La vérification de l'hypothèse sur la constance des probabilités.

Par

Stanisław Kołodziejczyk.

(Laboratoire Biométrique de l'Institut Nencki, Soc. Scient. de Varsovie).

1. Dans un mémoire récent Mrs. J. Neyman et E. S. Pearson proposent une méthode générale [1]) de vérification des hypothèses statistiques.

Ayant donné un fait observè F et une hypothèse statistique H, on calcule ce qu'on appelle la vraisemblance de cette hypothèse λ_H. Si λ_H est petite, par ex. si $\lambda_H \leqslant \lambda_0$, on convient de conclure que l'hypothèse est probablement fausse. Si au contraire $\lambda_0 < \lambda_H$, on admet qu'on n'a pas de raison suffisante pour une conclusion pareille. Le nombre λ_0 doit être choisi de manière que le danger de rejection d'une hypothèse vraie ne soit pas trop grand. Ce danger est mesuré par P_{λ_0} — la valeur de la probabilité pour qu'on ait $\lambda_H \leqslant \lambda_0$, déterminée par l'hypothèse considérée H, ou la borne supérieure de cette probabilité, si celle ci n'est pas déterminée par H. Il est évident d'ailleurs, que si l'on rejette l'hypothèse H lorsque $\lambda_H \leqslant \lambda_0$, la probabilité π pour qu'on rejette une hypothèse vraie ne surpasse pas P_{λ_0}. Or, si λ_0 est choisi de telle façon que $P = \varepsilon$, ε étant un nombre arbitraire entre zèro et un, et qu'on accepte la règle de rejeter l'hypothèse H lorsque $\lambda_H \leqslant \lambda_0$ et de l'accepter dans les autres cas, on peut être sur qu'une hypothèse vraie sera rejetée avec une fréquence moyenne, qui ne surpasse pas ε.

---

[1]) J. Neyman and E. S. Pearson: *On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference.* Biometrika Vol. XX—A. P. 175—240 et 264—294.

Le point essentiel dans la méthode de Mrs. J. Neyman et E. S. Pearson consiste dans le principe que le degré de notre confiance en une hypothèse statistique H peut être mesuré par la valeur de la vraisemblance λ_H. Evidemment ce principe peut être admis ou rejeté, selon qu'il semble intuitif ou non. De ce point de vue il est intéressant que toutes les méteodes de vérification des hypothèses, qui sont entrées en usage général et qui ont été examinées [1]), sont des conséquences du principe mentionné.

Le but de la note présente est de démontrer, que la méthode bien connue de Lexis-Bortkiewicz de vérification de l'hypothèse sur la constance des probabilités dans plusieurs séries des épreuves indépendantes est — elle aussi — une conséquence du principe des Mrs. J. Neyman et E. S. Pearson [2]).

2. Rappelons la terminologie. Soit F un fait, déterminé par les coordonnées

$$x_1, x_2, \ldots x_s \tag{1}$$

qui peuvent varier dans certaines limites, et h — une hypothèse concernant F. Si h détermine la probabilité (sensu stricto, ou la probabilité élémentaire) de F, nous dirons, que h est une hypothèse statistique simple. Toute hypothèse H, qui n'est pas simple est dite composée. Il est évident que l'hypothèse composée H peut être transformée en une hypothèse simple. Il suffit pour celà d'adjoindre à H arbitrairement quelques suppositions supplémentaires pour que l'ensemble de ces suppositions détermine la probabilité de F. Si h est une hypothèse simple, qui peut être obtenue de H par ce procédé, nous dirons, que h appartient à H.

Soit Ω l'ensemble des hypothèses simples qu'on considère dans un cas donné comme admissibles. Soit h une de ces hypo-

[1]) J. Neyman and E. S. Pearson: *On the Use and Interpretation...* Loc. cit.

J. Neyman and E. S. Pearson: *On the Problem of Two Samples.* Bull. de l'Académie Polonaise des Sciences et des Lettres. A, 1930, p: 73—96.

J. Neyman: *Contribution to the Theory of Certain Test Criteria.* Bull. de l'Institut International de Statistique, 1929. P. 44—88.

J. Neyman: *Sur la limite de la vraisemblance de l'hypothèse.* C. R., t. 188, p. 1560.

J. Neyman: *Sur une méthode de vérification des hypothèses.* Ibidem p. 1467.

[2]) Ce problème m'a été posé par Mr. J. Neyman, à qui je dois aussi quelques indications sur la méthode.

thèses et P_{hF} — la probabilité (sensu stricto ou la probabilité élémentaire) de F, déterminée par h. Nous allons supposer que l'ensemble des valeurs de P_{hF}, qui correspondent à un même fait F est borné, quel que soit F. Soit P_F la borne supérieure de P_{hF} par rapport a l'ensemble Ω et pour un fait fixé F.

Si l'on a observé F, on appelle la vraisemblance de l'hypothèse simple h le rapport

$$\lambda_h = \frac{P_{hF}}{P_F}. \tag{2}$$

Si H est une hypothèse composée, pour définir la vraisemblance de H lorsqu'on a observé F, on considère le sousensemble Ω_H des hypothèses simples qui appartiennent à H. Soit H_{HF} la borne supérieure des nombres P_{hF} par rapport à l'ensemble Ω_H et le fait F. La vraisemblance λ_H de H est alors

$$\lambda_H = \frac{P_{HF}}{P_F}. \tag{3}$$

3. Le problème, que nous allons considérer, peut être précisé comme suit.

Nous considérons les épreuves indépendantes qui peuvent donner lieu à un des deux événements: E ou sa négation $\overline{E}$. Ces épreuves sont effectuées en s séries et soit n_i le nombre des épreuves, qui appartiennent à l'i-ème série ($i = 1, 2, \ldots s$). Supposons qu'on sait, que la probabilité de E dans les épreuves appartenant à une même série est constante. Soit p_i la valeur de cette probabilité qui correspond à l'i-ème série des épreuves.

Le fait observé F consiste dans les nombres k_i des épreuves de l'i-ème série, qui ont donné lieu à l'événement E, pour $i = 1, 2 \ldots s$.

L'hypothèse H, qu'on doit vérifier, affirme que

$$p_1 = p_2 = \ldots = p_s \tag{4}$$

c'est-à-dire que la probabilité de l'événement E était la même dans toutes les séries des épreuves. Désignons par p la valeur de cette probabilité. Il est à remarquer, que p n'est pas déterminée par l'hypothèse H, qui par conséquent est une hypothèse composée.

Nous allons vérifier cette hypothèse par rapport à l'ensemble Ω des hypothèses admisibles, qui renferme toute hypothèse h précisant les valeurs des probabilités p_i quelconques $0 \leqslant p_i \leqslant 1$ pour $i = 1, 2, \ldots s$.

4. La probabilité du fait observé F déterminée par une hypothèse h est

$$P_{hF} = \prod_{i=1}^{s} C_{n_i}^{k_i} p^{k_i} (1 - p_i)^{n_i - k_i} \tag{5}$$

La valeur maximum de P_{hF} est égale à

$$P_F = \prod_{i=1}^{s} C_{n_i}^{k_i} q_i^{k_i} (1 - q_i)^{n_i} \tag{6}$$

où

$$q_i = \frac{k_i}{n_i}. \tag{7}$$

La probabilité du fait F, déterminée par une hypothèse simple qui appartient à H est égale à

$$P'_{HF} = p^{k_0}(1 - p)^{n_0 - k_0} \prod_{i=1}^{s} C_{n_i}^{k_i} \tag{8}$$

où

$$k_0 = \sum_{i=1}^{s} k_i; \qquad n_0 = \sum_{i=1}^{s} n_i. \tag{9}$$

Posons encore

$$q_0 = \frac{k_0}{n_0}. \tag{10}$$

Alors le maximum de P'_{HF} sera

$$P_{HF} = q_0^{k_0}(1 - q_0)^{n_0 - k_0} \prod_{i=1}^{s} C_{n_i}^{k_i} \tag{11}$$

et la vraisemblance λ_H de l'hypothèse H

$$\lambda_H = \frac{q_0^{k_0}(1 - q_0)^{n_0 - k_0}}{\prod_{i=1}^{s} q_i^{k_i}(1 - q_i)^{n_i - k_i}}. \tag{12}$$

La formule (12) présente la solution de la première partie du problème. Pour la compléter il faudrait considérer la probabilité

$$P\{\lambda_H \leqslant \lambda_0\} \tag{13}$$

pour qu'on ait $\lambda_H \leqslant \lambda_0$, λ_0 étant un nombre positif quelconque. En cas où cette probabilité serait déterminée par l'hypothèse H, c'est-à-dire si elle était indépendante de la valeur commune p des probabilités de l'évenement E dans toutes les s séries des épreuves, qui n'est pas déterminée par H, la solution serait complétée par le calcul de (13). Si au contraire $P\{\lambda_H \leqslant \lambda_0\}$ n'était pas déterminée par H, il serait nécessaire de calculer la borne supérieure de $P\{\lambda_H \leqslant \lambda_0\}$.

Evidemment

$$P\{\lambda_H \leqslant \lambda_0\} = \sum \prod_{i=1}^{s} C_{n_i}^{k_i} p^{k_i} (1-p)^{n_i - k_i}, \tag{14}$$

où la somme Σ s'étend sur tous les systèmes des valeurs des nombres k_i pour lesquels

$$o \leqslant k_i \leqslant n_i \qquad (i = 1, 2, \ldots s) \tag{15}$$

$$\lambda_H = \prod_{i=1}^{s} \frac{q_0^{k_i}(1-q_0)^{n_i - k_i}}{q_i^{k_i}(1-q_i)^{n_i - k_i}} \leqslant \lambda_0. \tag{16}$$

Le calcul de la somme (14) semble inabordable et nous nous bornerons au calcul de sa limite, lorsque (a) le nombre n_0 des épreuves effectuées croît indéfiniment de telle manière que

$$n_i \geqslant \nu n_0 \tag{17}$$

ν étant un nombre positif fixe. Nous supposerons de plus que (b)

$$o < p < 1, \tag{18}$$

ce qui est d'ailleurs évidemment une limitation sans importance.

Désignons par $I(A)$ l'intégrale

$$I(A) = \frac{\prod_{i=1}^{s} \sqrt{n_i}}{(2\pi p(1-p))^{\frac{s}{2}}} \int \ldots \int_A e^{-\frac{\sum_{i=1}^{s} n_i (q_i - p)^2}{2p(1-p)}} dq_1 dq_2 \ldots dq_s \tag{19}$$

étendue sur un domaine A dans l'espace à s dimensions, et où $q_1, q_2, \ldots q_s$ désignent les variables continues. Considérons λ_H comme une fonction de ces variables et désignons par W_0 le domaine défini par l'inégalité

$$\lambda_H > \lambda_0. \tag{20}$$

Si les conditions (a) et (b) sont remplies, et si le nombre n_0 est assez grand, on peut appliquer le théorème de Laplace et écrire

$$P\{\lambda_H > \lambda_0\} = I(W_0) + \eta. \tag{21}$$

$|\eta|$ étant si petit que l'on veut. Fixons un nombre positif arbitraire ε et choisissons N assez grand pour qu'on aît $|\eta| < \varepsilon$, lorsque $n_0 > N$.

Vu les propriétes connues de la fonction sous le signe de l'intégrale (19) on peut trouver un tel nombre positif χ_0 que

$$I(W_1) = \left(\frac{1}{\sqrt{2\pi}}\right)^s \int \dots \int_{W_1'} e^{-\frac{\sum_{i=1}^{s} y_i^2}{2}} dy_1 dy_2 \dots dy_s > 1 - \varepsilon, \tag{22}$$

où W_1 désigne le domaine défini par l'inégalité

$$\chi^2 = \frac{\sum_{i=1}^{s} (q_i - p)^2}{p(1-p)} \leqslant \chi_0^2 \tag{23}$$

et W_1' est la transformation de W_1 par les formules

$$q_i = p + y_i \sqrt{\frac{p(1-p)}{n_i}}, \qquad i = 1, 2, \dots s. \tag{24}$$

Remarquons que le domaine W_1', et par conséquent l'intégrale $I(W_1)$, sont indépendants de n_i, $(i = 1, 2, \dots, s)$.

Considérons W_2 — la partie commune des domaines W_0 et W_1. Il est aisé de voir que

$$P\{\lambda_H > \lambda_0\} = I(W_2) + \eta_2 \tag{25}$$

où $|\eta_2| < 2\varepsilon$, lorsque

$$n_0 > N. \tag{26}$$

Les raisonnements qui suivent ont pour but de trouver deux domaines W_5 et W_6 tels que

$$W_5 \subset W_2 \subset W_6 \tag{27}$$

et que les intégrales $I(W_5)$ et $I(W_6)$ tendent vers une même limite $1 - P_{\lambda_0}$ lorsque $n_0 \to \infty$. Il est clair qu'alors le même nombre sera aussi la limite de la probabilité $P\{\lambda_H > \lambda_0\}$.

Dans tout point du domaine W_2 on a

$$|q_i - p| \leqslant \frac{\chi_0 \sqrt{p(1-p)}}{\sqrt{n_i}} \leqslant \frac{\chi_0}{\sqrt{n_0}} \sqrt{\frac{p(1-p)}{\nu}} \tag{28}$$

$$|q_0 - p| \leqslant \frac{\chi_0}{\sqrt{n_0}} s \sqrt{\nu p(1-p)} \tag{29}$$

$$|q_i - q_0| \leqslant \frac{\chi_0}{\sqrt{n_0}} \frac{(s\nu + 1)\sqrt{p(1-p)}}{\nu}. \tag{30}$$

Posons

$$q_i = q_0 + x_i \sqrt{\frac{q_0(1-q_0)}{n_i}} \tag{31}$$

Si le nombre N est assez grand il résulte de (26) et (29) qu'il existe un nombre positif $\alpha < \frac{1}{2}$ tel que

$$\alpha < q_0 < 1 - \alpha \tag{32}$$

donc que

$$\frac{\alpha}{1-\alpha} < \frac{q_0}{1-q_0} < \frac{1-\alpha}{\alpha} \tag{33}$$

et

$$\frac{\alpha}{1-\alpha} < \frac{1-q_0}{q_0} < \frac{1-\alpha}{\alpha}. \tag{34}$$

Supposons que N est assez grand et fixons α satisfaisant (32). Il résulte alors de (30), (33) et (34) que les variables x_i $(i = 1, 2, \ldots s)$ sont bornées dans leur ensemble et p. ex. que $|x_i| < M$ pour $i = 1, 2, \ldots s$. Un calcul simple donne alors

$$-\log \lambda_H = \sum_{i=1}^{s} n_i \left(q_0 + x_i \sqrt{\frac{q_0(1-q_0)}{n_i}} \right) \left[x_i \sqrt{\frac{1-q_0}{n_i q_0}} - \right.$$

$$\left. - \frac{1}{2} x_i^2 \frac{1-q_0}{n_i q_0} + \frac{1}{\left(1 + \Theta_i' \sqrt{\frac{1-q_0}{n_i q_0}}\right)^3} \frac{x_i^3}{3} \left(\frac{1-q_0}{n_i q_0} \right)^{3/2} \right] -$$

$$- \sum_{i=1}^{s} n_i \left(1 - q_0 - x_i \sqrt{\frac{q_0(1-q_0)}{n_i}} \right) \left[x_i \sqrt{\frac{q_0}{n_i(1-q_0)}} + \right.$$

$$+\frac{1}{2}x_i^2\frac{q_0}{n_i(1-q_0)}+\frac{1}{\left(1-\Theta_i''\sqrt{\frac{q_0}{n_i(1-q_0)}}\right)^3}\frac{x_i^3}{3}\left(\frac{q_0}{n_i(1-q_0)}\right)^{3/2}\Bigg]=$$

$$(35)\qquad =\frac{1}{2}\sum_{i=1}^{s}x_i^2+\eta_3\,,\qquad 0\leqslant\Theta_i',\,\Theta_i''\leqslant 1$$

où $|\eta_3|<\frac{\varepsilon}{2}$ dans tout point du domaine W_2 pourvu que N soit assez grand et l'inégalité (26) — satisfaite.

Posons

$$(36)\qquad Q_0^2=-2\,lg\lambda_0$$

et considérons deux domaines W_3 et W_4 définis par des inégalités

$$(37)\qquad Q^2=\sum_{i=1}^{s}x_i^2=\sum_{i=1}^{s}\frac{(q_i-q_0)^2n_i}{q_0(1-q_0)}\leqslant Q_0^2-\varepsilon$$

$$(38)\qquad \chi^2=\sum_{i=1}^{s}\frac{(q_i-p)^2n_i}{p(1-p)}\leqslant\chi_0^2$$

et

$$(39)\qquad Q^2\leqslant Q_0^2+\varepsilon$$

$$(40)\qquad \chi^2\leqslant\chi_0^2$$

respectivement. Il est clair que

$$(41)\qquad W_3\subset W_2\subset W_4$$

et par conséquent

$$(42)\qquad I(W_3)\leqslant I(W_2)\leqslant I(W_4).$$

Observons que les inégalités (37) et (39) sont équivalentes à

$$(43)\qquad \sum_{i=1}^{s}\frac{(q_i-q_0)^2n_i}{p(1-p)}\leqslant(Q_0^2-\varepsilon)\left(1+\frac{q_0-p}{p}\right)\left(1-\frac{q_0-p}{1-p}\right)$$

et

$$(44)\qquad \sum_{i=1}^{s}\frac{(q_i-q_0)^2n_i}{p(1-p)}\leqslant(Q_0^2+\epsilon)\left(1+\frac{q_0-p}{p}\right)\left(1-\frac{q_0-p}{1-p}\right)$$

respectivement. A cause de (29) et au moyen de l'augmentation éventuelle du nombre N on peut satisfaire les inégalités:

$$(45) \qquad Q_0^2 > (Q_0^2 - \varepsilon)\left(1 + \frac{q_0 - p}{p}\right)\left(1 - \frac{q_0 - p}{1 - p}\right) > Q_0^2 - 2\varepsilon$$

et

$$(46) \qquad Q_0^2 < (Q_0 + \varepsilon)\left(1 + \frac{q_0 - p}{p}\right)\left(1 - \frac{q_0 - p}{1 - p}\right) < Q_0^2 + 2\varepsilon.$$

Désignons par W_5 et W_6 les domaines définis par les inégalités

$$(47) \qquad \sum_{i=1}^{s} \frac{(q_i - q_0)^2 n_i}{p(1-p)} \leqslant Q_0^2 - 2\varepsilon$$

$$(48) \qquad \chi^2 \leqslant \chi_0^2$$

et

$$(49) \qquad \sum_{i=1}^{s} \frac{(q_i - q_0) n_i}{p(1-p)} \leqslant Q_0^2 - 2\varepsilon$$

$$(50) \qquad \chi^2 \leqslant \chi_0^2$$

respectivement. Il est clair que si N est assez grand, on a pour tout $n_0 > N$

$$(51) \qquad W_5 \subset W_3 \subset W_2 \subset W_4 \subset W_6$$

et par conséquent

$$(52) \qquad I(W_5) \leqslant I(W_2) \leqslant I(W_6).$$

Pour calculer les limites des intégrales $I(W_5)$ et $I(W_6)$, considérons les domaines, soit W_7 et W_8, qui correspondent aux inégalités (47) et (49) respectivement. A cause des propriétés du nombre χ_0 nous aurons

$$(53) \qquad I(W_7) = I(W_5) + \eta_4$$

$$(54) \qquad I(W_8) + I(W_6) + \eta_5$$

où $0 \leqslant \eta_4 < \varepsilon$ et $0 \leqslant \eta_5 < \varepsilon$.

Les intégrales $I(W_7)$ et $I(W_8)$ sont des cas particuliers de l'intégrale $I(W_\sigma)$ prise dans le domaine W_σ, où l'on a

$$\sum_{i=1}^{s} \frac{(q_i - q_0)^2 n_i}{p(1-p)} \leqslant \sigma^2. \tag{55}$$

Le calcul de l'intégrale $I(W_\sigma)$ ne présente aucune difficulté. La transformation (24) donne immédiatement

$$I(W_\sigma) = \left(\frac{1}{\sqrt{2\pi}}\right)^s \int \cdots \int e^{-\frac{1}{2}\sum_{i=1}^{s} y_i^2} \, dy_1 dy_2 \ldots dy_s \tag{56}$$

où l'intégration s'étend à un domaine, où l'on a

$$\sum_{i=1}^{s} y_i^2 - \frac{\sqrt{n_l}}{n_0}\left(\sum_{i=1}^{s} y_i \sqrt{n_i}\right)^2 \leqslant \sigma^2 \tag{57}$$

Or la valeur de (56) est connue, savoir

$$I(W_\sigma) = C \int_0^\sigma t^{s-2} e^{-\frac{1}{2}t^2} dt \tag{58}$$

où

$$\frac{1}{C} = \int_0^\infty t^{s-2} e^{-\frac{1}{2}t^2} dt. \tag{59}$$

En combinant les résultats (25), (48), (51), (52) et (58), on peut écrire

$$C \int_0^{\sqrt{Q_{2l}+2\varepsilon}} t^{s-2} e^{-\frac{1}{2}t^2} dt + \eta_2 - \eta_4 \leqslant P\{\lambda_H > \lambda_0\} \leqslant C \int_0^{\sqrt{Q_{2l}-2\varepsilon}} t^{s-2} e^{-\frac{1}{2}t^2} dt, \tag{60}$$

d'où on conclut que

$$\lim_{n_0 \to \infty} P\{\lambda_H > \lambda_0\} = C \int_0^{Q_0} t^{s-2} e^{-\frac{1}{2}t^2} dt \tag{61}$$

donc que

$$\lim P\{\lambda_H \leqslant \lambda_0\} = C \int_{Q_0}^\infty t^{s-2} e^{-\frac{1}{2}t^2} dt = P_{\lambda_0}, \tag{62}$$

ce qu'il fallait démontrer. Il est à remarquer, que la limite de la

probabilité $P\{\lambda_H \leqslant \lambda_0\}$ est parfaitement déterminée par l'hypothèse H.

Si le nombre des épreuves effectuées n_0 est très grand, l'égalité (62) permet de mesurer la probabilité $P\{\lambda_H \leqslant \lambda_0\}$ par la valeur de sa limite P_{λ_0}.

La téchnique de la vérification de l'hypothèse H sur la constance de la probabilité pendant s séries des épreuves indépendantes consiste dans le calcul du nombre

$$Q_H^2 = -2\,lg\lambda_H. \tag{63}$$

Si ce nombre est grand, on conclut que la vraisemblance de l'hypothèse H est petite et on hésite d'accepter H. On considère ensuite les tables [1]) de l'intégrale (62) et on y trouve la valeur de

$$P_H = C\int_{Q_H}^{\infty} t^{s-2}e^{-\frac{1}{2}t^2}\,dt. \tag{64}$$

Si P_H est jugé petit, on n'hésite plus de rejeter H, vu que la probabilité pour qu'on rejette une hypothèse vraie, est dans ce cas plus petite que P_H.

Pour le calcul de Q_H on peut s'adresser à la formule

$$\lambda_H = \frac{q_0^{k_0}(1-q_0)^{n_0-k_0}}{\prod_{i=1}^{s} q_i^{k_i}(1-q_i)^{n_i-k_i}} \tag{65}$$

ou à la formule approchée

$$Q_H^2 = \frac{\sum_{i=1}^{s}(q_i-q_0)^2 n_i}{q_0(1-q_0)}. \tag{66}$$

On voit sans peine que cette dernière formule est en connexion simple avec le coefficient de dispersion D introduit pour la vérification de l'hypothèse H par Lexis et étudié par M. L. v. Bortkiewicz, savoir

$$Q_H^2 = (s-1)D^2. \tag{67}$$

On sait que si D est sensiblement plus grand que l'unité, d'après la méthode de Lexis il faudrait rejeter H. Le verdict serait le même si l'on partait du principe général des Mrs. J. Ney-

[1]) Karl Pearson: *Tables for Statisticians and Biometricians*. Cambridge, 1924.

man et E. S. Pearson. Ainsi dans le cas où l'on sait que la probabilité de l'événement E était constante pendant chaque série des épreuves, la méthode classique de la vérification de l'hypothèse H est une conséquence du principe mentionné.

Il est à remarquer que la circonstance que la loi de probabilité élémentaire de l'expression

$$z = \sqrt{\frac{\sum_{i=1}^{s} (q_i - q_0)^2 n_i}{q_0(1 - q_0)}} \tag{68}$$

peut être représentée approximativement par la fonction

$$Cz^{s-2} e^{-\frac{1}{2}z^2} \tag{69}$$

est connue [1]). Au contraire le résultat (62) semble être nouveau.

---

[1]) Voir p. ex. R. A. Fisher: *Statistical Methods for Research Workers*. Oliver and Boyd. 1925.