Zeitschrift: Bulletin de l'Association suisse des électriciens

Herausgeber: Association suisse des électriciens

Band: 45 (1954)

Heft: 16

Rubrik: Production et distribution d’énergie : les pages de l’UCS

Production et distribution d'énergie

Les pages de l'UCS

Adaptation des méthodes statistiques modernes aux besoins des électriciens

Par *Ch. Morel, Zurich*

519.24 : 658.8.03 : 621.311

Les progrès réalisés au cours des dernières années dans le domaine de la statistique mathématique justifient l'application toujours plus étendue des méthodes d'investigation que cette science encore jeune met à disposition, à l'étude des problèmes les plus divers de l'exploitation des entreprises de production et de distribution d'énergie électrique. Dans le cadre limité qui lui est imposé, le présent exposé essaie de dégager les principes fondamentaux des méthodes statistiques modernes et de montrer, par quelques exemples simples, leur application pratique. Il insiste en particulier sur les dangers d'une interprétation par trop étroite des résultats qui sont toujours affectés d'une erreur possible, chiffrable il est vrai, mais dont il faut tenir compte. Le problème de la corrélation y est traité en détail, ce qui permet de montrer également comment fonctionne l'analyse de la variance. Un bref aperçu de l'application des distributions statistiques à la solution de problèmes que soulève l'exploitation termine cet exposé.

Die in der mathematischen Statistik in den letzten Jahren erzielten Fortschritte rechtfertigen eine immer weitere Anwendung der von dieser noch jungen Wissenschaft zur Verfügung gestellten Mittel, zur Lösung der mannigfaltigen Probleme des Elektrizitätswerkbetriebes. In beschränktem Rahmen dieses Aufsatzes wird versucht, die Grundprinzipien der modernen statistischen Methoden herauszuschälen, und ihre praktische Anwendung anhand einiger einfacher Beispiele zu zeigen. Es wird insbesondere auf die Gefahren einer allzu engen Auslegung der Ergebnisse hingewiesen, die immer mit einem möglichen Fehler behaftet sind, der aber zahlenmässig feststellbar ist und so berücksichtigt werden kann. Die Korrelationsrechnung wird eingehend behandelt, was gleichzeitig erlaubt, zu zeigen, wie die Streuungszerlegung vor sich geht. Schliesslich wird ein kurzer Überblick über die Anwendung der statistischen Verteilungen zur Lösung von Problemen des Betriebes gegeben.

Avant-propos

La statistique joue dans l'exploitation un rôle incontestable, rôle double qui est de *récapituler* ce qui s'est passé pour en tirer des *conclusions* (par exemple amélioration de l'exploitation) ou des *prévisions* pour l'avenir (par exemple nouvel équipement). Elle n'est pas un but en elle-même, mais un instrument dont l'emploi n'est pas exempt de risques. La conception actuelle de la statistique et les méthodes modernes d'analyse et de contrôle permettent d'éviter ces risques et, ce qui importe avant tout, de chiffrer le degré de certitude des résultats acquis.

La première tâche du statisticien à consiste *recueillir* les données et, si possible, à les exprimer par des chiffres. Ce travail peut paraître ennuyeux, mais il n'est pas sans importance, car c'est de lui que dépend en grande partie le résultat final des investigations. Il faut ensuite *présenter* les données recueillies, les faire parler, leur donner de la vie. On formera des sommes et des moyennes, on étudiera la variabilité des séries de données et les relations entre ces séries. C'est seulement alors qu'on pourra *interpréter* les résultats en se servant des différentes méthodes de contrôle (tests) créées au cours de ces dernières décades. Ce dernier travail en particulier exige une collaboration étroite entre le praticien ou le spécialiste et le statisticien.

1. Comment recueillir des données

Il semble utile de rappeler au début que les séries de chiffres recueillies par le statisticien ne sont pas des masses anonymes, mais des groupements de valeurs individuelles dont chacune a sa vie et ses caractéristiques propres, dont il faut tenir compte au cours des calculs. Les valeurs individuelles (ou individus) peuvent paraître hétérogènes à première vue. Elles relèvent toutefois d'un univers dont elles forment un échantillon. Or, les lois qui régissent l'univers se reproduisent dans chaque échantillon entre certaines limites données par la probabilité. On dit que les caractéristiques d'un échantillon sont des estimations de celles de l'univers dont il est issu, estimations dont l'exactitude peut être chiffrée.

Cette conception a une valeur immédiatement pratique. Il n'est plus nécessaire de rassembler indéfiniment des chiffres pour obtenir des renseignements précis. Il suffit d'opérer sur un échantillon dont l'extension dépendra, outre des caractéristiques statistiques du groupe considéré, de la précision que l'on se propose d'atteindre. Nous ne nous arrêterons pas ici au problème de l'échantillonnage, car celui-ci déborde du cadre de l'étude que nous avons en vue.

La plupart des données susceptibles de faire l'objet d'études statistiques sont fournies par l'exploitation — prise au sens le plus large du mot — c'est-à-dire par les services techniques, commerciaux et comptables, et ceci sous forme de chiffres. Le statisticien peut cependant être appelé à résoudre des problèmes où il ne peut pas puiser dans le réservoir de l'exploitation. Pour obtenir les données nécessaires, il doit alors procéder par enquêtes. Ce chapitre à lui seul pourrait remplir des pages. Nous nous bornerons ici à quelques considérations générales. Les questionnaires seront aussi simples que possible; ils éviteront les questions suggestives ou ambiguës. Au besoin, on soumettra préalablement le questionnaire à une dizaine ou à une vingtaine de personnes, ce qui permettra de préciser les questions peu claires. Pour la disposition des questionnaires, on tiendra compte du mode de dépouillement prévu, manuel ou mécanique, et prévoira, pour autant que faire se peut, des possibilités de recoupement et de contrôle.

La question la plus délicate qui se pose est celle de l'extension à donner à l'échantillon, car il est clair que l'on ne pourra pour ainsi dire jamais étendre ses investigations à l'ensemble des individus considérés. Ceci est d'ailleurs valable non seulement pour les enquêtes, mais pour toutes les études statistiques. Nous renvoyons à ce sujet à la documentation figurant à la fin de la présente étude.

Un autre problème que nous tenons à mentionner ici est celui du contrôle et du réceptionnement du matériel. Citons, par exemple, le contrôle de la longévité des lampes à incandescence, le contrôle des isolateurs, le contrôle de la conductibilité du cuivre, les essais d'huiles isolantes, le contrôle de la qualité du charbon, etc. La première question qui se pose est la suivante: le matériel à contrôler étant généralement détruit à l'essai, quel pourcentage du lot à contrôler faut-il prélever pour obtenir, avec une précision donnée, un maximum d'information. Nous ne pouvons nous y arrêter dans le cadre de cette étude et renvoyons également à la littérature scientifique.

2. La présentation des données

Les données recueillies se présentent généralement sous la forme d'une masse amorphe qu'il s'agit d'ordonner. On commencera par étudier leur distribution. Cette opération descriptive permettra, la plupart du temps, de reconnaître quelle loi de distribution régit l'ensemble de chiffres considéré. La seconde opération, la réduction des données, consiste à calculer les caractéristiques ou paramètres de la distribution. Si l'on a plusieurs séries de données dont on suppose qu'elles sont liées par certaines relations, on sera amené à rechercher les dépendances statistiques.

Description des données

Pour pouvoir se rendre compte de la distribution, il faut tout d'abord ordonner les données; on les groupera d'abord par classes puis, à l'intérieur des classes, par ordre croissant ou décroissant. Le critère à choisir pour le classement sera le caractère de la variable à étudier. L'intervalle de variation (écart entre la plus faible et la plus forte valeur individuelle) sera divisé en un nombre de classes d'égale largeur tel que chaque classe comporte un nombre suffisant d'individus, mais que le caractère de la distribution ressorte quand même avec une précision suffisante.

Pour la représentation graphique, on se servira de préférence de la courbe en escalier (histogramme), pour bien marquer qu'il ne s'agit pas d'une fonction continue, mais de valeurs discrètes.

La courbe de distribution (fréquence par classe) sera complétée par la courbe de répartition ou courbe des fréquences cumulées (somme des fréquences de toutes les classes inférieures à la classe considérée y compris celle-ci).

Suivant le cas, on choisira les fréquences absolues ou les fréquences relatives (quotient de la fréquence considérée par la somme de toutes les fréquences ou extension de l'échantillon indiqué en pourcents).

Distributions statistiques

La distribution la plus connue est la distribution normale ou distribution de Gauss-Laplace, appelée aussi loi des erreurs ou courbe en cloche. C'est une fonction symétrique dont l'expression mathématique est

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Les deux paramètres de cette fonction sont μ (moyenne) et σ^2 (variance ou dispersion). Par un changement de variable, on peut la ramener à une forme plus simple, la forme standard, de moyenne zéro et de variance égale à l'unité.

On pose

$$u = \frac{x-\mu}{\sigma}$$

et obtient

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

L'expression $\varphi(x)\,dx$ est la probabilité pour qu'un écart soit compris entre x et $x + dx$. La somme des probabilités élémentaires $\int_{-\infty}^{x} \varphi(x)\,dx$ exprime la probabilité pour que l'écart soit inférieur ou égal à x. Ce chiffre est toujours inférieur à 1, car on démontre facilement que $\int_{-\infty}^{+\infty} \varphi(x)\,dx = 1$.

Une autre distribution, dont la distribution normale n'est qu'un cas-limite, est la distribution binomiale, dont l'expression mathématique est

$$\varphi(x) = \binom{m}{x} p^x (1-p)^{m-x}$$

où

$$\binom{m}{x} = \frac{m!}{x!\,(m-x)!}$$

C'est la loi qui indique la probabilité pour qu'un évènement favorable, de probabilité élémentaire $p < 1$, se produise x fois dans une série de m essais (par exemple probabilité d'extraire x boules rouges en une série de m tirages, lorsque l'urne contient 100 boules rouges sur 1000 au total; $p = 1/10$).

La moyenne ou espérance mathématique est:

$$\mu = m\,p$$

et la variance:

$$\sigma^2 = m\,p\,(1-p)$$

Lorsque m tend vers l'infini et que p reste fini, la distribution binomiale passe à la distribution normale.

Par contre, lorsque m tend vers l'infini et p vers zéro, le produit $mp = \lambda$ restant fini, on obtient une autre forme de distribution: la distribution de Poisson ou distribution des évènements rares, dont l'expression mathématique est

$$\varphi(x) = \frac{\lambda^x\, e^{-\lambda}}{x!}$$

Pour cette fonction on a

$$\mu = \lambda \quad \text{et} \quad \sigma^2 = \lambda$$

Réduction des données

La méthode statistique permet de remplacer un ensemble de données numériques par un certain nombre de paramètres ou caractéristiques résumant le dit ensemble. Cette opération permet soit de comparer entre eux plusieurs ensembles par la mise en regard de leurs paramètres, soit de réduire les données afin d'en donner une description simplifiée.

On commencera par déterminer le paramètre de position qui est généralement la moyenne arithmétique $\bar{x}$. C'est celui qui répond le mieux aux critères que l'on pose à l'égard des caractéristiques. Elle est consistante, correcte et efficace. D'autres paramètres de position moins efficaces, mais qui peuvent fournir certains renseignements lorsque la distribution s'écarte par trop de la normale sont la médiane et le mode.

Si l'échantillon compte N valeurs individuelles x_i, réparties en M classes de centre x_j et de fréquence f_j la moyenne s'écrit

$$\bar{x} = \frac{1}{N} \mathop{\mathrm{S}}_{i=1}^{N} x_i = \frac{1}{N} \mathop{\mathrm{S}}_{j=1}^{M} f_j\, x_j, \text{ où } N = \mathop{\mathrm{S}}_{j=1}^{M} f_j$$

La mesure la plus efficace de la variabilité est la variance s^2 (ou l'écart-type s) qui a pour expression

$$s^2 = \frac{1}{N-1} \mathop{\mathrm{S}}_{i=1}^{N} (x_i - \bar{x})^2 = \frac{1}{N-1} \mathop{\mathrm{S}}_{j=1}^{M} f_j\, (x_j - \bar{x})^2$$

Pour le calcul, on se servira des relations

$$S\,(x_i - \bar{x})^2 = S\,x_i^2 - \frac{1}{N} (S\,x_i)^2 =$$

$$= S\,x_i^2 - \bar{x}\, S\,x_i = S\,x_i^2 - N\,\bar{x}^2$$

et

$$S\,f_j\,(x_j - \bar{x})^2 = S\,f_j\,x_j^2 - \frac{1}{N} (S\,f_j\,x_j)^2 =$$

$$= S\,f_j\,x_j^2 - \bar{x}\, S\,f_j\,x_j = S\,f_j\,x_j^2 - N\,\bar{x}^2$$

Une autre mesure de la variabilité est l'intervalle de variation $w = x_N - x_1$ qui est d'un emploi courant dans le contrôle de la fabrication, par exemple.

Ces calculs ont un sens précis lorsque la distribution est normale, ou presque. Très souvent, on pourra ramener une distribution dissymétrique à la normale par un changement de variable. Bien des cas se laissent réduire de la sorte en posant $z = \log\,(a + b\,x)$ ou plus simplement $z = \log x$. Dans d'autres cas, il faudra poser $z = x^2$ ou $z = \frac{1}{x}$.

Le moyen le plus simple de reconnaître une distribution normale consiste à dresser la courbe des fréquences cumulées sur du papier à échelle dite de probabilité. Si la courbe ainsi obtenue est une droite ou presque, on pourra admettre que la distribution est normale ou presque.

Dépendance statistique

Très souvent, l'étude statistique ne porte pas sur une seule variable, mais en fait intervenir plusieurs dont il s'agit de faire ressortir les relations mutuelles.

Il est peut-être indiqué de rappeler ici que les relations entre séries statistiques ne sont généralement pas des relations fonctionnelles. A chaque valeur d'une des variables peuvent correspondre plusieurs valeurs différentes de l'autre variable. La relation, plus ou moins lâche, est dite stochastique. Elle permet de conclure, en partant d'une variable, à la valeur probable de l'autre, ou plutôt aux limites entre lesquelles on peut s'attendre avec une probabilité donnée à voir varier l'autre. La fonction qui lie des deux variables est dite de régression. Elle peut souvent être considérée, en première approximation, comme une droite (régression linéaire), mais il existe aussi des cas où la régression est d'un ordre supérieur. L'intensité du groupement des valeurs individuelles autour de la fonction de régression se mesure par le coefficient de détermination, indépendant de la forme de la relation.

La régression linéaire peut être simple

$$Y = a + b\,(x - \bar{x})$$

ou multiple

$$Y = a + b_1\,(x_1 - \bar{x}_1) + b_2\,(x_2 - \bar{x}_2) + \ldots$$

où x_1, x_2 ... sont les variables indépendantes.

Dans ces équations, la constante a est égale à la moyenne $\bar{y}$ et le coefficient de régression b se détermine par la méthode des moindres carrés. On obtient ainsi, pour la régression linéaire simple en x,

$$b = \frac{S\,(x - \bar{x})\,(y - \bar{y})}{S\,(x - \bar{x})^2}$$

et pour la régression linéaire double,

$$b_1 = \frac{S\,(x_1 - \bar{x}_1)\,(x_2 - \bar{x}_2)\, S\,(x_2 - \bar{x}_2)\,(y - \bar{y}) - S\,(x_2 - \bar{x}_2)^2\, S\,(x_1 - \bar{x}_1)\,(y - \bar{y})}{[S\,(x_1 - \bar{x}_1)\,(x_2 - \bar{x}_2)]^2 - S\,(x_1 - \bar{x}_1)^2\, S\,(x_2 - \bar{x}_2)^2}$$

et

$$b_2 = \frac{S\,(x_1 - \bar{x}_1)\,(x_2 - \bar{x}_2)\, S\,(x_1 - \bar{x}_1)\,(y - \bar{y}) - S\,(x_1 - \bar{x}_1)^2\, S\,(x_2 - \bar{x}_2)\,(y - \bar{y})}{[S\,(x_1 - \bar{x}_1)\,(x_2 - \bar{x}_2)]^2 - S\,(x_1 - \bar{x}_1)^2\, S\,(x_2 - \bar{x}_2)^2}$$

Nous renonçons à donner ici les formules encore plus compliquées pour les coefficients b lorsqu'il s'agit de régression multiple à plus de deux variables indépendantes.

Le coefficient de détermination B, qui n'est autre que le rapport de la variabilité de la fonction de régression à la variabilité totale de la variable dépendante

$$B = \frac{S\,(Y - \bar{y})^2}{S\,(y - \bar{y})^2}$$

s'exprime de la sorte:

— pour la régression simple

$$B = \frac{[S\,(x - \bar{x})\,(y - \bar{y})]^2}{S\,(x - \bar{x})^2\,S\,(y - \bar{y})^2} = b\,\frac{S\,(x - \bar{x})\,(y - \bar{y})}{S\,(y - \bar{y})^2}$$

— et pour la régression double

$$B = \frac{b_1\,S\,(x_1 - \bar{x}_1)\,(y - \bar{y}) + b_2\,S\,(x_2 - \bar{x}_2)\,(y - \bar{y})}{S\,(y - \bar{y})^2}$$

Pour la régression linéaire simple, le coefficient de corrélation r est égal à la racine carrée du coefficient de détermination B

$$r^2 = B \quad \text{ou} \quad r = \sqrt{B}$$

La régression non linéaire, par exemple

$$Y = a + b\,x + c\,x^2 + \ldots$$

se traite de la façon suivante.

On pose

$$Y = \bar{y} + b_1\,(x_1 - \bar{x}_1) + b_2\,(x_2 - \bar{x}_2) + \ldots$$

où

$$x_1 = x, \quad x_2 = x^2, \text{ etc.}$$

La solution est alors identique à celle de la régression multiple.

Pour déterminer la forme de la courbe, on se sert avantageusement de la méthode des polynomes orthogonaux établie par Fisher. Cependant, on pourra se passer de ces opérations, si l'on peut reconnaître à première vue le degré de l'équation de régression.

3. L'interprétation des résultats

L'interprétation permet, en premier chef, de formuler des jugements inductifs sur l'univers en cause à partir de l'information apportée par les échantillons qui en sont extraits, et ceci avec l'aide des tests statistiques. Elle permet également de rapporter la variabilité des phénomènes aux causes qui les ont engendrés; ceci est le domaine de l'analyse de la variance.

Les tests statistiques

La méthode appliquée ici consiste à formuler une hypothèse sur un ou plusieurs univers et à vérifier par un test approprié dans quelle mesure cette hypothèse est infirmée par l'examen d'un ou de plusieurs échantillons tirés de cet (ou de ces) univers. Si cette mesure est forte, c'est-à-dire si l'écart est significatif, l'hypothèse doit être rejetée. Si la mesure est faible, ou l'écart fortuit, on peut seulement conclure à l'absence de raisons suffisantes pour la rejeter. On tâchera donc de formuler l'hypothèse de façon à ce que son rejet et non son acceptation donne la certitude que l'on recherche.

Les seuils de signification

Avant de passer en revue les principales lois qui régissent les relations entre les caractéristiques d'un univers et celles des échantillons qui en dérivent, il nous semble utile de décrire succinctement la méthode généralement utilisée pour les comparaisons. La plupart des grandeurs statistiques ou de leurs écarts sont distribués selon des lois connues (distribution normale ou autre), de sorte qu'à chaque grandeur ou écart constaté correspond une probabilité bien déterminée. L'expérience a conduit à fixer une limite appelée seuil de signification et définie comme la valeur de l'écart correspondant à la probabilité 1 % que cet écart soit atteint ou dépassé. On admet généralement deux seuils de signification correspondant respectivement aux probabilités 5 % et 1 % (ou 0,05 et 0,01). Un écart dont la probabilité est supérieure à 5 % est dit fortuit, un écart supérieur, c'est-à-dire de probabilité inférieure à 5 %, significatif. — Lorsque la probabilité d'un écart est inférieure à 1 %, on dit que cet écart est fortement significatif.

Dans une distribution normale, par exemple, le seuil 5 % ($u_{0,05}$) correspond à 1,96 σ et le seuil 1 % ($u_{0,01}$) à 2,576 σ. On peut donc dire, en première approximation, que tout écart supérieur au double de l'écart-type ($u > 2\,\sigma$) est significatif. On verra plus loin l'application pratique de ces seuils de signification.

Application des tests statistiques

Les situations nécessitant l'application des tests statistiques concernent le plus souvent des com-

Tableau I

Caractéristique observée (échantillon)	Valeur exacte (univers)	Variance	Distribution de la caractéristique observée	
			Petit échantillon	Grand échantillon
Moyenne $\bar{x} = \frac{1}{N}\,S\,x$	μ	$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$	normale	normale
Variance $s^2 = \frac{1}{N-1}\,S\,(x - \bar{x})^2$	σ^2	$\sigma_{s^2}^2 \cong \frac{2\,\sigma^4}{N}$	dissymétrique	normale
Ecart-type $s = \sqrt{\frac{1}{N-1}\,S\,(x - \bar{x})^2}$	σ	$\sigma_s^2 = \frac{\sigma^2}{2\,N}$	dissymétrique	normale
Coefficient de régression linéaire $b = \frac{S\,(x - \bar{x})\,(y - \bar{y})}{S\,(x - \bar{x})^2}$	β	$\sigma_b^2 = \frac{S\,(y - Y)^2}{N - 2}$	dissymétrique	normale
coefficient de corrélation $r = \frac{S\,(x - \bar{x})\,(y - \bar{y})}{\sqrt{S\,(x - \bar{x})^2 \cdot S\,(y - \bar{y})^2}}$	ϱ	$\sigma_r^2 = \frac{(1 - \varrho^2)^2}{N - 1}$	dissymétrique	à peu près normale

Tableau II

Grand échantillon (N > 100)		Petit échantillon (N < 100)	
Écart entre valeur observée et valeur exacte	Écart entre deux valeurs observées	Écart entre valeur observée et valeur exacte	Écart entre deux valeurs observées
1. Moyennes			
$u = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ $\sigma^2_{\bar{x}} = \dfrac{s^2}{N}$	$u = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma_d}$ $\sigma_d^2 = \dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}$	$t = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ $\sigma^2_{\bar{x}} = \dfrac{s^2}{N}$ $n = N - 1$	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma_d}$ $\sigma_d^2 = \dfrac{S(x_1 - \bar{x}_1)^2 + S(x_2 - \bar{x}_2)^2}{N_1 + N_2 - 2} \left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)$ $n = N_1 + N_2 - 2$
2. Ecarts-types			
$u = \dfrac{s - \sigma}{\sigma_s}$ $\sigma_s^2 = \dfrac{s^2}{2N}$	$u = \dfrac{s_1 - s_2}{\sigma_d}$ $\sigma_d^2 = \dfrac{s_1^2}{2N_1} + \dfrac{s_2^2}{2N_2}$	$F = \dfrac{s^2}{\sigma^2}$ $n_1 = N - 1 \quad n_2 = \infty$	$F = \dfrac{s_1^2}{s_2^2}$ $n_1 = N_1 - 1 \quad n_2 = N_2 - 1$
3. Coefficients de régression			
$u = \dfrac{b - \beta}{\sigma_b} \sqrt{S(x - \bar{x})^2}$ $\sigma_b^2 = \dfrac{S(y - Y)^2}{N - 2}$	$u = \dfrac{b_1 - b_2}{\sigma_d}$ $\sigma_d^2 = \dfrac{S(y_1 - Y_1)^2 + S(y_2 - Y_2)^2}{N_1 + N_2 - 4} \cdot$ $\left(\dfrac{1}{S(x_1 - \bar{x}_1)^2} + \dfrac{1}{S(x_2 - \bar{x}_2)^2}\right)$	$t = \dfrac{b - \beta}{\sigma_b} \sqrt{S(x - \bar{x})^2}$ $\sigma_b^2 = \dfrac{S(y - Y)^2}{N - 2}$ $n = N - 2$	$t = \dfrac{b_1 - b_2}{\sigma_d}$ $\sigma_d^2 = \dfrac{S(y_1 - Y_1)^2 + S(y_2 - Y_2)^2}{N_1 + N_2 - 4} \cdot$ $\left(\dfrac{1}{S(x_1 - \bar{x}_1)^2} + \dfrac{1}{S(x_2 - \bar{x}_1)^2}\right)$ $n = N_1 + N_2 - 4$
4. Coefficients de corrélation			
$u = \dfrac{r - \varrho}{\sigma_\varrho}$ $\sigma_\varrho^2 = \dfrac{(1 - \varrho^2)^2}{N - 1}$	$u = \dfrac{r_1 - r_2}{\sigma_d}$ $\sigma_d^2 = s^2_{r1} + s^2_{r2}$ $s_r^2 = \dfrac{(1 - r^2)^2}{N - 1}$	$t = \dfrac{r}{\sqrt{1 - r^2}} \sqrt{N - 2}$ $n = N - 2$	$u = \dfrac{z_1 - z_2}{\sigma_d}$ $z = \dfrac{1}{2} [l\,n\,(1 + r) - l\,n\,(1 - r)]$ $\sigma^2_d = s^2_{z1} + s^2_{z2}$ $s_z = \dfrac{1}{N - 3}$

Dans ce tableau:

u = distribution normale
t = distribution de Student-Fisher; n = degré de liberté
F = distribution de Fisher-Snedecor; n_1 et n_2 = degrés de liberté.

Pour les calculs pratiques on peut remplacer σ et ϱ par leurs estimations s et r tirées du ou des échantillons.

paraisons entre un échantillon et l'univers dont il est tiré (comparaison des paramètres estimés par l'observation aux valeurs exactes de l'univers) ou des comparaisons entre deux ou plusieurs échantillons (comparaison des paramètres estimés sans connaissance des valeurs exactes).

Dans les deux cas, on testera à l'hypothèse nulle, c'est-à-dire que l'on admettra a priori que le ou les échantillons en cause sont issus du même univers. Si le test est négatif, on pourra conclure soit que l'échantillon n'est pas issu de l'univers considéré, soit que les échantillons en présence émanent d'univers différents.

Comparaison de paramètres statistiques

Les erreurs d'échantillonnage des principales grandeurs statistiques, qu'il importe de connaître pour pouvoir procéder aux différents contrôles ou tests, sont résumées dans le tableau I.

La distribution des écarts entre deux caractéristiques observées, ou entre une valeur observée et sa valeur théorique ou exacte est différente, suivant l'extension de l'échantillon.

Les formules applicables à ces comparaisons sont résumées dans le tableau II.

Analyse de la variance

La variabilité d'une population statistique est imputable à une multitude de causes produisant chacune un petit effet. Or, parmi ces causes, il y en a certainement quelques-unes qui produisent chacune un effet appréciable. L'analyse de la variance consiste à dégager ces causes, dites causes contrôlées et à en mesurer l'effet par comparaison aux effets des causes incontrôlables que l'on désigne par l'expression variabilité ou erreur résiduelle.

Ces méthodes exigent le groupement des données observées en groupes et en sous-groupes homogènes faisant ressortir l'intervention systématique d'une ou de plusieurs des causes dont on recherche l'influence. Elles nécessitent très souvent l'adoption de plans d'expérimentation dont l'établissement à lui seul demande déjà des connaissances approfondies. Nous nous bornerons plus loin à donner quelques exemples simples de l'analyse de la variance.

4. Exemples pratiques

A. Comparaison de moyennes et de variances

On opère souvent, dans le domaine de la tarification comme dans les services d'exploitation et les services commerciaux, avec des moyennes. On parle de consommations moyennes, de prix moyens, d'abonnés moyens. On compare entre elles les moyennes de plusieurs années ou périodes successives; on tire de ces comparaisons des conclusions plus ou moins nettes, servant souvent à étayer des décisions ou des arguments. Or, ces comparaisons sont la plupart du temps aléatoires et mêmes trompeuses, car elles ne tiennent pas compte d'un facteur dont le rôle est pourtant essentiel: la dispersion.

Prenons, par exemple, un groupe d'abonnés domestiques et examinons quelle répercussion aura, sur leurs factures, l'introduction d'un nouveau tarif. Pour simplifier les calculs, nous en choisirons vingt au hasard. Le principe reste le même pour un échantillon plus étendu; seule la distribution à la base du test varie suivant qu'on considère une série de quelques valeurs ou un échantillon de plusieurs centaines ou milliers d'individus.

Les données relatives à ce groupe d'abonnés figurent au tableau III, qui permet aussi de se rendre compte de la marche des calculs. Les montants des factures individuelles en francs suisses sont indiqués en chiffres ronds et se rapportent, pour chaque abonné, à la consommation effective durant l'année servant de base à l'étude.

Tableau III

Abonné N°	Facture à l'ancien tarif		Facture au nouveu tarif		Différence nouveau-ancien tarif	
	x'	x'^2	x''	x''^2	$x_i = x'' - x'$	x_i^2
1	91	8 281	64	4 096	— 27	729
2	63	3 969	43	1 849	— 20	400
3	243	59 049	207	42 849	— 36	1 296
4	185	34 225	171	29 241	— 14	196
5	164	26 896	154	23 716	— 10	100
6	205	42 025	199	39 600	— 6	36
7	246	60 516	226	51 076	— 20	400
8	120	14 400	126	15 876	+ 6	36
9	187	34 969	190	36 100	+ 3	9
10	141	19 881	146	21 316	+ 5	25
11	112	44 944	121	14 641	+ 9	81
12	51	2 601	57	3 249	+ 6	36
13	24	576	22	484	— 2	4
14	30	900	37	1 369	+ 7	49
15	35	1 225	38	1 444	+ 3	9
16	36	1 296	38	1 444	+ 2	4
17	38	1 444	55	3 025	+ 17	289
18	27	729	45	2 025	+ 18	324
19	16	256	27	729	+ 11	121
20	11	121	26	676	+ 15	225
Sommes	2 025	358 303	1 992	294 803	— 33	4 369
Moyenn.	101,25	—	99,6	—	— 1,65	—

Pour que la différence entre les deux tarifs soit significative, il faut que la moyenne des différences individuelles s'écarte significativement de zéro. Si l'on prélevait un nombre quelconque d'échantillons, chacun d'étendue restreinte N, de l'univers dont le groupe d'abonnés à l'étude représente également un sondage, on constaterait que les valeurs de l'expression $t = \frac{\bar{x}}{s}\sqrt{N}$, calculées pour chaque échantillon, obéissent à une loi bien déterminée, la fonction t, qui a l'avantage d'être indépendante de l'écart-type de l'univers. Son allure est différente suivant le degré de liberté n, qui est ici égal à $N-1 = 19$. Plus le degré de liberté augmente, plus la distribution t se rapproche de la distribution gaussienne qui en est un cas particulier pour $n = \infty$. On trouve dans des tables spéciales les valeurs de t correspondant à $p = 0{,}05$ et $p = 0{,}01$, pour chaque valeur de n.

Le calcul donne

$$s^2 = \frac{1}{19}\left(4369 - \frac{33^2}{20}\right) = 227{,}0 \text{ et}$$

$$s = \sqrt{227} = 15{,}07$$

de sorte que

$$t = \frac{1{,}65\sqrt{20}}{15{,}07} = 0{,}49$$

Dans la table relative à la distribution t, on trouve sous $n = 19$:

pour $p = 0{,}05$ $t_{0,05} = 2{,}093$

et pour $p = 0{,}01$ $t_{0,01} = 2{,}861$

La valeur calculée pour t étant sensiblement inférieure à $t_{0,05}$, la différence entre les deux tarifs n'est pas significative. Autrement dit, la différence reste dans les limites de la dispersion inhérente à la population considérée. Le tarif n'aura donc, pour l'entreprise, aucune répercussion financière sensible, à consommation égale, bien entendu, et à condition que les abonnés ne modifient pas leurs habitudes de consommation après l'introduction du nouveau tarif. Ce calcul suppose, en outre, que le nouveau tarif sera obligatoire pour tous les abonnés.

Cet exemple confirme ce que nous avancions plus haut au sujet des moyennes. Si, pour juger de l'effet du nouveau tarif, on s'était contenté de considérer les sommes ou les moyennes des factures, on aurait simplement constaté que l'introduction du nouveau tarif entraîne une chute de recettes de l'ordre de 1,5 pour 100. On aurait alors été tenté d'admettre que la baisse aurait le même effet pour chaque abonné. Or, un regard sur le tableau 1 suffit pour voir que les abonnés seront touchés de façon très différente. Les uns subiront une augmentation substantielle de leur facture, alors que d'autres seront fortement privilégiés. La forte turbulence au sein de ce groupe d'abonnés ressort d'ailleurs clairement de la haute valeur de l'écart-type, qui est environ dix fois plus élevé que la moyenne des différences individuelles.

On peut aussi comparer entre elles les variances des deux séries de montants «ancien» et «nouveau» tarif. On calcule pour chaque série la variance s^2, puis on forme le quotient. Pour cette comparaison, on se sert de la fonction F, qui ne dépend que du degré de liberté des deux variances, et dont les valeurs, correspondant à $P = 0{,}05$ ou $0{,}01$, figurent également dans des tables. On pose $F = \frac{s'^2}{s''^2}$.

Pour les séries du tableau I, le quotient des deux variances s'^2 et s''^2 est identique au quotient des sommes des carrés, les deux séries ayant la même extension. Ce quotient se chiffre par 1,197. Dans les tables, on trouve pour $P = 0{,}05$:

$$F_{0,05} = 2{,}308 \text{ pour } n_1 = 12 \text{ et } n_2 = 19;$$

$$F_{0,05} = 2{,}114 \text{ pour } n_1 = 24 \text{ et } n_2 = 19.$$

La valeur exacte pour $n_1 = 19$ et $n_2 = 19$ peut se calculer par interpolation, mais cela n'est pas nécessaire dans ce cas, car la valeur effective de F (1,197) est sensiblement inférieure aux deux valeurs encadrant la valeur exacte. On en déduit que la différence entre les deux variances n'est pas significative. En d'autres termes, l'introduction du nouveau tarif n'entraînera pas un bouleversement des factures sortant du cadre de la dispersion régnant déjà sous le régime de l'ancien tarif.

B. Comparaison de régressions linéaires simples

Dans cet exemple, le problème consiste à rechercher l'influence de l'introduction d'un nouveau tarif sur le développement de la consommation d'énergie des abonnés. Si l'on connaît par exemple, pour les abonnés considérés, les chiffres de consommation pour plusieurs années avant et après le changement de tarif, on peut comparer entre elle les régressions (consommation en fonction des années) de chacune des deux périodes caractéristiques.

Les données numériques figurent au tableau IV.

Tableau IV

Abonné Nº	Consommation d'énergie en kWh à l'ancien tarif (y_i')					Consommation d'énergie en kWh au nouveau tarif (y_i'')				
	Année 1	Année 2	Année 3	Année 4	Total	Année 1	Année 2	Année 3	Année 4	Total
1	170	218	296	278	962	483	423	571	516	1 993
2	88	95	101	100	384	134	118	125	146	523
3	117	132	124	109	482	134	121	152	166	573
4	104	113	116	120	453	170	171	138	176	655
5	216	226	247	265	954	245	242	261	228	976
6	151	167	167	144	629	179	212	208	247	846
7	72	82	75	84	313	135	134	163	161	593
8	134	124	146	162	566	175	209	205	225	814
9	81	81	81	78	321	89	87	103	115	394
10	140	104	96	93	433	156	211	299	262	928
11	119	105	101	144	469	166	160	153	124	603
12	161	177	185	201	724	197	199	194	184	774
13	164	172	175	170	681	244	236	259	291	1 030
14	153	159	155	156	623	221	220	224	167	832
15	71	92	101	101	365	130	146	182	238	696
Total	1 941	2 047	2 166	2 205	8 359	2 858	2 889	3 227	3 246	12 230
Moyennes	129	136	144	147	139	191	193	216	216	204

a) calcul de la régression «ancien tarif»

$$Y' = \bar{y}' + b'(x' - \bar{x}')$$

$$\bar{y}' = 8359 : 15 = 139$$

$$\bar{x}' = (1 + 2 + 3 + 4) : 4 = 2{,}5$$

$$b' = (1941 + 2 \cdot 2047 + 3 \cdot 2166 + 4 \cdot 2205 -$$
$$- 2{,}5 \cdot 8539) : (450 - 150^2 : 60) = 6{,}07$$
$$Y' = 139 + 6{,}07\,(x' - 2{,}5) = 6{,}07\,x' + 124$$

b) calcul de la régression «nouveau tarif»

$$Y'' = \bar{y}'' + b''\,(x'' - \bar{x}'')$$
$$\bar{y}'' = 12230 : 15 = 204$$
$$\bar{x}'' = (1 + 2 + 3 + 4) : 4 = 2{,}5$$
$$b'' = (2858 + 2 \cdot 2889 + 3 \cdot 3227 + 4 \cdot 3246 -$$
$$- 2{,}5 \cdot 12230) : (450 - 150^2 : 60) = 10{,}08$$
$$Y'' = 204 + 10{,}08\,(x'' - 2{,}5) = 10{,}08\,x'' + 179$$

c) comparaison des deux coefficients de régression

$$t = \frac{b'' - b'}{\sigma_d}$$
$$\sigma_d^2 = \frac{S\,(y' - Y')^2 + S\,(y'' - Y'')^2}{N' + N'' - 4} \cdot$$
$$\cdot \left(\frac{1}{S\,(x' - \bar{x}')^2} + \frac{1}{S\,(x'' - \bar{x}'')^2}\right)$$
$$S\,(y' - Y')^2 = S\,y'^2 - \frac{1}{N}\,(S\,y')^2 -$$
$$- \frac{\left[S\,x'\,y' - \frac{1}{N}\,S\,x'\,S\,y'\right]^2}{S\,x'^2 - \frac{1}{N}\,(S\,x')^2}$$
$$S\,(y' - Y')^2 = 1\,330\,877 - \frac{8359^2}{60} -$$
$$- \frac{[21353 - 150 \cdot 8359 : 60]^2}{450 - 150^2 : 60} = 163563$$
$$S\,(y'' - Y'')^2 = 3277086 - \frac{12230^2}{60} -$$
$$- \frac{[31331 - 150 \cdot 12230 : 60]^2}{450 - 150^2 : 60} = 776591$$
$$\sigma_d^2 = \frac{163563 + 776591}{60 + 60 - 4} \cdot \left(\frac{1}{75} + \frac{1}{75}\right) = 216{,}13$$
$$\sigma_d = \sqrt{216{,}13} = 14{,}701$$
$$t = \frac{10{,}08 - 6{,}07}{14{,}701} = 0{,}273$$

Dans la table, on trouve, pour $n = 116$

$$t_{0,05} = 1{,}98 \qquad t_{0,01} = 2{,}617$$

Les deux coefficients de régression ne diffèrent pas entre eux de manière significative. Le nouveau tarif n'a donc pas influencé de manière sensible l'accroissement annuel de la consommation d'énergie. Par contre, la différence entre les moyennes $\bar{y}'$ et $\bar{y}''$ est fortement significative, comme le montre le contrôle suivant:

$$t = \frac{\bar{y}'' - \bar{y}'}{\sigma_d} \sqrt{\frac{N'\,N''}{N' + N''}}$$

où

$$\sigma_d^2 = \frac{1}{N' + N'' - 2}\,[S\,(y' - \bar{y}')^2 + S\,(y'' - \bar{y}'')^2]$$
$$\sigma_d^2 = \frac{1}{60 + 60 - 2}\,[166462 + 784204] = 8058{,}2$$
$$\sigma_d = \sqrt{8058{,}2} = 89{,}767$$
$$t = \frac{204 - 139}{89{,}767} \cdot \sqrt{\frac{60^2}{60 + 60}} = 3{,}966$$

pour

$n = 118,\; t_{0,05} = 1{,}98;\; t_{0,01} = 2{,}617;\; t_{0,001} = 3{,}373$

En d'autres termes, la probabilité d'une telle différence est si faible ($< 0{,}001$) que celle-ci ne peut être fortuite.

Le changement de tarif a donc provoqué une augmentation brusque de la consommation très significative.

C. Etude détaillée de la régression simple

Dans l'exemple choisi, il s'agissait d'établir, au cours d'une étude de tarif, la relation entre la chute de recettes et la surface réduite des locaux professionnels d'un certain nombre d'abonnés, artisans ou petits commerçants. Par chute de recettes, il faut entendre la différence entre la facture totale pour l'éclairage à l'ancien tarif et le produit de la consommation totale par le taux unique (prix du kWh) du nouveau tarif augmenté de la redevance fixe due pour les locaux d'habitation. La surface réduite des locaux professionnels est égale à leur surface réelle multipliée par un facteur dépendant de l'éclairement exigé pour le travail effectué dans le local. Les locaux sont classés en 4 catégories, pour lesquelles les facteurs sont 1, $1/2$, $1/4$ et $1/8$ respectivement.

L'échantillon comprend 86 abonnés.

Posons maintenant:

y = chute de recettes en fr.
x = surface réduite totale des locaux professionnels en m^2.
N = 86.

La valeur x variant entre 1 et 72, nous ferons des classes de largeur 3. Les centres de classe x_j seront 2, 5, 8 ...

Pour commencer, nous groupons les données en un tableau (V) auquel on donnera avantageusement la forme suivante qui permet de calculer d'emblée les y^2.

Tableau V

$x_j =$	2		5		8		....
	y_j	y_j^2	y_j	y_j^2	y_j	y_j^2	....
	39	1521	22	484	60	3 600	....
	19	361	75	5 625	48	2 304	....
	.	.	.	.	.	.	
	.	.	.	.	.	.	
	.	.	.	.	.	.	
$S =$	53	2259	411	24 607	1139	114 821	....

De ce tableau, nous passons au tableau VI dont les colonnes se retrouveront au cours des calculs qui suivent.

b) *Coefficient de détermination*

$$B = \frac{[S\,(x-\bar{x})\,(y-\bar{y})]^2}{S\,(x-\bar{x})^2\,S\,(y-\bar{y})^2} = \frac{b\,S\,(x-\bar{x})\,(y-\bar{y})}{S\,(y-\bar{y})^2}$$

$$S\,(y-\bar{y})^2 = S\,y_j^2 - \frac{1}{N}\,(S\,y_j)^2 =$$

$$= 3\,461\,123 - \frac{1}{86}\,12\,077^2 = 1\,765\,147$$

$$B = \frac{9{,}83 \cdot 139\,723}{1\,765\,147} = 0{,}778 = 77{,}8\,\%$$

c) *Test du coefficient de régression.*

Question: b est-il significativement différent de 0 ?

Tableau VI

x_j	f_j	$f_j\,x_j$	x_j^2	$f_j\,x_j^2$	$\overset{j}{S}\,y_j$	$(\overset{j}{S}y_j)^2$	$\overset{j}{S}\,y_j^2$	$\frac{1}{f_j}\,(\overset{j}{S}\,y_j)^2$	$x_j\,\overset{j}{S}\,y_j$
2	5	10	4	20	53	2 809	2 259	561,80	106
5	12	60	25	300	411	168 921	24 607	14 076,75	2 055
8	18	144	64	1 152	1 139	1 297 321	114 821	72 073,39	9 112
11	10	110	121	1 210	1 154	1 331 716	185 190	133 171,60	12 694
14	7	98	196	1 372	622	386 884	65 336	55 269,14	8 708
17	6	102	289	1 734	924	853 776	169 052	142 296,00	16 708
20	8	160	400	3 200	1 302	1 695 204	229 012	211 900,50	26 040
23	6	138	529	3 174	1 174	1 378 276	259 844	229 712,67	27 002
26	3	78	679	2 037	682	465 124	167 426	155 041,33	17 732
29	1	29	841	841	396	156 816	156 816	156 816,00	11 484
32	3	96	1 024	3 072	1 006	1 012 036	341 860	337 345,33	32 192
35	2	70	1 225	2 450	823	677 329	396 125	338 664,50	28 805
38	2	76	1 444	2 888	633	400 689	200 889	200 344,50	24 050
59	1	59	3 481	3 481	618	381 924	381 924	381 924,00	36 462
62	1	62	3 844	3 844	329	108 241	208 241	108 241,00	20 398
71	1	71	5 041	5 041	811	657 721	657 721	657 721,00	57 581
—	86	1 363	—	35 816	12 077	—	3 461 123	3 225 159,51	331 129
	$S\,f_j = N$	$S\,f_j\,x_j = S\,x$		$S\,f_j\,x_j^2 = S\,x^2$	$\overset{M}{S}\,\overset{j}{S}\,y_j = S\,y$		$\overset{M}{S}\,\overset{j}{S}\,y_j^2 = S\,y^2$	$\overset{M}{S}\,\frac{1}{f_j}\,(\overset{j}{S}\,y_j)^2$	$\overset{M}{S}\,(x_j\,\overset{j}{S}\,y_j) = S\,x\,y$

a) *Formule de régression*

$$Y = \bar{y} + b\,(x-\bar{x})$$

$$\bar{y} = \frac{1}{N}\,S\,f_j\,y_j = \frac{1}{86} \cdot 12\,077 = 140{,}43$$

$$\bar{x} = \frac{1}{N}\,S\,f_j\,x_j = \frac{1}{86} \cdot 1363 = 15{,}85$$

$$b = \frac{S\,(x-\bar{x})\,(y-\bar{y})}{S\,(x-\bar{x})^2}$$

$$S\,(x-\bar{x})\,(y-\bar{y}) = S\,x_j\,y_j - \frac{1}{N}\,(S\,f_j\,x_j)\,(S\,y_j) =$$

$$= 331\,129 - \frac{1}{86}\,1363 \cdot 12\,077 = 139\,723$$

$$S\,(x-\bar{x})^2 = S\,f_j\,x_j^2 - \frac{1}{N}\,(S\,f_j\,x_j)^2 =$$

$$= 35\,816 - \frac{1}{86}\,1363^2 = 14\,214$$

$$b = \frac{139\,723}{14\,214} = 9{,}83$$

$$Y = 140{,}43 + 9{,}83\,(x - 15{,}85) \qquad Y = 9{,}83\,x - 15{,}37$$

Nous posons:

$$t = \frac{b}{s_b}\sqrt{S\,(x-\bar{x})^2} \quad \text{où} \quad s_b^2 = \frac{1}{N-2}\,S\,(y-Y)^2$$

$$S\,(y-Y)^2 = S\,(y-\bar{y})^2 - S\,(Y-\bar{y})^2 =$$

$$= S\,y_j^2 - \frac{1}{N}\,(S\,y_j)^2 - \frac{\left[S\,x_j\,y_j - \frac{1}{N}\,S\,x_j\,S\,y_j\right]^2}{S\,f_j\,x_j^2 - \frac{1}{N}\,(S\,f_j\,x_j)^2} =$$

$$= 1\,765\,147 - \frac{139\,723^2}{14\,214} =$$

$$= 1\,765\,147 - 1\,373\,471 = 391\,676$$

$$s_b^2 = \frac{1}{84} \cdot 391\,676 = 4662{,}8 \qquad s_b = 68{,}285$$

$$t = \frac{9{,}83}{68{,}285} \cdot \sqrt{14\,214} = 17{,}163^{**}$$

Pour un degré de liberté

$$n = N - 2 = 84$$

$$t_{0,05} \cong 1{,}99 \quad \text{et} \quad t_{0,01} \cong 2{,}64$$

La valeur trouvée pour t étant sensiblement supérieure à celle de $t_{0,01}$, la régression est fortement significative.

d) *Analyse de la variance*

L'analyse de la variance repose sur le fait que la somme des carrés des écarts totaux est égale à la somme des sommes des carrés des écarts dus aux différentes causes contrôlés, augmentée de la somme des carrés des écarts dus au caractère aléatoire de la variable (erreur résiduelle). Chacune de ces sommes est affectée d'un degré de liberté. La comparaison des carrés moyens ou variances obtenus en divisant les sommes des carrés par leur degré de liberté avec la variance résiduelle permet de tester si les causes ont un effet significatif.

Dans notre cas, la variance totale peut être décomposée en une variance entre classes et une variance à l'intérieur des classes ou variance résiduelle. La variance entre classes elle-même, se compose de deux variances: l'une provenant de la régression (écarts entre les points de la droite de régression et la moyenne générale) et l'autre des déviations des moyennes de classes par rapport à la droite de régression. Nous obtiendrons ainsi le tableau VII (où M est égal au nombre de classes en x, soit 24).

Tableau VII

Origine de la variation	Somme des carrés	Degré de liberté
Régression linéaire	$S f_j (Y_j - \bar{y})^2 = SC_r$	1
Déviation par rapport à la droite de régression	$S f_j (\bar{y}_j - Y_j)^2 = SC_d$	$M - 2$
Totale entre classes	$S f_j (\bar{y}_j - \bar{y})^2 = SC_c$	$M - 1$
Résiduelle (intérieur des classes)	$S (y - \bar{y}_j)^2 = SC_e$	$N - M$
Totale	$S (y - \bar{y})^2 = SC_t$	$N - 1$

Nous connaissons déjà les sommes de carrés SC_t et SC_r; il suffit de calculer une des deux sommes SC_e ou SC_d pour trouver l'autre, car $SC_r + SC_d = SC_c$ et $SC_c + SC_e = SC_t$. La plus simple à calculer est SC_e, car

$$S (y - \bar{y}_j)^2 = S y^2 - S \frac{1}{f_j} (S y_j)^2 =$$

$$= 3461123 - 3225160 = 235963$$

L'analyse de la variance a la forme suivante (tableau VIII):

Le test $\dfrac{CM_c}{CM_e}$ fait ressortir que la corrélation entre y et x est fortement significative, indépendamment de toute hypothèse sur le caractère de cette liaison.

Le test $\dfrac{CM_d}{CM_e}$ indique que les déviations de la droite de régression sont légèrement significatives, de sorte que la régression peut être admise linéaire en première approximation.

e) *Interprétation*

La formule de régression indique que la chute de recettes est en moyenne de fr. 9,83 par m² de surface réduite, mais qu'il faut réduire de ce montant une somme fixe de fr. 15.37.

La valeur du coefficient de détermination fait ressortir que la corrélation est relativement étroite, car 78 % en chiffre rond de la variabilité des chutes de recettes s'explique par la variabilité de la surface réduite des locaux.

Le test du coefficient de régression montre que celui-ci est fortement différent de zéro. La régression est donc fortement significative.

La même conclusion ressort de l'analyse de la variance, de laquelle on déduit ensuite que la variabilité des déviations par rapport à la droite de régression n'est que faiblement significative. On peut donc admettre que la régression est linéaire en première approximation. Pour être exact, il faudrait maintenant rechercher le degré de l'équation de régression.

D. *Etude de la régression multiple*

Dans l'exemple précédent, nous avons admis, à titre provisoire, que la redevance pour les locaux d'habitation était donnée et que la chute de recettes n'avait aucune relation avec elle. Or, ce n'est pas le cas, de sorte qu'il faut étudier de quelle façon dépend la chute de recettes et des locaux d'habitation (nombre) et des locaux professionnels (surface réduite). Cela conduit à l'équation

$$Y = a + b_1 x_1 + b_2 x_2$$

où Y est la chute des recettes en fr.

x_1 le nombre de locaux d'habitation

x_2 la surface réduite des locaux professionnels en m²

Tableau VIII

Variance	Dégre de liberté n	Somme des carrés SC	Carré moyen CM	$F = \dfrac{CM}{CM_e}$
Régression	1	1 373 471	729 634 (CM_r)	—
Déviation de la régression	22	155 713	7 078 (CM_d)	1,8597*
Totale entre classes	23	1 529 184[1])	69 095 (CM_c)	18,154**
Résiduelle	62	235 963	3 806 (CM_e)	—
Totale	85	1 765 147	—	—

Des tables, on tire pour

$$n_1 = 22 \quad \text{et} \quad n_2 = 60$$

$$F_{0,05} = 1{,}70 \quad \text{et} \quad F_{0,01} = 2{,}12$$

[1]) On peut contrôler ce chiffre au moyen de la formule directe pour SC_c. On a

$$S f_j (\bar{y}_j - \bar{y})^2 = S f_j \bar{y}^2 - \frac{1}{N} (S f_j \bar{y}_j)^2 = \overset{M}{S} \frac{1}{f_j} (\overset{j}{S} y_j)^2 - \frac{1}{N} (\overset{j}{S} y_j)^2$$

Le calcul donne: $SC_c = 3\,225\,160 - \frac{1}{86} \cdot 12\,077^2 = 1\,529\,184.$

a, b_1 et b_2 des constantes à déterminer
b_1 et b_2 sont les coefficients de régression multiple.

L'équation peut se mettre sous la forme

$$Y - \bar{y} = b_1 (x_1 - \bar{x}_1) + b_2 (x_2 - \bar{x}_2)$$

Pour les calculs, on dresse avantageusement un tableau contenant:

a) les variables x_1, x_2, y et leur somme T
b) les carrés des variables et de leur somme
c) les produits des variables deux par deux
d) les sommes de chacune des colonnes.

Cela permet de calculer les sommes des carrés suivantes:

$$S_1 = S(x_1 - \bar{x}_1)^2 = S x_1^2 - \frac{1}{N}(S x_1)^2$$

$$S_2 = S(x_2 - \bar{x}_2)^2 = S x_2^2 - \frac{1}{N}(S x_2)^2$$

$$S_3 = S(y - \bar{y})^2 = S y^2 - \frac{1}{N}(S y)^2$$

$$S_4 = S(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) = S x_1 x_2 - \frac{1}{N}(S x_1)(S x_2)$$

$$S_5 = S(x_1 - \bar{x}_1)(y - \bar{y}) = S x_1 y - \frac{1}{N}(S x_1)(S y)$$

$$S_6 = S(x_2 - \bar{x}_2)(y - \bar{y}) = S x_2 y - \frac{1}{N}(S x_2)(S y)$$

$$S_7 = S(T - \bar{T})^2 = S T^2 - \frac{1}{N}(S T)^2$$

La dernière somme sert au contrôle, car on démontre facilement que

$$S_1 + S_2 + S_3 + 2(S_4 + S_5 + S_6) = S_7$$

Pour les coefficients de régression partielle on obtient alors

$$b_1 = \frac{S_6 \cdot S_4 - S_5 \cdot S_2}{S_4^2 - S_1 \cdot S_2}$$

$$b_2 = \frac{S_5 \cdot S_4 - S_6 \cdot S_4}{S_4^2 - S_1 \cdot S_2}$$

$$B = \frac{b_1 \cdot S_5 + b_2 \cdot S_6}{S_3}$$

(à suivre)

Compteur à indicateur de puissance maximum à double cadran

Par *A. Berner, H. Feuz* et le Service de l'électricité de Neuchâtel

621.317.785

Les auteurs décrivent un compteur de leur conception comportant deux cadrans entrant alternativement en fonction à la fin d'une période de facturation.

Die Verfasser beschreiben einen von ihnen entwickelten Zähler mit zwei Messwerken, die abwechslungsweise auf Ende einer Abrechnungsperiode in Betrieb gesetzt werden können.

L'inconvénient majeur des compteurs à indicateur de puissance maximum courants réside dans la remise à zéro de l'index.

Celle-ci étant faite lors de la lecture des compteurs, il ne subsiste aucun témoin en cas de contestation. La lecture peut certes se faire en présence de l'abonné mais il en résulte une perte de temps considérable pour les releveurs, perte de temps d'autant plus appréciable que le nombre de ces compteurs est plus élevé.

En outre, les contrats de vente d'énergie réclamant de tels compteurs (tarifs binômes en particulier) ayant souvent un caractère mensuel, les releveurs devraient être simultanément chez tous les abonnés le dernier jour du mois, ce qui est pratiquement impossible.

Il existe bien des compteurs enregistrant graphiquement la puissance maximum tels que «maxigraphe» et «printo-maxigraphe», mais le prix de ces appareils est souvent disproportionné à l'importance de l'abonnement.

Le compteur à double cadran représenté à la fig. 1 (brevet suisse n° 294 742) remédie à ces inconvénients.

Il possède deux cadrans fonctionnant alternativement et portant chacun:

— un indicateur de puissance maximum (kW),
— deux numéroteurs, l'un pour le haut tarif, l'autre pour le bas tarif (kWh),
— un signal indiquant si le cadran est en service ou non.

Admettons qu'il s'agisse de relevés mensuels, que nous soyons en mi-juin et que kWh et kW s'enregistrent sur le cadran supérieur «S». Le cadran inférieur «I» est à l'arrêt depuis le 31 mai; une plaque signalétique «Arrêt» l'indique. Ce cadran «I» hors service donne les index en kW et kWh arrêtés au 31 mai.

Le 30 juin à minuit, l'enregistrement sera commuté du cadran «S» au cadran «I» par télécommande ou par horloge. Au moment où cette commutation s'effectue, l'index kW du cadran «I» revient automatiquement à zéro, la plaquette «Arrêt» du cadran «I» disparaît, et apparaît au cadran «S» lequel est immobilisé et l'enregistrement s'effectue sur le cadran «I» du 1er juillet à 00.00 h au 31 juillet à minuit.

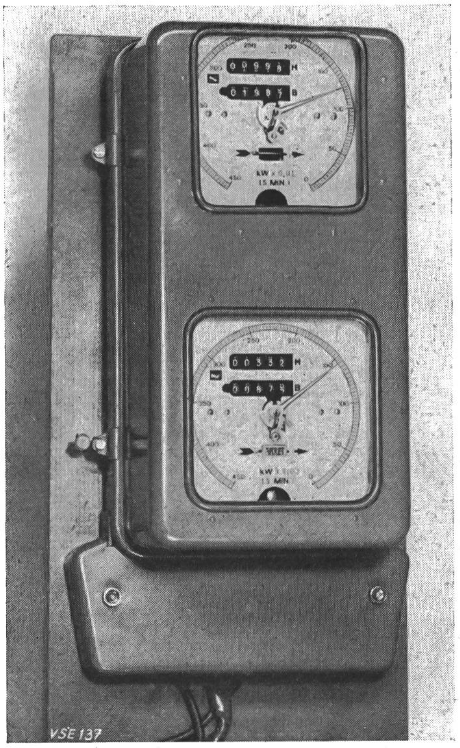

Fig. 1
Compteur à indicateur de puissance maximum à deux cadrans

Ainsi la position des index kW et kWh haut et bas tarif du 30 juin à minuit reste sur le cadran «S» durant tout le mois de juillet. Le releveur dispose d'un mois pour relever les index du 30 juin et l'abonné dispose aussi d'un mois pour contrôler son compteur.

Le fait que les index kWh sont aussi bloqués à date et à heure fixes est particulièrement intéressant aux changements saisonniers de tarifs, pour les abonnés importants surtout.

De plus, si l'abonnement est conçu de manière que la puissance maximum ne soit enregistrée que pendant certaines heures, les heures de pointe par exemple, il est possible de mettre l'enregistrement de la puissance maximum hors service pendant les heures creuses, sur le cadran en activité.

Toutes les commandes:

- commutation des cadrans,
- passage du haut tarif au bas tarif et inversement,
- hors service et en service de l'indicateur de puissance maximum

peuvent s'effectuer sans difficulté par télécommande centralisée ou par horloge, les horloges existantes pouvant sans difficulté être munies d'un dispositif additionnel pour la commande mensuelle.

Communications de nature économique

Unsere Kraft — Die Elektrizität

C'est sous ce titre qu'a paru, en langue allemande, aux éditions Orell Füssli à Zurich, une brochure de Josef Jaeger traitant de l'économie électrique suisse. Cet opuscule s'adresse avant tout au grand public. L'auteur s'y est fixé comme tâche de mettre les principaux problèmes de notre économie électrique à la portée de tout le monde et de donner une idée claire de la nature de cette économie. Les développements concernant les exportations d'énergie électrique sont tout spécialement d'actualité, de même que ceux qui traitent des relations entre l'économie électrique et la protection de la nature et des sites.

La brochure a été richement illustrée par M. W. Bär.

Communications des organes de l'UCS

Commission fédérale des poids et mesures

Dans le rapport de gestion de l'UCS (Bull. ASE Bd. 45 (1954), n° 11, p. 427/15), nous avons communiqué par erreur, que M. le directeur Thorens, depuis décédé, avait été élu président de la Commission fédérale des poids et mesures. Nous prions nos lecteurs d'excuser cette erreur.

Le Conseil fédéral a élu entre temps comme nouveau président de cette commission, à la place de M. le professeur Paul Joye, Fribourg, qui s'était retiré à la fin de 1953, M. le directeur *Karl Bretscher*, Dr. h. c., administrateur-délégué de la maison Winkler, Fallert, S. A. à Berne.

Fonds du Centenaire de l'EPF 1955; dons des entreprises d'électricité

La collecte, que l'UCS a organisée parmi ses membres au bénéfice du fonds du Centenaire de l'EPF 1955, a donné jusqu'ici des résultats très réjouissants. Nous avons reçu environ 200 dons d'un montant total de 390 000.— fr. en chiffre rond. On relève parmi eux 3 dons de 50 000.— fr. chacun, 10 dons entre 10 000.— fr. et 30 000.— fr. et 29 dons entre 1000.— fr. et 7000.— fr. Nous remercions vivement tous les donateurs.

La collecte continue; les dons sont à verser au Compte de chèques postaux VIII 4417 à Zurich.

Rédaction des «Pages de l'UCS»: Secrétariat de l'Union des Centrales Suisses d'Electricité, Seefeldstrasse 301, Zürich 8, téléphone (051) 34 12 12; compte de chèques postaux VIII 4355; adresse télégraphique: Electrunion Zürich.

Rédacteur: *Ch. Morel,* ingénieur.

Des tirés à part de ces pages sont en vente au secrétariat de l'UCS, au numéro ou à l'abonnement.