de production.

Nous avons toujours opéré sur **1000 cigarillos consécutifs** . Le nombre 1.000 a été choisi parce qu'il correspond à peu près à une période de la fluctuation du poids moyen, qui est grossièrement sinusoïdale.

Première méthode - Pesée individuelle de 1000 Ninas

Nous avons tout d'abord établi un contrôle de la dispersion basé sur la détermination de l'écart-type empirique des poids individuels. La pesée des cigarillos est effectuée sur un peson à lecture directe dont le cadran est divisé en 51 plages (-25, -24, ..., -1, 0, +1, +2, + ... +25) d'amplitude 1cg. Le zéro est réglable.

Les opérations à effectuer sont les suivantes :

- dénombrement des 1.000 cigarillos
- pesée globale pour détermination du poids moyen
- réglage du 0 du peson au poids moyen
- pesée individuelle des Ninas et enregistrement des mesures sous forme d'histogramme
- calcul de l'écart-type.

Ces opérations demandent au total environ 4 heures. Une vérificatrice ne peut donc, par cette méthode, contrôler que deux machines par jour ouvrable.

Par conséquent, en affectant une vérificatrice en permanence à ce contrôle, on ne dispose d'une mesure par machine qu'environ toutes les 6 semaines.

Après quelques mois de fonctionnement de ce contrôle, on s'est aperçu :

- que la distribution des poids était toujours sensiblement normale,
- que l'écart-type variait d'une machine à l'autre,
- que l'écart-type, pour une même machine, avait tendance à croître avec le temps.

Ainsi est apparue la nécessité d'établir un nouveau contrôle donnant une information plus rapide.

Deuxième méthode basée sur l'intervalle de variation w_2

a) Mode opératoire :

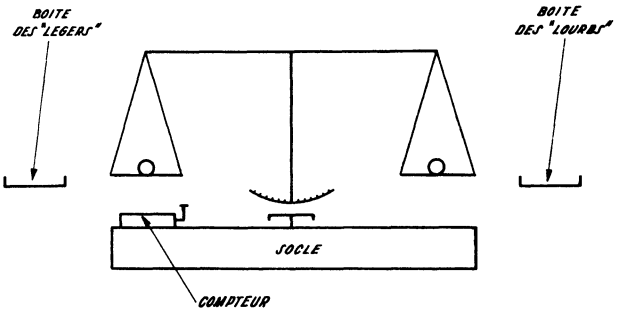
La vérificatrice dispose d'un trébuchet choisi aussi sensible que possible, et sur le socle duquel a été installé un petit compteur. Elle place devant elle un échantillon d'environ 1.000 cigarillos.

Elle prélève simultanément deux unités, une dans chaque main, et les dépose sur les plateaux du trébuchet. Tandis que de la main droite elle appuie sur un bouton pour libérer le fléau, elle appuie de la main gauche sur la touche du compteur. Elle repère ainsi le cigarillo le plus lourd et le cigarillo le plus léger, et enregistre automatiquement le nombre de couples comparés. Le cigarillo le plus lourd est ensuite transporté de la main droite dans une "boîte des lourds" et le plus léger, de la main gauche, dans une "boîte des légers".

Si, par exemple, on a disposé la "boîte des lourds" à droite du trébuchet et l'autre boîte à gauche, les gestes s'effectuent selon la règle suivante et deviennent rapidement un réflexe :

- Si l'aiguille se dirige vers la gauche : transporter les Ninas sans croiser les mains.

- Si l'aiguille se dirige vers la droite : enlever les Ninas des plateaux en croisant les mains et transporter dans les boîtes en décroisant.



Les deux boîtes ont été préalablement tarées et ont le même poids.

On poursuit l'expérience jusqu'à ce que le compteur indique 500. On pose ensuite les deux boîtes sur les plateaux d'une balance de comparaison à lecture directe, et on obtient ainsi, **par simple lecture**, la différence de poids totale entre les 500 "lourds" et les 500 "légers" de chaque couple. On peut en déduire facilement l'intervalle de variation moyen (en multipliant par 2/1.000) et également, si on le désire, l'écart-type en multipliant par le facteur de proportionnalité 0.8862. Pratiquement, nous avons retenu comme indice de dispersion la différence totale de poids, lue directement, qui est proportionnelle à l'écart-type.

b) Avantages.

Les opérations sont grandement simplifiées par rapport à la 1ère **Méthode** :

- suppression du dénombrement des 1.000 Ninas,
- suppression de leur pesée globale,
- suppression du réglage du peson. Il faut bien entendu s'assurer que le trébuchet indique 0 lorsque les plateaux sont vides, mais il n'y a plus lieu de procéder à un réglage systématique pour chaque échantillon,
- remplacement des 1.000 **pesées** individuelles, qui exigent d'attendre l'équilibre du peson, et l'enregistrement des résultats par 500 **comparaisons**, qui suppriment ces deux exigences.
- Enfin, suppression de tout calcul.

Le temps nécessaire à une mesure est à peu près 1 heure, et une même vérificatrice peut contrôler environ 8 machines par jour.

La rapidité d'exécution compense donc largement la perte d'information résultant de l'utilisation de cette méthode, puisque (voir Chapitre II) cette perte d'information n'est qu'à peine supérieure à 50 %.

Cependant la méthode présente l'inconvénient mentionné au début de cet article, à savoir qu'il n'existe qu'un seul risque d'erreur, celui de classer un "lourd" en "léger" et vice-versa, soit par manque de sensibilité de la balance, soit par erreur de manipulation; la conséquence en est une sous-estimation systématique de l'intervalle de variation ou de l'écart-type.

c) Précision :

Malgré l'instauration du contrôle par la seconde méthode, nous avons poursuivi le contrôle basé sur la 1ère Méthode et sur chaque échantillon de 1.000 cigarillos soumis à la pesée individuelle fut effectuée ensuite une mesure de l'intervalle de variation.

Nous disposons ainsi de plusieurs mesures de $\bar{w}_2$ sur des échantillons dont l'écart-type σ est connu.

Trois vérificatrices effectuant, à tour de rôle, le contrôle par la seconde méthode, nous avons enregistré sous forme de graphiques, les résultats relatifs à chacune d'elles en portant en abscisses les valeurs observées de σ, et en ordonnées, celles de $\bar{w}_2$ moyen (σ et $\bar{w}_2$ en milligrammes).

Sur chaque graphique, nous avons de plus tracé deux droites :

- la droite théorique : $\bar{w}_2 = \frac{1}{0.8862}\,\sigma = 1.1284\,\sigma$
- la droite des moindres carrés calculés à partir des résultats observés.

Les équations de ces droites sont les suivantes :

Vérificatrice TG : $\bar{w}_2 = 1.1284\,\sigma - 11.20$ (153 points)
" AT : $\bar{w}_2 = 1.0846\,\sigma - 1.93$ (106 ")
" MT : $\bar{w}_2 = 1.0885\,\sigma - 1.59$ (58 ")

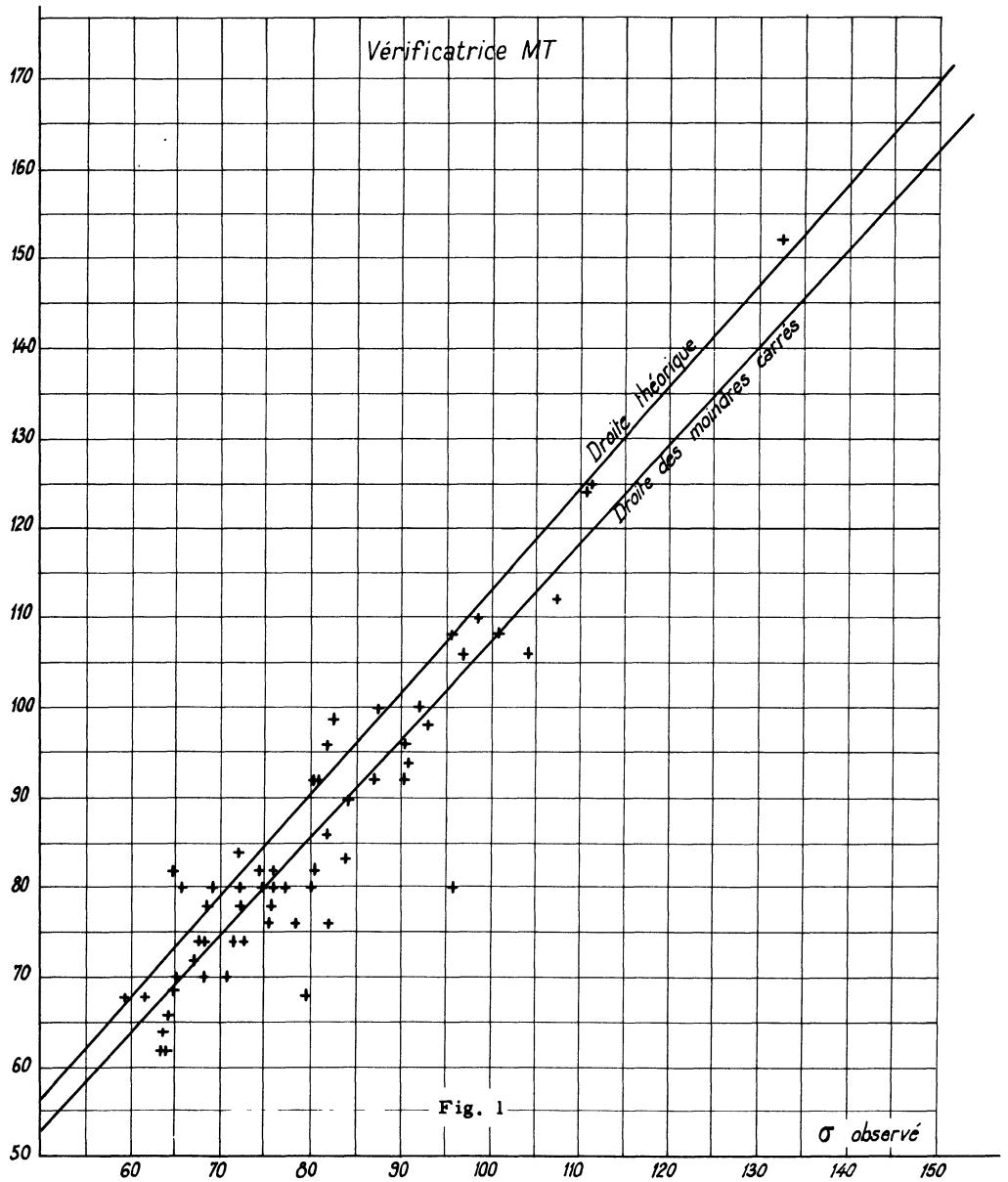


Fig. 1

Pour chaque graphique, on remarque :

- que le nuage de points est très aplati sur la droite des moindres carrés,
- que les points sont en majorité situés au-dessous de la droite théorique, ce qui correspond au risque d'erreur signalé plus haut dans la détermination de $\bar{w}_2$,
- que les deux droites ne sont pas exagérément écartées l'une de l'autre,
- enfin que le facteur personnel de la vérificatrice semble intervenir assez peu.

En définitive, l'examen des graphiques montre que la valeur observée de $\bar{w}_2$ apporte une information suffisante sur la valeur réelle de σ .

Nous donnons page 50 , à titre d'exemple, le graphique relatif à la vérificatrice MT. (fig. 1)

Résultats obtenus sur le plan industriel

Sur le plan industriel, ces contrôles ont entraîné une amélioration de la dispersion des poids des cigarillos. Nous donnons comme exemple la reproduction de la fiche de la machine N° 12 relative à une période de 8 mois environ : chaque trait horizontal a une longueur proportionnelle au w_2 observé. On constate une tendance régulière à la diminution de la dispersion dans le temps. (fig. 2)

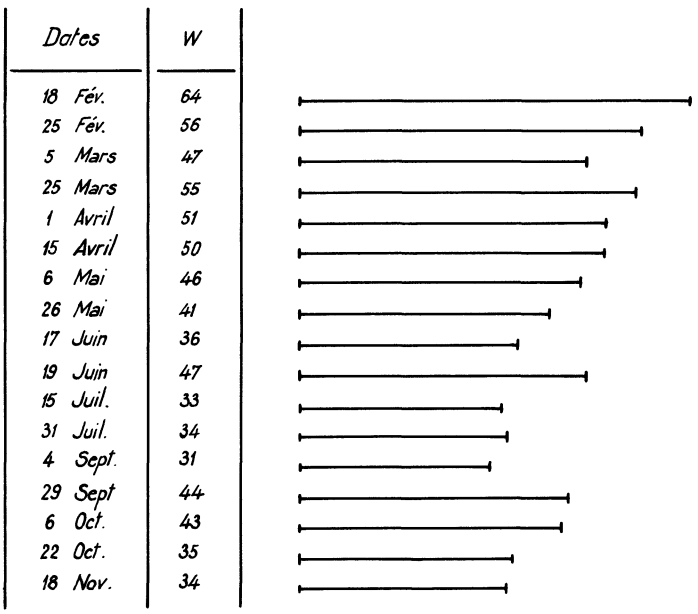


Fig. 2

Cependant, malgré l'établissement de ces fiches sur lesquelles sont enregistrés les résultats des contrôles par la 2ème Méthode, nous n'avons jamais abandonné totalement la 1ère Méthode. Le contrôle par l'intervalle de variation est un excellent moyen de **dépistage** des machines à mettre en observation ou en révision mais ses résultats sont peu éloquents pour le personnel chargé de l'entretien des machines. Celui-ci accorde en général peu de valeur à la différence moyenne de poids entre deux cigarillos prélevés au hasard, et objecte fréquemment que l'on aurait pu trouver un tout autre résultat si les couples à comparer avaient été constitués différemment.

Par contre, la présentation de l'histogramme donné par la 1ère Méthode est frappante, et il n'est jamais venu à l'idée de personne de nous dire qu'un autre échantillon de 1.000 Ninas aurait pu donner un autre résultat.

C'est pourquoi, lorsque par l'intervalle de variation, nous constatons qu'une machine a une grande dispersion, nous procédons de suite à la pesée individuelle de 1.000 autres cigarillos. Nous avons ainsi confirmation du fait, et nous possédons l'histogramme grâce auquel l'action sur le personnel d'entretien est facilitée. Une nouvelle mesure de l'écart-type, par la 1ère Méthode, est effectuée après révision, et le mécanicien est plus sensible à la présentation des deux histogrammes "avant" et "après" qu'à tous les indices chiffrés de dispersion qu'on peut lui donner.

En définitive, la méthode de contrôle de la dispersion basée sur l'intervalle de variation w_2, malgré la perte d'information qu'elle entraîne, constitue pour l'Ingénieur, grâce à sa rapidité, un moyen efficace de surveillance d'un nombre assez considérable de machines.

Cependant, la mesure directe de l'écart-type par pesée individuelle conserve des avantages, non seulement sur le plan statistique, mais également sur le plan psychologique.

II. - THÉORIE DE LA MÉTHODE

loi de w_2 . - x_1 et x_2 étant deux nombres pris au hasard (dans l'ordre 1, 2) dans une même distribution normale d'écart-type σ, la loi de $z = (x_1 - x_2)$ est une loi normale de moyenne nulle et d'écart-type $\sigma\sqrt{2}$:

$$f(z)\,dz = \frac{1}{2\sigma\sqrt{\pi}}\, e^{-\frac{z^2}{4\sigma^2}}\, dz \qquad (-\infty \leqslant z \leqslant +\infty)$$

La loi de $w_2 = |z|$ est donc :

$$f(w_2)\,dw_2 = \frac{1}{\sigma\sqrt{\pi}}\, e^{-\frac{w_2^2}{4\sigma^2}}\, dw_2 \qquad (o \leqslant w_2 \leqslant +\infty)$$

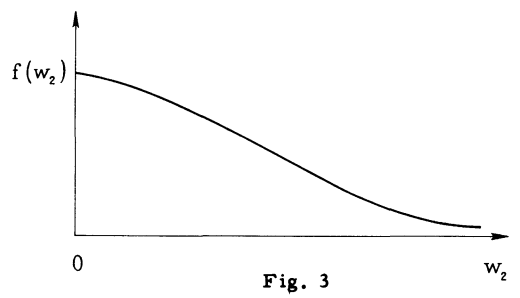


Fig. 3

Elle est représentée par la moitié droite de la courbe normale de moyenne nulle et d'écart-type $\sigma\sqrt{2}$ dont on a doublé les ordonnées. (fig.3)

L'espérance mathématique de w_2 est :

$$E(w_2) = \mu_1' = \int_o^\infty \frac{1}{\sigma\sqrt{\pi}}\, e^{-\frac{w_2^2}{4\sigma^2}}\, w_2\, dw_2 = \frac{2\sigma}{\sqrt{\pi}} \int_o^\infty e^{-\left(\frac{w_2}{2\sigma}\right)^2}\, d\left(\frac{w_2}{2\sigma}\right)^2$$

$$\mu_1' = \frac{2\sigma}{\sqrt{\pi}}$$

d'où l'on tire :

$$\sigma = \frac{\sqrt{\pi}}{2} E(w_2) = 0{,}8862\, E(w_2)$$

relation utilisée pour l'estimation de σ à partir de la moyenne arithmétique de plusieurs valeurs indépendantes de w_2 .

On calcule sans difficulté les premiers moments centrés de la loi de w_2 . On trouve :

$$\text{(Loi de } \bar{w}_2\text{)} \begin{cases} \text{moment du 2ème ordre (variance)} \quad \mu_2 = \dfrac{2\pi - 4}{\pi}\sigma^2 \\ \text{écart-type} \quad \ldots\ldots\ldots\ldots \sqrt{\mu_2} = \sqrt{\dfrac{2\pi - 4}{\pi}}\,\sigma \\ \text{moment du 3ème ordre} \quad \mu_3 = \dfrac{16 - 4\pi}{\pi}\sigma^3 \\ \text{moment du 4ème ordre} \quad \ldots\ldots\ldots\ldots \mu_4 = \left(12 - \dfrac{16}{\pi} - \dfrac{42}{\pi^2}\right)\sigma^4 \end{cases}$$

Loi de la moyenne arithmétique de n valeur w_2 indépendantes (loi de $\bar{w}_2$)

$$\bar{w}_2 = \frac{(w_2)_1 + (w_2)_2 + \ldots + (w_2)_n}{n}$$

L'expression de la loi de $\bar{w}_2$ n'est pas simple, mais on peut calculer facilement ses premiers moments.

Tout d'abord, on a évidemment :

$$E(\bar{w}_2) = E(w_2) = \mu'_1 = \frac{2\sigma}{\sqrt{\pi}}$$

Pour calculer les moments centrés, on passe par l'intermédiaire des **cumulants** de la **loi de w_2** , dont les expressions sont :

$$K_2 = \mu_2 \qquad K_3 = \mu_3$$

$$K_4 = \mu_4 - 3\mu_2^2 = \frac{32(\pi - 3)}{\pi^2}\sigma^4$$

Le cumulant d'ordre r de la loi de $\bar{w}_2$ s'obtient en divisant par n^{r-1} le cumulant d'ordre r de la loi de w_2 . Les cumulants de $\bar{w}_2$ sont donc :

$$K_2 = \frac{2\pi - 4}{n\pi}\sigma^2$$

$$K_3 = \ldots\ldots\ldots \frac{16 - 4\pi}{n^2\pi\sqrt{\pi}}\sigma^3$$

$$K_4 = \frac{32(\pi - 3)}{n^3\pi^2}\sigma^4$$

Les cumulants K_2 et K_3 se confondent avec les moments centrés d'ordre 2 et 3. On obtient le moment d'ordre 4 par la relation :

$$\mu_4 = K_4 + 3K_2^2$$

On trouve ainsi, pour moments de la loi de $\bar{w}_2$:

$$\text{(Loi de } \bar{w}_2\text{)} \begin{cases} \text{moment du 2ème ordre (variance)} \quad \mu_2 = \dfrac{2\pi - 4}{n\pi}\sigma^2 \\ \text{écart-type} \quad \ldots\ldots\ldots\ldots \sqrt{\mu_2} = \sqrt{\dfrac{2\pi - 4}{n\pi}}\,\sigma \\ \text{moment du 3ème ordre} \quad \mu_3 = \dfrac{16 - 4\pi}{n^2\pi\sqrt{\pi}}\sigma^3 \\ \text{moment du 4ème ordre} \quad \ldots\ldots\ldots\ldots \mu_4 = \dfrac{12n(\pi - 2)^2 + 32(\pi - 3)}{n^3\pi^2}\sigma^4 \end{cases}$$

Les coefficients de Pearson ont ainsi pour valeur :

$$\begin{cases} \beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{2(\pi - 4)^2}{n(\pi - 2)^3} = \dfrac{0,9859}{n} \\ \beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3 + 8\dfrac{\pi - 3}{n(\pi - 2)^2} = 3 + \dfrac{0,8692}{n} \end{cases}$$

Lorsque n augmente, ces coefficients tendent respectivement vers 0 et 3, qui **sont les valeurs de β_1 et β_2 dans une loi normale**

Pour n = 10, on a	$\beta_1 = 0,099$	$\beta_2 = 3,087$
n = 50, "	$\beta_1 = 0,020$	$\beta_2 = 3,017$
n = 100, "	$\beta_1 = 0,010$	$\beta_2 = 3,009$

On a ainsi l'exemple de la convergence vers la loi normale de la moyenne arithmétique de variables aléatoires (w_2) à distribution individuelle complètement dissymétrique. L'approximation à la loi normale est déjà très acceptable pour n = 30 ($\beta_1 = 0.033$, $\beta_2 = 3.029$).

Précision de l'écart-type estimé à partir de $\bar{w}_2$

On estime l'écart-type σ par :

$$\hat{\sigma}_{(w_2)} = \frac{\sqrt{\pi}}{2} \bar{w}_2$$

On vient de voir que la variance de $\bar{w}_2$ est :

$$\text{var } (\bar{w}_2) = \frac{2\pi - 4}{n\pi} \sigma^2$$

la variance de l'estimation $\hat{\sigma}_{(w_2)}$ est donc :

$$\text{var } \left[\hat{\sigma}_{(w_2)} \right] = \frac{\pi - 2}{2n} \sigma^2$$

n désignant le nombre de **couples d'observations,** le nombre d'observations utilisées est en fait 2 n. Lorsqu'on estime l'écart-type par la racine carrée de l'écart quadratique moyen d'un échantillon unique de (2n) mesures :

$$\hat{\sigma}_{(s^2)} = \sqrt{\frac{\sum (x - \bar{x})^2}{2n - 1}}$$

la variance de cette estimation (n étant suffisamment grand) est

$$\text{var } \left[\hat{\sigma}_{(s^2)} \right] = \frac{\sigma^2}{4n}$$

Le rapport des deux variances est :

$$I = \frac{\text{var } \left[\hat{\sigma}_{(s^2)} \right]}{\text{var } \left[\hat{\sigma}_{(w_2)} \right]} = \frac{1}{2(\pi - 2)} = \frac{44}{100}$$

La perte d'information, par l'emploi de $\bar{w}_2$, dépasse 50 %, mais elle peut se trouver compensée par une plus grande facilité, ou rapidité des opérations.

D'autre part $\hat{\sigma}_{(s^2)}$ et $\hat{\sigma}_{(w_2)}$ ne représentent la même chose que si l'on a rassemblé 2n objets, que l'on a calculé leur écart-type empirique $\hat{\sigma}_{(s^2)}$ et que d'autre part on a calculé $\hat{\sigma}_{(w_2)}$ en constituant n couples de deux, par une suite de prélèvements aléatoires et exhaustifs dans l'ensemble des 2n objets. Si $\hat{\sigma}_{(s^2)}$ a été calculé à partir de 2n objets consécutifs dans la production d'une machine, et que $\hat{\sigma}_{(w_2)}$ résulte de n couples espacés dans le temps, ces deux indices pourront différer : $\hat{\sigma}_{(w_2)}$ sera généralement plus faible que $\hat{\sigma}_{(s^2)}$ à cause des liaisons qui existent fréquemment entre les caractéristiques d'objets consécutifs.

Conséquences d'erreurs dans le classement des deux objets

Comme il a été dit au début, on considère le cas où les deux objets du couple ne sont pas mesurés, mais seulement comparés : celui dont la caractéristique est la plus grande (le plus lourd s'il s'agit de poids) est placé d'un côté, celui dont la caractéristique est la plus petite (le plus léger) d'un autre côté. Ce n'est qu'après avoir classé les n couples qu'on détermine la différence moyenne entre l'ensemble des plus grands (lourds) et l'ensemble des plus petits (légers).

Dans cette suite d'opérations, des erreurs de classement peuvent se produire : elles sont même inévitables lorsqu'on compare des poids au moyen d'une balance insuffisamment sensible : deux objets de poids très voisins auront à peu près une chance sur deux d'être incorrectement classés, le plus léger étant dirigé vers la "boîte des lourds", le plus lourd vers la "boîte des légers".

On va examiner quelles sont les conséquences de ces erreurs de classement.

Pour cela, nous admettrons que tout couple pour lequel la différence des mesures est inférieure à un nombre "a" est classé strictement au hasard; lorsque cette différence est supérieure à a, le couple est correctement classé.

Soit y la mesure **jugée la plus grande** dans le couple,
z la mesure **jugée la plus petite**

La différence $x = y - z$ varie de $-a$ à $+\infty$ suivant la loi suivante :

$-a < x < a$.- La loi de x est une portion de loi normale de moyenne nulle et d'écart-type $\sigma\sqrt{2}$.

$a < x < +\infty$.- La loi de x est une portion de la loi de w_2 .

On a donc :

$$\text{pour } -a < x < +a \qquad f_1(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\frac{x^2}{4\sigma^2}}$$

$$\text{pour } a < x < +\infty \qquad f_2(x) = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{x^2}{4\sigma^2}}$$

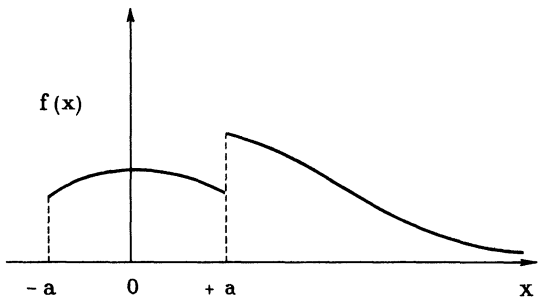


Fig. 4

1) Valeur moyenne de x

On a :

$$\mu'_1 = E(x) = \int_{-a}^{+a} x\, f_1(x)\, dx + \int_{a}^{+\infty} x\, f_2(x)\, dx = \frac{2\sigma}{\sqrt{\pi}} e^{-\frac{a^2}{4\sigma^2}}$$

d'où en posant

$$\lambda = \frac{a}{\sigma} :$$

$$E(x) = \frac{2\sigma}{\sqrt{\pi}} e^{-\frac{\lambda^2}{4}}$$

$$\sigma = \frac{\sqrt{\pi}}{2} e^{\frac{\lambda^2}{4}} E(x) \# \frac{\sqrt{\pi}}{2}\left(1 + \frac{\lambda^2}{4}\right) E(x)$$

l'approximation étant valable si le rapport λ est assez petit.

Ainsi, en prenant pour estimation de σ :

$$\hat{\sigma} = \frac{\sqrt{\pi}}{2}\bar{x}$$

on obtient une valeur systématiquement trop faible (estimation biaisée).

Si l'on connaît a et une valeur approchée de σ, on peut calculer une valeur approchée de λ et faire la correction nécessaire.

Par exemple :

si $a = 0$, la formule d'estimation correcte est : $\hat{\sigma} = 0{,}8862\ \bar{x}$

$a = \frac{\sigma}{8}$, " $\hat{\sigma} = 0{,}8897\ \bar{x}$

$a = \frac{\sigma}{4}$, " $\hat{\sigma} = 0{,}9001\ \bar{x}$

$a = \frac{\sigma}{2}$, " $\hat{\sigma} = 0{,}9434\ \bar{x}$

etc ....

2) Variance de x et de $\bar{x}$:

Le moment d'ordre 2 par rapport à l'origine est :

$$\mu'_2 = E(x^2) = \int_{-a}^{+a} x^2 f_1(x)\,dx + \int_{a}^{+\infty} x^2 f_2(x)\,dx = 2\sigma^2$$

et le moment centré (variance) :

$$\mu_2 = \mu'_2 - \mu'^2_1 = 2\sigma^2 - \frac{4\sigma^2}{\pi} e^{-\frac{\lambda^2}{2}} = 2\left(1 - \frac{2}{\pi} e^{-\frac{\lambda^2}{2}}\right)\sigma^2$$

La variance de **la moyenne $\bar{x}$** de n valeurs indépendantes x est :

$$\text{var}(\bar{x}) = \frac{2}{n}\left(1 - \frac{2}{\pi} e^{-\frac{\lambda^2}{2}}\right)\sigma^2$$

a) Si, ignorant λ , on prend pour estimation de σ :

$$\hat{\sigma} = \frac{\sqrt{\pi}}{2}\bar{x}$$

la variance de cette quantité est :

$$\frac{\pi - 2\, e^{-\frac{\lambda^2}{2}}}{2n}\sigma^2$$

L'estimation est **biaisée** et **moins précise** que lorsqu'il n'y a pas d'erreurs de classement. La nouvelle "perte d'information" par rapport à ce cas ($\lambda = o$) est :

$$1 - \frac{\pi - 2}{\pi - 2e^{-\frac{\lambda^2}{2}}} = \frac{2\left(1 - e^{-\lambda^2/2}\right)}{\pi - 2\, e^{-\lambda^2/2}}$$

soit approximativement, si $\lambda = \frac{a}{\sigma}$ est petit :

$$\frac{\lambda^2}{\lambda^2 + \pi - 2}$$

Par exemple, pour :

$a = 0$, cette perte d'information est de 0 %

$a = \frac{\sigma}{8}$, " 1.3 %

$a = \frac{\sigma}{4}$, " 5.1 %

$a = \frac{\sigma}{2}$, " 17.1 %

b) Si, connaissant une valeur approchée de λ , on prend pour estimation de σ :

$$\hat{\sigma} = \frac{\sqrt{\pi}}{2} e^{\frac{\lambda^2}{4}}\ \bar{x}$$

la variance de cette quantité est :

$$\left(\frac{\pi\, e^{\frac{\lambda^2}{2}} - 2}{2n}\right)\sigma^2$$

L'estimation est alors à **peine biaisée,** mais toujours moins précise que s'il n'y a pas d'erreurs de classement. La "perte d'information" par rapport à ce cas est :

$$1 - \frac{\pi - 2}{\pi e^{\frac{\lambda^2}{2}} - 2} = \frac{\pi\left(e^{\frac{\lambda^2}{2}} - 1\right)}{\pi e^{\frac{\lambda^2}{2}} - 2}$$

soit approximativement, si $\lambda = \frac{a}{\sigma}$ est petit :

$$\frac{\lambda^2}{\lambda^2 + 2 - \frac{4}{\pi}}$$

Par exemple,

pour a = 0 , cette perte d'information est de 0 %

$a = \frac{\sigma}{8}$, " 2,1 %

$a = \frac{\sigma}{4}$, " 8 %

$a = \frac{\sigma}{2}$, " 26,8 %

Avantages et inconvénients de la méthode

Nous les décrirons dans le cas où la caractéristique mesurée est un poids.

Les **inconvénients** sont les suivants :

- une perte d'information, par rapport à l'estimation de σ utilisant l'écart quadratique moyen des mesures, qui dépasse 50 % dans le cas le plus favorable.
- une estimation biaisée (dans un sens connu) lorsqu'il y a des erreurs de classement, ce qui oblige :
  - soit à opérer avec un appareil de pesée très sensible, (mais pas nécessairement parfaitement juste)
  - soit à effectuer une correction qui exige que l'on connaisse, au moins de façon approchée, la sensibilité de l'appareil et l'ordre de grandeur de l'écart-type recherché.

Les **avantages** sont :

- la grande rapidité des opérations : on compare deux poids sans avoir à les mesurer,
- la suppression des erreurs de lecture,
- la suppression presque totale des calculs,
- la possibilité d'imaginer un appareil effectuant le classement de façon automatique.

On a vu au paragraphe I, que les avantages peuvent l'emporter nettement sur les inconvénients, en particulier dans des contrôles industriels qui doivent être confiés à un personnel peu qualifié, et où l'on ne recherche pas une très grande précision de l'indice de dispersion.

VÉRIFICATIONS EXPÉRIMENTALES

Avant d'appliquer industriellement la méthode, nous avons procédé "in vitro" à une série de vérifications expérimentales dont les résultats vont être donnés.

Toutes les vérifications ont porté sur la dispersion en poids de lots de cigarettes prélevées à la sortie d'une machine. La distribution des poids de cigarettes obéit presque toujours à une loi normale, ou très voisine de la normale.

Le poids moyen d'une cigarette est d'environ 1 gr 2. L'écart-type, exprimé en milligrammes, varie, suivant le modèle, l'ancienneté et l'état de la machine, de 50 à 100 et même plus.

Les lots étudiés - provenant de plusieurs Manufactures et dans chacune d'elles de plusieurs machines - sont de 1000 cigarettes.

a) Estimation de σ à partir de w_2, sans erreur de classement

Les cigarettes, numérotées de 1 à 1000, ont été pesées individuellement avec une grande précision, et l'on a calculé l'écart-quadratique moyen.

On a ensuite associé les numéros deux par deux, strictement au hasard. On a calculé les 500 différences de poids w_2 des cigarettes ainsi associées, puis leur moyenne $\overline{w_2}$. On en a déduit l'estimation $\hat{\sigma}_{(w_2)}$ par la relation :

$$\hat{\sigma}_{(w_2)} = 0,8862\ \overline{w}_2$$

Les résultats obtenus figurent dans le tableau ci-après (on avait choisi volontairement des lots d'écarts-types très différents).

Référence		$\hat{\sigma}_{(s^2)}$	$\hat{\sigma}_{(w_2)}$	Différence
Manuf. X	Machine 1	54,1	54,8	+ 0,7
	" 2	55,1	57,-	+ 1,9
	" 3	58,4	57,4	- 1,-
	" 4	63,8	64,8	+ 1,-
	" 5	67,2	65,7	- 1,5
		59,7	59,9	+ 0,2
Manuf. Y	Machine 1	67,3	63,6	- 3,7
	" 2	76,3	73,-	- 3,3
	" 3	79,6	79,7	+ 0,1
	" 4	87,9	86,-	- 1,9
	" 5	115,-	112,6	- 2,4
		85,2	83,-	- 2,2
Manuf. Z	Machine 1	90,5	86,-	- 4,5
	" 2	97,8	98,1	+ 0,3
	" 3	101,8	105,6	+ 3,8
	" 4	112,7	111,1	- 1,6
	" 5	117,9	112,5	- 5,4
		104,1	102,7	- 1,4
ENSEMBLE		83,-	81,9	- 1,1

La concordance des deux estimations peut être considérée comme satisfaisante. Les plus gros écarts sont presque toujours dans le sens $\hat{\sigma}_{(w_2)} < \hat{\sigma}_{(s^2)}$, et sur l'ensemble $\hat{\sigma}_{(w_2)}$ est légèrement inférieur à $\hat{\sigma}_{(s^2)}$. Ces écarts sont dans le sens qu'expliquent de légères erreurs de classement.

b) Influence des erreurs de classement

Sur 3 lots de 1.000 cigarettes, on a procédé à la vérification suivante :

1) On détermine l'écart-type $\hat{\sigma}_{(s^2)}$ (désigné ci-dessous par σ).

2) Comme indiqué ci-dessus (numérotage des cigarettes de 1 à 1000, constitution au hasard de 500 couples) - on détermine le poids total des plus lourdes de chaque couple, et le poids total des plus légères, d'où par différence et division par 500 l'intervalle de variation moyen $\overline{w}_2$; on en déduit le rapport $\left(\frac{\sigma}{\overline{w}_2}\right)$ observé (la valeur théorique est 0,8862).

3) On repère tous les couples dont la différence w_2 est inférieure ou égale à a = 5 mg. Pour chacun de ceux-ci, on dirige, d'après le résultat d'une épreuve de "pile ou face", la cigarette la plus lourde, soit vers le groupe des cigarettes "les plus lourdes" soit vers le groupe des cigarettes "les plus légères" et inver-

sement. Ainsi, le poids total des cigarettes considérées "comme les plus lourdes dans chaque couple" est normalement diminué; le poids total des cigarettes "considérées comme les plus légères" est normalement augmenté. On calcule ensuite la nouvelle valeur de $(\frac{\sigma}{\bar{w}_2})$ qui est normalement plus élevée que la précédente - et on la compare à la valeur théorique dont on a établi précédemment qu'elle est :

$$0,8862\ e^{\lambda^2/4} \quad , \quad \text{avec } \lambda = \frac{5}{\sigma}.$$

4) On recommence la même opération avec des "erreurs de classement" caractérisées par des valeurs de a :

a = 10 a = 15 a = 20 a = 25

et l'on compare chaque fois le coefficient observé $(\frac{\sigma}{\bar{w}_2})$ - qui normalement augmente de plus en plus - et le coefficient théorique calculé avec la valeur convenable de λ.

Le résultat de ces opérations figure dans les trois tableaux ci-après. On y constate qu'il y a une concordance satisfaisante entre coefficients observés et coefficients théoriques.

1er LOT
1.000 cigarettes - 500 couples - σ = 70,21

Valeur de a	0	5	10	15	20	25
Valeur de $\lambda = \frac{a}{\sigma}$	0	0.071	0.142	0.214	0.285	0.356
% de couples séparés au hasard	0%	3.6%	8.2%	11.4%	15.-%	19.8%
$\bar{w}_2$ observé	78.458	78.322	77.970	77.694	76.978	76.014
$\sigma / \bar{w}_2$ observé	0.8949	0.8964	0.9005	0.9037	0.9121	0.9236
Coefficient théorique	0.8862	0.8873	0.8907	0.8964	0.9044	0.9147

2ème LOT
1.000 cigarettes - 500 couples - σ = 80.23

Valeur de a	0	5	10	15	20	25
Valeur de $\lambda = \frac{a}{\sigma}$	0	0.062	0.125	0.187	0.249	0.312
% de couples séparés au hasard	0%	3.2%	7%	10%	13.2%	16.6%
$\bar{w}_2$ observé	90.444	90.404	89.964	89.660	88.852	88.008
$\sigma / \bar{w}_2$ observé	0.8871	0.8875	0.8918	0.8948	0.9029	0.9108
Coefficient théorique	0.8862	0.8871	0.8897	0.8940	0.9000	0.9080

3ème LOT
1.000 cigarettes - 500 couples - σ = 87.13

Valeur de a	0	5	10	15	20	25
Valeur de $\lambda = \frac{a}{\sigma}$	0	0.057	0.115	0.172	0.230	0.287
% de couples séparés au hasard	0%	2.6%	6.6%	9.2%	10.4%	15%
$\bar{w}_2$ observé	96.000	95.928	95.628	94.948	94.024	94.056
$\sigma / \bar{w}_2$ observé	0.9076	0.9083	0.9111	0.9177	0.9267	0.9264
Coefficient théorique	0.8862	0.8869	0.8891	0.8928	0.8980	0.9046