Proc. Estonian Acad. Sci. Humanities and Social Sciences, 1992, 41, 2, 87—94
https://doi.org/10.3176/hum.soc.sci.1992.2.04

Aasa MAAMÄGI *

L'ANALYSE STATISTIQUE DES PARTITIONS

Soit $\mathfrak{M}=\{S^i\}_{j=1}^{N(n,k)}$ l'ensemble de toutes les partitions possibles de n objets $(O=\{o_j\}_{j=1}^{n})$ en k classes (les classes sont disjointes et l'union des classes donne l'ensemble de tous les objets).

Si les classes sont descriptibles à l'avance (en ce cas nous les appelons classes nommées), il est aisé de voir, que le nombre des partitions $N(n,k)=N_1(n,k)=k^n$ $(\mathfrak{M}=\mathfrak{M}_1)$.

Si les classes sont indescriptibles à l'avance (en ce cas nous les appelons classes non-nommées), le nombre des partitions $N(n,k)=N_2(n,k)=$ $=S(n,k)$, où $S(n,k)$ sont les nombres de Stirling de deuxième espèce, parce que dans ces conditions il n'existe pas de classes sans objets $(\mathfrak{M}=\mathfrak{M}_2)$.

Si nous voulons aussi ranger les classes, alors $N(n,k)=N_3(n,k)=$ $=k!S(n,k)$ $(\mathfrak{M}=\mathfrak{M}_3)$.

On peut remarquer ici, que si $k=n$, $\mathfrak{M}_3$ est l'espace des rangements.

Le problème à résoudre est le suivant: comment tester l'hypothèse H:

$$P(R=S^j)=p_j, \qquad S^j\subset\mathfrak{M}; \qquad j=1, 2, \ldots, N(n,k),$$

où $P(R=S^j)$ est la probabilité de l'événement associé à l'apparition de la partition S^j dans l'épreuve; $\{p_j\}_{j=1}^{N(n,k)}$ sont fixés.

Les résultats obtenus suivent.

Nous examinons principalement les cas où $p_j=\dfrac{1}{N(n,k)}$ (H_0) pour tout j et le cas $N(n,k)=N_i(n,k)$; $i=2,3$ (également $N(n,k)=$ $=\sum_{i=1}^{k} S(n,i)$, $N(n,k)=\sum_{i=1}^{k} i!S(n,i)$); k est donné à l'avance.

Nous voulons accentuer encore une fois que dans $\mathfrak{M}_2$ et $\mathfrak{M}_3$ il n'y a pas de classes vides et par conséquent il n'y a pas de raison de parler de la probabilité de l'événement, qu'un objet quelconque se trouve dans une certaine classe parce que les classes ne sont déterminées que par les objets qui les composent.

Le problème a été posé par professeur Aïvazian.

Il faut ajouter qu'on examine habituellement les problèmes de l'analyse statistique des partitions dans l'espace $\mathfrak{M}_1$, $\mathfrak{M}_3$. Nous examinons principalement le cas de l'espace $\mathfrak{M}_2$. Nous ne pouvons pas citer la recherche dans une direction pareille.

I. La métrique. Chaque partition R se présente comme une matrice associée $\{r_{ij}\}_{i,j=1}^{mn}$:

a) Pour $\mathfrak{M}_1$ $(m=k)$:

$$r_{ij}=\begin{cases} 1, \text{ si l'objet } j \text{ se trouve dans la} \\ \quad \text{classe numérotée par } i; \\ 0 \text{ au contraire.} \end{cases}$$

* Eesti Teaduste Akadeemia Majanduse Instituut (Institut d'Economie de l'Académie des Sciences d'Estonie). Estonia pst. 7, Tallinn EE0105. Estonia.

b) Pour $\mathfrak{M}_2$ $(m=n)$:

$$r_{ij}=\begin{cases}1, & \text{si l'objet } i \text{ se trouve dans la}\\ & \text{classe où se trouve l'objet } j;\\ 0 & \text{au contraire.}\end{cases}$$

c) Pour $\mathfrak{M}_3$ $(m=n)$:

$$r_{ij}=\begin{cases}1, & \text{si le numéro de la classe, dans laquelle se}\\ & \text{trouve l'objet } j \text{ n'est pas moins que le nu-}\\ & \text{méro de la classe dans laquelle se trouve}\\ & \text{l'objet } i;\\ 0 & \text{au contraire.}\end{cases}$$

Nous examinons les distances suivantes:

$$d_1(R^1, R^2)=\frac{1}{2}\sum_{u=1}^{k}\sum_{v=1}^{n}|r^1_{uv}-r^2_{uv}| \quad \text{pour } \mathfrak{M}_1,$$

$$d_2(R^1, R^2)=\sum_{u=1}^{n}\sum_{v=1}^{n}|r^1_{uv}-r^2_{uv}| \quad \text{pour } \mathfrak{M}_2,$$

$$d_3(R^1, R^2)=\sum_{u=1}^{n}\sum_{v=1}^{n}|r^1_{uv}-r^2_{uv}| \quad \text{pour } \mathfrak{M}_3,$$

$$d_4(R^1, R^2)=\sum_{u=1}^{n}\left(\sum_{v=1}^{n}(r^1_{uv}-1)r^2_{uv}\right)\left(\sum_{l=1}^{n}(r^2_{ul}-1)r^1_{ul}\right) \quad \text{pour } \mathfrak{M}_2.$$

Il est aisé de voir qu'on peut calculer ces distances aussi à l'aide des tableaux de contingence. Il en résulte que nous devons connaître uniquement les nombres $\{n_{ij}\}_{i,j=1}^{k}$ où n_{ij} est le nombre des objets qui dans une partition se trouvent dans la classe i et simultanément, dans la deuxième partition se trouvent dans la classe j.

Ci-après nous argumenterons le choix des distances et la nécessité de la recherche des problèmes posés.

II. L'analyse statistique des données non-numériques (qualitatives). Comme nous le savons, l'utilisation des méthodes statistiques dans les domaines non-traditionnels (comme par example les sciences humaines, les estimations qualitatives des experts etc.) s'élargit actuellement de plus en plus; de sorte qu'il faut souvent poser les problèmes d'une manière non-traditionelle. Cela concerne en particulier l'analyse des données qualitatives, qui se rencontrent fréquemment en médecine, en économie, dans les sciences humaines. Les méthodes de l'analyse des données catégorielles (c.-à-d. l'analyse des tableaux de contingence) ne sont pas toujours utilisables dans de tels cas mais si l'utilisation des données catégorielles est possible, il nous reste parfois le doute que les résultats ne soient pas exacts. Le contenu statistique de tels problèmes peut en même temps être clair et bien s'exprimer à l'aide d'une terminologie différente. Il faut encore noter que les statisticiens sont habitués à exprimer les données à l'aide des nombres, même si cela ne correspond pas complètement au contenu du problème qu'on examine, c'est-à-dire si les données sont par leur nature indescriptibles à l'aide des nombres (par exemple le niveau du talent n'est pas quantifiable). Cela concerne particulièrement les données mesurées par une échelle nominale (les professions par exemple). La recherche de telles données est insuffisante. Nous voudrions avoir pour l'analyse des données non-numeriques une théorie aussi souple que celle dont nous disposons pour l'analyse des données numeriques.

Exemple. Supposons que les médecins qualifiés partitionnent un nombre suffisamment grand de malades d'après leur état de santé. Supposons que les progrès en médecine aient obligé les médecins à revoir le système du diagnostic en usage. Pour éviter les erreurs dans la classification future des maladies, les médecins sont priés de séparer leurs malades en se basant seulement sur sa propre expérience et les nouvelles connaissances en médecine. Il est claire que les médecins ne donnent pas des partitions identiques. On peut penser qu'en observant les partitions de cet exemple dans l'espace $\mathfrak{M}_2$, nous faisons moins d'erreurs dans les résultats, qu'en les observant dans l'espace $\mathfrak{M}_1$.

Au cours de l'analyse il nous reste à résoudre par exemple les problèmes suivants:

1) Tester la validité de l'hypothèse H_0, c'est-à-dire vérifier si les distinctions entre les partitions obtenues ne sont pas signifiantes. Cela veut dire, qu'il faut contrôler qu'il n'existe en effet aucune nouvelle classification des maladies (ou bien les malades ont tel ou tel état de santé ou bien les médecins ne sont pas capables de reconnaître les différences entre les maladies à la base de leur experience).

2) Il faut tirer au clair s'il existe parmi les médecins des groupes dont les opinions se distinguent profondément (les différences entre les opinions à l'intérieur d'un même groupe sont insignifiantes).

3) Il faut trouver une estimation pour la partition «juste» («vraie») S^0.

4) Si c'est possible, il faut trouver un ensemble de confiance pour la partition S^0.

5) Nous devons donner des noms aux classes, c'est-à-dire nous devons les caractériser (cela veut dire qu'il faut élaborer un système de diagnostic tout à fait nouveau).

Nous pouvons formuler les quatre premiers problèmes à l'aide des notions que nous utilisons dans cet exposé. Nous ne donnons pas ici d'analyse plys détaillée. Pour résoudre le cinquième problème, il faut avoir plus d'information que les partitions seulement. Les résultats des quatre premiers problèmes sont aussi les données de départ (initiales) pour résoudre le problème numéro cinq.

Quelques mots sur les distances. Nous croyons que malgré la grande quantité de distances (de coefficients de ressemblance etc.) en usage [1] utilisées pour l'analyse des classifications, le problème du choix de la distance est le moins examiné. Nous insistons sur le fait que les distances, qu'on utilise pour l'analyse des données, doivent toujours correspondre à la nature de l'espace, qu'on observe. Mais on ne fait pas toujours attention à cette exigence. Expliquons cette pensée. En examinant l'espace $\mathfrak{M}_2$, nous croyons, qu'en changeant les numéros des colonnes et (ou) des lignes dans le tableau de contingence (d'une manière arbitraire), la distance ne change pas. Dans l'espace $\mathfrak{M}_1$, la distance ne doit pas changer si nous changeons simultanément (de la même manière) les numéros des lignes et des colonnes.

Mais dans l'espace $\mathfrak{M}_3$, généralement parlant, la distance doit varier en cas de changements de toutes genres de numéros des lignes et des colonnes.

Aussi examine-t-on les distances $d_2(R^1, R^2)$ et $d_3(R^1, R^2)$ déjà définées dans les espaces $\mathfrak{M}_2$ et $\mathfrak{M}_3$. Ces distances sont argumentées car l'unicité de ces distances pour un système d'axiomes raisonnables est démontrée [2, 3]. Pour comparer les résultats, nous examinons aussi les distances $d_1(R^1, R^2)$ et $d_4(R^1, R^2)$. L'unicité de ces distances pour les autres systemes d'axiomes est aussi démontrée [4—6].

Il faut noter, que les mathématiciens s'intéressent à l'analyse des distances [7—9], car la recherche des espaces $\{\mathfrak{M}_i\}_{i=1}^3$ n'épuise pas tous les besoins pratiques.

Nous pouvons, par exemple, ne pas limiter le nombre des classes, les classes peuvent être conjointes; on peut examiner les classes composées par un nombre limité d'objets (c'est-à-dire les classes qui ont un volume limité). En ce dernier cas de même, le nombre de classes ne peut pas être borné car le nombre d'objets augmentant, le nombre des classes augmente aussi. Il faut parfois observer les partitions dont les classes sont rangées partiellement etc.

Nous ne savons pas dans quel espace il faut analyser les données obtenues à condition qu'avant la classification les objets-étalons aient été données comme exemples des classes.

La dernière question qui nous reste à aborder est la suivante. Ce que nous ne savons pas, ce sont les nombres $\{p_i\}_{i=1}^{N(n,k)}$, qui se manifestent en effet. Pour faire des modèles, il faut les connaître. Les recherches nécessaires exigent beaucoup d'observations et de calculs. Pour le moment, nous ne pouvons citer aucune recherche dans une direction pareille. Nous n'avons fait que quelques calculs pour estimer à l'oeil si notre supposition était vraisemblable ou non.

L'unique supposition qui semble naturelle, est la suivante:

$$P(R=S^i) \geqslant P(R=S^j), \quad \text{si} \quad d(S^i, S^0) \leqslant d(S^j, S^0),$$

où S_0 est la partition «juste» («vraie») [10, 11].

Nous examinons essentiellement l'hypothèse H_1:

$$P(R=S^i) = f_\theta(d(S^i, S^0)), \quad i=1, 2, \ldots \ldots, N(n,k);$$

$f_\theta(x) = C_{S^0\theta}^{nk} e^{-\theta x}$ ($C_{S^0j}^{nk}$ est le facteur de pondération) mais cela ne restreint pas essentiellement la généralité car tous les résultats obtenus pour le cas où l'hypothèse H_1 est vraie se basent sur les résultats correspondants qu'on a obtenu en supposant que l'hypothèse H_0 ait lieu. On peut aussi appliquer ces résultats obtenus pour l'examen d'autres hypothèses. Il est manifesté que pour $\theta=0$ H_1 donne H_0.

III. Les résultats. Réduisons notre problème. Nous cherchons la distribution des distances $d(R^1, R^2)$ et $d(R^1, S^\Phi)$, où R^1, R^2 sont deux variables aléatoires, quantitatives, indépendantes[1] (S^Φ est n'importe quelle partition fixée), en supposant que R^1 et R^2 soient distribués conformément à l'hypothèse H (la variable peut être remplacée par certaine fonction de ces variables, nous l'examinerons en détail peu après). Il est clair que pour avoir des critères pour le test d'hypothèse à l'aide des distances, il faut connaître les distributions de ces distances.

Nous nous bornons à l'examen de ce problème parce que nous ne pouvons rien dire à propos de propriétés des tests, que l'on peut élaborer actuellement sur la base de ces distributions. Notre but était d'avoir au moins quelques résultats nous permettant de tester les hypothèses les plus simples. Nous avons commencé par l'examen des moments de ces distances.

Pour formuler les résultats, supposons que des variables $\{R^u\}_{u=1}^m$ et S sont indépendantes (au sens statistique) et que $\{R^u\}_{u=1}^m$ soient distribués suivant l'hypothèse H_0, S par H_1. La «vraie» partition S^0 peut coïncider ou non avec la partition S^Φ fixée. U est la partition à une classe. Considérons $d_p(R^u, R^v)$, $d_p(R^u, S^\Phi)$, $d_p(S, S^0)$, $d_p(S, S^\Phi)$, $d_2(U, R_m)$, où

[1] La variable discrète aléatoire est donnée si ses valeurs et les probabilités de chaque valeur sont données.

$R_m = \bigcap_{u=1}^{m} R^u$ est la variable dont la distribution est induite par les distributions des variables $\{R^u\}_{u=1}^m$ (en tenant compte de l'indépendance)[2]. Il est clair qu'en ce cas le nombre des classes n'est plus que k^m.

La distribution exacte n'est calculée que pour $d_2(R^u, R^v)$ ($n \leqslant 42$; $k=2$), $d_3(R^u, R^v)$ ($n \leqslant 47$; $k=2$), $d_4(R^u, R^v)$ ($n \leqslant 25$; $k=2$) (précisément: pour certaines transformations linéaires des distances). Le calcul de la distribution $d_2(R^u, R^v)$ ne présente aucune difficulté. Pour le calcul des distributions exactes, les difficultés sont liées au temps (du calculateur) seulement. Pour $d_2(R^u, R^v)$ ($k=2$) il suffit des valeurs $n \leqslant 42$. Ce sont les moments, que nous examinons pour les autres cas. Quant aux moments $d_2(S, S^\Phi)$, il est possible (pour $k=2$) de les exprimer par les moments $d_2(S, S^0)$. Plus précisément: nous examinons

$$E\{a_p(n,k)\, d_p(R^u, R^v) + b_p(n,k)\}^r, \tag{1}$$

$$E\{c_p(n,k,S^\Phi)\, d_p(R^u, S^\Phi) + l_p(n,k,S^\Phi)\}^r, \tag{2}$$

$$E\{g_p(n,k,S^0)\, d_p(S, S^0) + h_p(n,k,S^0)\}^r, \tag{3}$$

$$E\{s(n,k,m)\, d_2(U, \bigcap_{u=1}^{m} R^u) + q(n,k,m)\}^r \tag{4}$$

où $a_p(n,k)$, $b_p(n,k)$, ..., $q(n,k,m)$ sont certains facteurs.

Si les limites (pour $n \to \infty$) des moments (1)—(4) pour les facteurs concrets sont calculées, on pourra estimer les limites des transformations linéaires correspondantes des distances. Nous avons fait cela pour tous les p ($p=1,2,3,4$; $k=2$). La distribution normale et la distribution du *chi deux* sont les distributions limites. Il faut noter, que pour $p=3,4$ les distributions limites $d_p(R^u, S^\Phi)$, $d_p(S, S^0)$ dependent de la limite S^0, S^Φ. Nous avons examiné le cas où une classe de ces partitions avait un nombre borné d'objets et le cas contraire. Les distributions limites sont différentes. Avant de passer au cas, où le nombre de classes est arbitraire (fixée), nous donnons comme exemple deux propositions. Nous considérons l'espace $\mathfrak{M}_2$.

Proposition 1. *Soit* $z^n_{R^u S^\Phi} = \frac{1}{n}(n^2 - \delta - 2d_2(R^u, S^\Phi))$ (*où* $\delta=0$ *pour n pair et* $\delta=1$ *pour n impair*). *Les limites des moments* $z^n_{R^u S^\Phi}$ *pour* $n \to \infty$ *sont*

$$\lim_{n\to\infty} E(z^n_{R^u S^\Phi})^r = (2r-1)!!$$

Il en résulte qu'en ce cas la distribution limite est χ^2 (1).

Examinons l'espace $\mathfrak{M}_3$.

Proposition 2. *Soient* $\hat{y}^n_{R^u S^\Phi} = \frac{1}{2n}(4d_3(R^u, S^\Phi) - n^2 - d_3(U, S^\Phi))$ *et m le nombre borné des objets dans une classe de la partition* S^Φ. *Nous avons:*

$$\lim_{n\to\infty} E(\hat{y}^n_{R^u S^\Phi})^r = \frac{d^r}{dx^r}\left\{\frac{1}{\sqrt{1+x}}\left(\frac{e^x+e^{-x}}{2}\right)^m\right\}_{x=0}.$$

La démonstration de ces résultats n'est pas compliquée. Il faut faire quelques transformations.

---

[2] Les valeurs R_m sont les intersections ensemblistes des valeurs des variables $\{R^u\}_{u=1}^m$; la probabilité de chaque valeur est le produit des probabilités des valeurs correspondantes de $\{R^u\}_{u=1}^m$.

IV. Nombre arbitraire de classes. Les premiers moments sont calculés. Nous examinons de nouveau l'espace $\mathfrak{M}_2$.

Théorème 1. *Soient* $\varphi^{nk}_{R^u R^v} = (k-1)n - \frac{k^2}{2n} d_1(R^u, R^v)$, ζ et η *deux variables aléatoires dont les distributions sont respectivement* $\chi^2((k-1)^2)$ *et* $\chi^2(2(k-1))$. *Les limites des moments* $\varphi^{nk}_{R^u R^v}$ *sont:*

$$\lim_{n\to\infty} E(\varphi^{nk}_{R^u R^v})^r = E\left(\zeta - \frac{1}{2}(k-2)\eta\right)^r$$

si, comme auparavant, R^u, R^v *sont indépendantes et distribuées suivant* H_0.

Théorème 2. *Soit* $x^{nk}_m = (k^m - 1) - \frac{k^m}{n} d_2(U, \bigcap_{u=1}^{m} R^u)$. *Pour* $n \to \infty$ *la variable* x^{nk}_m *converge en distribution vers une variable aléatoire, de loi* $\chi^2(k^m - 1)$.

Pour l'espace $\mathfrak{M}_3$ on a les théorèmes suivants.

Théorème 3. *Soit* $y^{nk}_{R^u R^v} = \frac{k^2}{(k-1)\sqrt{n^3}} d_3(R^u, R^v) - \frac{\sqrt{n}}{2}(k+1)$. *Pour* $n \to \infty$ *la variable* $y^{nk}_{R^u R^v}$ *converge en distribution vers une variable aléatoire normalement distribuée (centrée réduite) si* R^u, R^v *sont indépendantes et distribuées suivant* H_0.

Introduisons les notations suivantes: A_i — le nombre d'objets dans la classe i de la partitions S^Φ,

$$B^n_{S^\Phi} = \frac{1}{n^3}\left(n^3 - \sum_{i=1}^{k} A_i\right), \qquad \frac{A_i}{n} \xrightarrow[n\to\infty]{} a_i.$$

Théorème 4. *Soit* $Y^{nk}_{R^u S^\Phi} = \frac{3}{2\sqrt{B^n_{S^\Phi}\ (k^2-1)n^3}} [2(kd_3(R^u, S^\Phi) - d_3(U, S^\Phi) - n^2(k-1))]$. *Si* $a_i \neq 0$, $a_i \neq 1$ *(pour tout* $i = 1, 2, \ldots, k$*), la variable* $Y^{nk}_{R^u S^\Phi}$ *converge en distribution vers une variable aléatoire normale (centrée réduite).*

Soit

$$P(S = S^i) = \frac{e^{-\frac{\theta d_3(S^i, S^0)}{n^{3/2}}}}{\sum_{v=1}^{N_3(n,k)} e^{-\frac{\theta d_3(S^v, S^0)}{n^{3/2}}}}.$$

Théorème 5. *Pour* $n \to \infty$ $(a_i \neq 0,\ a_i \neq 1)$ $Y^{nk}_{SS^0}$ *converge en loi vers une variable aléatoire normale (d'espérance mathématique* $-\theta \frac{\sqrt{k^2-1}}{3k} \lim_{n\to\infty} \sqrt{B^n_{S^0}}$ *et de variance 1).*

Pour comparer les résultats, il faut noter que si on examine le cas de l'espace $\mathfrak{M}_1$ et

$$P(S=S^i)=\frac{e^{-\frac{\theta d_1(S^i,S^0)}{n^{1/2}}}}{\sum_{v=1}^{N_1(n,k)} e^{-\frac{\theta d_1(S^v,S^0)}{n^{1/2}}}}$$

la variable aléatoire $\frac{kd_1(S^0,S)-(k-1)n}{\sqrt{n(k-1)}}$ pour $n\to\infty$ converge en distribution vers une variable aléatoire normale (d'espérance mathématique $-\theta\frac{\sqrt{k-1}}{k}$ et de variance 1).

Ces résultats sont obtenus essentiellement à partir des résultats de l'Analyse Combinatoire (les problèmes d'énumeration de là théorie des graphes).

Les démonstrations sont compliquées [12]. C'est pourquoi pour le moment nous les avons donné entièrement dans la thèse [13] seulement.

Il semble que pour développer la théorie statistique des données non-numériques, il faut résoudre non seulement de nombreux problèmes théoriques (par exemple obtenir pour $\mathfrak{M}_2$ des théorèmes semblables aux théorèmes 4, 5; examiner les propriétés de tests divers etc.) mais il faut résoudre aussi des problèmes de calcul et des problèmes pratiques.

Ainsi il est utile de connaître les distributions exactes des distances pour de petites valeurs de n, mais comme le calcul à l'aide de l'algorithme du dénombrement simple prend beaucoup de temps, il faut trouver des algorithmes plus rapides.

Il est utile aussi de parfaire la connaissance sur les distributions des erreurs que font les experts (la fonction $f_\theta(x)$). Il semble, que la solution de ce dernier problème exige beaucoup de temps.

BIBLIOGRAPHIE

1. *Goodman, L. A., Kruskal, W. H.* Measures of association for cross classifications: I—IV. — J. Amer. Stat. Assoc., 1954, **49**, 723—764; 1959, **54**, 123—163; 1963, **58**, 310—363.
2. *Kemeny, J., Snell, L.* Mathematical Models in the Social Sciences. Englewood Cliffs (N. J.). Prentice-Hall, 1962.
3. *Миркин Б. Г., Черный Л. Б.* Об измерении близости между разбиениями конечного множества объектов. — Автоматика и телемеханика, 1970, 5, 120—127.
4. *Плоткин А. А.* Мера независимости классификаций. — Автоматика и телемеханика, 1980, 4, 97—104.
5. *Тюрин Ю. Н.* Непараметрические методы статистики. — Новое в жизни, науке, технике. Серия «Математика, кибернетика», 1978, 4.
6. *Куликов С. М.* Структурные меры близости в пространстве классификаций и разбиений. — Dans: Прикладная статистика. Ученые записки по статистике, 45. Москва, Наука, 1983, 282—284.
7. *Орлов А. И.* Устойчивость в социально-экономических моделях. Москва, Наука, 1979.
8. *Литвак Б. Г.* Экспертная информация. Методы получения и анализа. Москва, Радио и связь, 1982.
9. *Раушенбах Г. В.* Об измерении близости между множествами в задачах кластерного анализа. — Dans: Статистика. Вероятность. Экономика. Ученые записки по статистике, **49**. Москва, Наука, 1985, 388—392.

10. *Тюрин Ю. Н.* О математических задачах в экспертных оценках. — Dans.: Экспертные оценки. Вопросы кибернетики. Москва, ИПУ, 1982, 7—16.
11. *Пинкова Я.* Вероятностное распределение в задачах статистического ранжирования. — Dans: Экспертные оценки. Вопросы кибернетики. Москва, ИПУ, 1982, 34—52.
12. *Маамяги А. В.* Некоторые задачи статистического анализа классификаций. Таллинн, АН ЭССР, 1982.
13. *Маамяги А. В.* Задачи статистического анализа разбиений. Москва, МГУ, 1985. (Автореферат.)

Présentée par J. Engelbrecht

Remise le 24 avril 1991,
acceptée le 31 mai 1991

Aasa MAAMÄGI

TÜKELDUSTE STATISTILINE ANALÜÜS

On vaadeldud kvalitatiivsete muutujate statistilisel töötlemisel tekkivaid matemaatilisi probleeme. Analüüsi põhiobjektiks on kauguste jaotusfunktsioon järjestatud ja järjestamata tükelduste ruumis (võrdluseks on analüüsitud ka eelnevalt kirjeldatud klasside juhtu). On eeldatud, et

$$P(R=S^j)=p_j,$$

kus $P(R=S^j)$ tähistab tükelduse S^j ilmumise tõenäosust, $\{p_j\}_{j=1}^{N}$ on fikseeritud ja N on tükelduste arv vaadeldavas ruumis. Toodud on kvalitatiivsete muutujate analüüsil lahendust vajavate statistiliste ülesannete püstitus. Töös on praktiline näide. Lõpus on saadud tulemuste täpne formuleering. Põhjalikult on analüüsitud probleeme, mis alles vajavad lahendamist.

Аза МААМЯГИ

СТАТИСТИЧЕСКИЙ АНАЛИЗ РАЗБИЕНИЙ

Рассматриваются проблемы статистической обработки качественных признаков. Основной объект исследования — распределение исследуемого расстояния (в основном — расстояния Хемминга) в пространстве упорядоченных и неупорядоченных разбиений (для сравнения рассматриваются и заранее описанные классы) при справедливости гипотезы H:

$$P(R=S^j)=p_j,$$

где $P(R=S^j)$ — вероятность выпадения в эксперименте разбиения S^j, $\{p_j\}_{j=1}^{N}$ — некоторые заранее фиксированные числа, N — число возможных разбиений в рассматриваемом пространстве. Приводятся постановки статистических задач, возникающих при анализе качественных признаков. Дан пример такой задачи. Сформулированы полученные результаты и очерчен круг проблем, требующих решения.