DISZKRÉT VALÓSZÍNŰSÉG-ELOSZLÁSOK KEVERÉKÉNEK FELBONTÁSA ÖSSZETEVŐIRE

MEDGYESSY PÁL

Bevezetés

Korábbi dolgozatainkban ([1], [2]) folytonos eloszlás-, illetve sűrűségfüggvények konstans súlyokkal vett keverékének felbontásával foglalkoztunk. Ebben a cikkben diszkrét valószínűség-eloszlásokra végzünk hasonló vizsgálatokat. Tudomásunk szerint ezzel a problémával eddig nem foglalkoztak, bár fontos gyakorlati alkalmazásai vannak.

Problémánk a következőképp fogalmazható meg: legyen $\psi_\nu(x)$ $(-\infty < x < \infty;\ \nu = a_1, a_2, \ldots)$ az x paramétertől függő diszkrét eloszlás, amelyet az $a_1, a_2, \ldots$ pontokban értelmeztünk; $\psi_\nu(x)$ az x-nek folytonos függvénye. Legyen továbbá $\Phi(x)$ egy eloszlásfüggvény. Bevezetjük a következő definíciót:

A $\psi_\nu(x)$ *komponensek* $\Phi(x)$ súlyfüggvénnyel való *keverékének* az

$$f_\nu = \int_{-\infty}^{\infty} \psi_\nu(x)\, d\Phi(x) \qquad (\nu = a_1, a_2, \ldots) \tag{1}$$

Stieltjes-integrállal értelmezett diszkrét eloszlást nevezzük.

A definíció az [1] dolgozat bevezetésében közölt, folytonos sűrűségfüggvények keverékére vonatkozó definíció értelemszerű kiterjesztése. (A diszkrét eloszlás eloszlásfüggvényét tartalmazó definíció hasonlóan adható meg, semmi előnnyel sem jár azonban, ha ebből indulunk ki s ezért mellőzzük.)

Az (1)-ben szereplő f_ν keverékeloszlás *felbontásán* általánosságban $\psi_\nu(x)$ vagy $\Phi(x)$ (esetleg mindkettő) meghatározását értjük az (1) összefüggés és az f_ν értékek ismerete alapján. Esetenként még azt is megengedjük, hogy a keresett függvények *típusát* is ismerjük, csak paramétereit nem.

Ha $\Phi(x)$ tiszta lépcsős függvény, (1) nyilván végtelen sorba megy át. Alkalmasan választott $\psi_\nu(x)$ és tiszta lépcsős $\Phi(x)$ függvényekkel elérhetjük, hogy f_ν az

$$\begin{cases} f_\nu = \sum_{k=1}^{N} A_k \binom{M_k}{\nu} p_k^\nu (1-p_k)^{M_k - \nu} & (\nu = 0, 1, \ldots, \max M_k) \\ \text{illetőleg} \\ f_\nu = \sum_{k=1}^{N} A_k \cdot e^{-\lambda_k} \frac{\lambda_k^\nu}{\nu!} & (\nu = 0, 1, \ldots) \end{cases} \tag{2}$$

alakot öltse $\left(A_k > 0, \sum_{k=1}^{N} A_k = 1, M_k > 0, 0 < p_k < 1, \lambda_k > 0 \text{ valós paraméterek}\right)$. A (2) alatti kifejezések közül az elsőt különböző M_k, p_k paraméterű binomiális eloszlások konstans A_k súlyokkal vett keverékének, a másodikat pedig λ_k paraméterű Poisson-eloszlások A_k súlyokkal vett keverékének nevezzük.

A dolgozatunkban tárgyalt problémák erre a két speciális keverékre vonatkoznak.

Az (1) alatti alakhoz viszonyítva látjuk, hogy a keverékek komponenseinek típusát ismerjük, de paramétereit nem, a súlyfüggvényről pedig csak azt tudjuk, hogy tiszta lépcsős függvény.

A következőkben ezen keverékek felbontásával fogunk foglalkozni, ami itt az ismeretlen A_k, M_k, p_k, λ_k paraméterek meghatározásával ekvivalens.

Az A_k mennyiségek valószínűségszámítási interpretációjához $\left(\sum_{k=1}^{N} A_k = 1\right)$ nem szükséges ragaszkodnunk s ezért a (2) alatti alakot egyszerűen mint binomiális, illetve Poisson-eloszlások szuperpozícióját kezeljük.

Binomiális eloszlások keveréke felbontásának gyakorlati szerepével és jelentőségével egy másik dolgozatban [3] már foglalkoztunk.

1. §.

Az

$$(3) \qquad f_\nu = \sum_{k=1}^{N} A_k \binom{M_k}{\nu} p_k^\nu (1 - p_k)^{M_k - \nu} \qquad (\nu = 0, 1, \ldots, \max M_k$$

($A_k > 0$, $M_k > 0$ egész, $0 < p_k < 1$, azonos M_k, p_k pár nincs) keverék felbontásánál, — ami itt az A_k, M_k, p_k értékek meghatározásával ekvivalens — az [1] dolgozatban említett »szóráscsökkentési« elvet alkalmazzuk: legyenek egyelőre a p_k-k mind különbözőek, s tegyük fel, hogy a (3)-ban szereplő f_ν-értékek ismeretében alkalmasan választott $\mu > 1$ valós paraméterrel elő tudunk állítani egy

$$(4) \qquad F_\nu(\mu) = \sum_{k=\nu}^{N} A_k \binom{M_k}{\nu} (\mu p_k)^\nu (1 - \mu p_k)^{M_k - \nu}$$

($\nu = 0, 1, \ldots, \max M_k$) eloszlást; ez (3)-tól csak abban különbözik, hogy benne p_k helyett μp_k szerepel ($0 < \mu p_k \leqq 1$).

A nem-negatív monoton növekvő $\mu_1, \mu_2, \ldots, \mu_i, \ldots$ értékekkel állítsuk elő sorban a megfelelő $F_\nu(\mu_i)$-ket. Tegyük fel, hogy valamelyik μ_i épp egyenlő $\max p_k = p_\tau$ reciprokával, vagyis $\mu_i = \frac{1}{p_\tau}$. Ennél a μ_i-nél a (4) a τ indexű komponens egyetlen ponttá elfajult binomiális eloszlás lesz, melynek 0-adik, 1-ső, ..., $M_{\tau-1}$-edik eleme zérus, az M_τ-adik pedig A_τ.

Ez az egyetlen pont nyilván ki fog ugrani az eloszlás többi pontja fölé és ha szomszédságában már igen kicsik az eloszlás tagjai, ordinátája jó közelítéssel megadja A_τ-t, abszcisszája pedig M_τ-t. p_τ nyilván az alkalmazott μ reciproka lesz.

Ezen körülmények mellett tehát egy komponens adatait meghatározhatjuk. Az M_τ-ban közel A_τ magasban kiugró pontot elhagyva és helyette a szomszédos adatokból interpolált új pontot vezetve be, kis hibával olyan keverékre jutunk, amelynek a kiindulónál eggyel kevesebb komponense van, és amelyben p_k helyett p_k/p_τ paraméterek szerepelnek. Ezen az eljárást megismételhetjük és újabb komponenst választhatunk le, és így tovább. A felbontással nyert adatok természetesen a komponensek számát, N-et is megadják.

A gyakorlatban legtöbbször egyetlen lépés (sőt, már egyetlen μ-értékkel való kísérlet) is elég, mert kimutatható, hogy az eljárás során az egyes komponensek maximumhelyeinek távolsága általában megnő, szórásnégyzeteik pedig általában csökkennek. Ennek a részletes tárgyalására nem térünk itt ki, mert előbbi okoskodásunk értelmében másodrendű fontosságúak[1]), csak azt jegyezzük meg, hogy szomszédjaitól erősen különvált komponenst egyetlen binomiális eloszlásként kezelhetünk, ekkor pedig paramétereit közelítőleg megállapíthatjuk. A komponensek száma legtöbbször egyetlen μ-vel való kísérletből is kiderül.

Állapodjunk meg abban, hogy a felbontást gyakorlatilag elintéztük, ha az f_ν értékek ismeretében általunk megadott μ-vel elő tudjuk állítani az $F_\nu(\mu)$ mennyiségeket.

Ezt az előállítást megadja a következő

Tétel: Ha az f_ν diszkrét eloszlás binomiális eloszlások

$$f_\nu = \sum_{k=1}^{N} A_k \binom{M_k}{\nu} p_k^\nu (1-p_k)^{M_k-\nu} \qquad (\nu = 0, 1, \ldots, \max M_k)$$

alakú keveréke, a komponensek szétválasztására alkalmas új

$$F_\nu(\mu) = \sum_{k=1}^{N} A_k \binom{M_k}{\nu} (\mu p_k)^\nu (1-\mu p_k)^{M_k-\nu}$$

keverékeloszlást az

$$F_\nu(\mu) = \mu^\nu \sum_{\varrho=\nu}^{\max M_k} \binom{\varrho}{\nu} (1-\mu)^{\varrho-\nu} f_\varrho \tag{5}$$

összefüggés adja meg.

[1]) A maximumhelyek távolságának változása annak az ismert ténynek az alapján vizsgálható, hogy egy M_k, μp_k paraméterű binomiális eloszlás maximuma kb. $M_k \mu p_k$-nál van; ilyenformán pl. két egymásra következő maximumhely távolsága $\mu(M_{k+1}\, p_{k+1} - M_k\, p_k)$, ami μ-vel együtt növekedik, hacsak μ tényezője $\neq 0$. (Analóg módon okoskodunk a szórásnégyzetek esetében.)

Bizonyítás: Felhasználva az

$$(1-\mu p_k)^{M_k-\nu} = [(1-\mu)\,p_k + (1-p_k)]^{M_k-\nu} =$$

$$= \sum_{\sigma=0}^{M_k-\nu} \binom{M_k-\nu}{\sigma} (1-\mu)^\sigma\, p_k{}^\sigma\, (1-p_k)^{M_k-\nu-\sigma}$$

azonosságot,

$$F_\nu(\mu) = \sum_{k=1}^{N} A_k \binom{M_k}{\nu} (\mu p_k)^\nu \sum_{\sigma=0}^{M_k-\nu} \binom{M_k-\nu}{\sigma} (1-\mu)^\sigma\, p_k{}^\sigma\, (1-p_k)^{M_k-\nu-\sigma}$$

Vezessük be a $\nu+\sigma=\varrho$ új összegezési változót; ekkor

$$F_\nu(\mu) = \mu^\nu \sum_{k=1}^{N} A_k \sum_{\varrho=\nu}^{M_k} \binom{M_k}{\nu} \binom{M_k-\nu}{\varrho-\nu} (1-\mu)^{\varrho-\nu}\, p_k^\varrho\, (1-p_k)^{M_k-\varrho}\,.$$

Használjuk fel az

$$\binom{M_k}{\nu}\binom{M_k-\nu}{\varrho-\nu} = \binom{M_k}{\varrho}\binom{\varrho}{\nu}$$

összefüggést, és vegyük figyelembe, hogy az $F_\nu(\mu)$ nem változik, ha M_k helyett max M_k-ig összegezünk. Ekkor azonban az összegezések felcserélhetők,

$$F_\nu(\mu) = \mu^\nu \sum_{\varrho=\nu}^{\max M_k} \binom{\varrho}{\nu} (1-\mu)^{\varrho-\nu} \sum_{k=1}^{N} A_k \binom{M_k}{\varrho} p_k^\varrho\, (1-p_k)^{M_k-\varrho}\,,$$

vagyis, (3) alapján

$$F_\nu(\mu) = \mu^\nu \sum_{\varrho=\nu}^{\max M_k} \binom{\varrho}{\nu} (1-\mu)^{\varrho-\nu} f_\nu$$

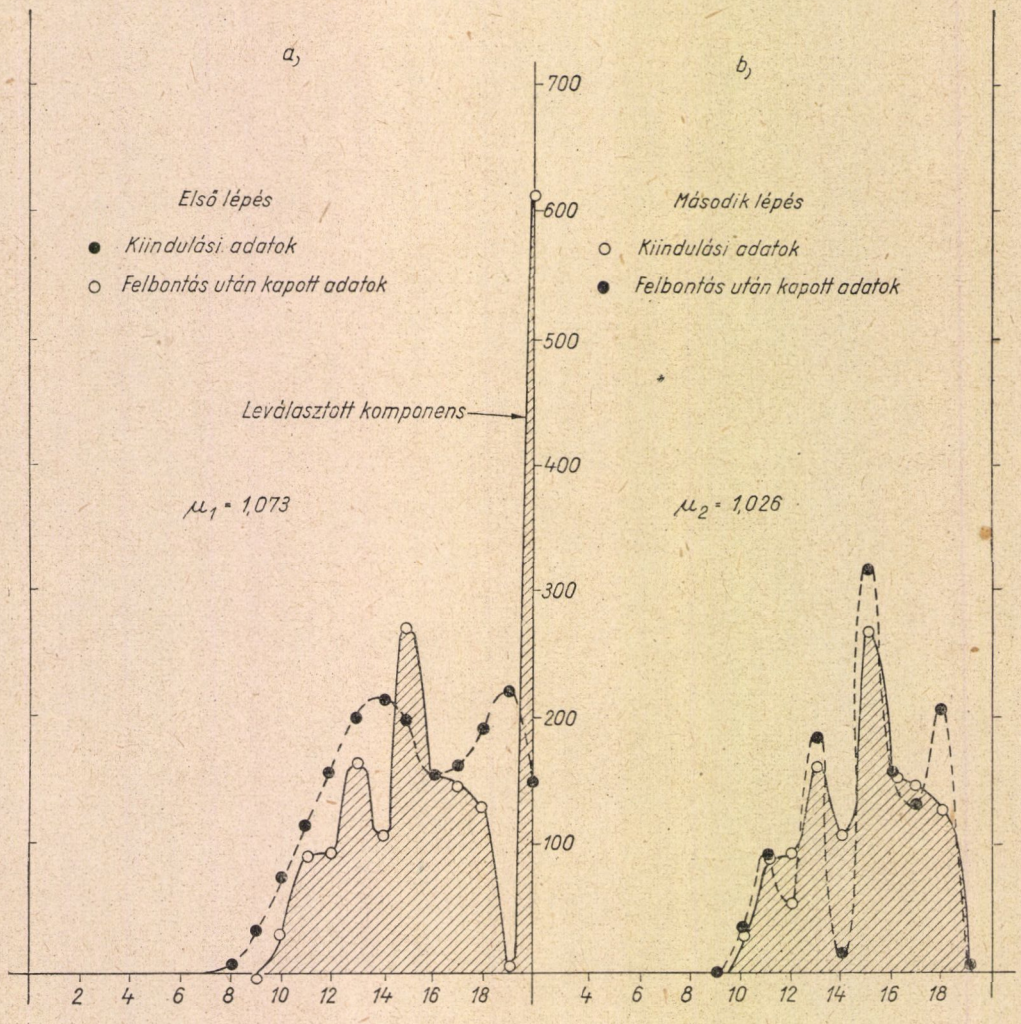
ahogy állítottuk.

Ezek után egyszerűen kínálkozik a keverék felbontásának gyakorlati módszere: Választunk valamilyen $\mu_1, \mu_2, \ldots$ sorozatot és a felbontandó binomiális keverék-eloszlás f_ν értékei segítségével (5) felhasználásával kiszámítjuk az $F_\nu(\mu_1)$, $F_\nu(\mu_2)$, ... eloszlásokat. (Ha max M_k értéke 10—20 körül van, ez géppel gyorsan megy.)

Vegyük észre, hogy $F_\nu(\mu)$ meghatározásához csak az $f_\nu, \ldots, f_{\max M_k}$ (a ν-nél nagyobb indexű) értékek szükségesek. Ha μ_i megközelíti az $1/\max p_k$ értéket, a vonatkozó komponens egyetlen pontként kiugrik a többi közül, ahogy azt fent már részleteztük. Ennek a ténynek a megállapítása az $F_\nu(\mu_1)$, $F_\nu(\mu_2)$, ... eloszlások ábrázolása segítségével könnyű. A kiugró komponens paramétereinek megállapítása ugyancsak a fentebb már tárgyalt módon történik. — A gyakorlatban az M_k értékek általában egyenlők.

(4) alakjából következik, hogy $\mu p_k > 1$ esetén az $F_\nu(\mu)$ komponensei közül egyesek negatív értékeket is felvesznek, vagyis várható, hogy maga $F_\nu(\mu)$ is helyenként negatív értékeket vesz fel: Ha tehát az eljárás során valamilyen μ értéknél már negatív tagokat is kapunk $F_\nu(\mu)$-ben, biztos, hogy túlléptük μ-vel az $1/\max p_k$ értéket. Ennek a fordítottja azonban már nem mondható ki, s ezért legbiztosabb eljárás több μ-vel végzett felbontások eredményeinek összehasonlítása.

Ha a p_k-k közt egyenlők is vannak, csak az lesz a különbség, hogy a szétválasztás valamelyik lépésénél egyszerre több pont »ugrik ki«. Mivel ez fenti okoskodásunkat nem befolyásolja és az eljárás menete is rögtön látható, ezért részletesen nem tárgyaljuk.



1. ábra

Binomiális eloszlások keverékének felbontása válik szükségessé a [3] dolgozatban (jelen kötet) tárgyalt fizikai-kémiai problémánál. Ilyen eredetű keveréket és felbontását mutatjuk be az 1. ábrán. *Áttekinthetőség kedvéért* folytonos görbét fektettünk át az adatokon. A közölt felbontás $\mu_1 = 1{,}073$, majd — a kiugró komponens leválasztása után nyert eloszlásnál, második lépésként — $\mu_2 = 1{,}026$-tal történt; jól láthatók a különvált komponensek (bizonyos különválás az eredeti felvételen is látszik). Vegyük észre, hogy nem két komponens van, mint azt a kiinduló eloszlásból gondolnánk, hanem öt! A keresett paraméterek a felbontás után kapott eloszlásból:

$$p_1 \approx 0,49 \quad p_2 \approx 0,59 \quad p_3 \approx 0,68 \quad p_4 \approx 0,82 \quad p_5 \approx 0,91$$

$$A_1 \approx 140 \quad A_2 \approx 220 \quad A_3 \approx 560 \quad A_4 \approx 280 \quad A_5 \approx 610,$$

amelyek jól egyeznek az itt történetesen gyakorlatból is ismert értékekkel.

Természetesen ajánlatos ellenőrzésként az így kiszámított paraméterekkel néhány eloszlás-pontot meghatározni, és egybevetni a kísérleti adatokkal.

Bemutatjuk még az

$$f_\nu = \binom{10}{\nu} 0,4^\nu (1-0,4)^{10-\nu} + 2\binom{10}{\nu} 0,5^\nu (1-0,5)^{10-\nu} \quad (\nu = 0, 1, \ldots, 10)$$

keverék felbontását is. Itt

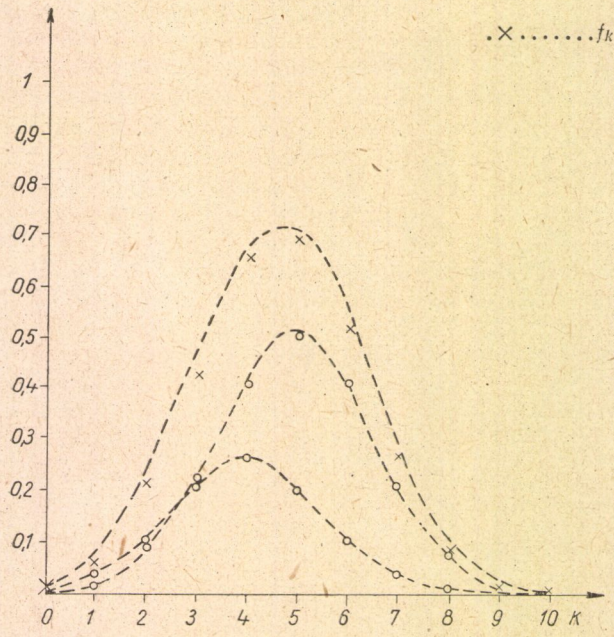
$$N = 2, \quad A_1 = 1, \quad A_2 = 2, M_1 = M_2 = 10, \quad p_1 = 0,4, \quad p_2 = 0,5 \cdot$$

A 2. a. ábra mutatja a két komponenst külön (o) és összegüket (×), a 2. b pedig összegüket (o) és a $\mu = \frac{10}{6}$-dal történt felbontás eredményét (×), vagyis az $F_\nu\left(\frac{10}{6}\right)$ értékeket (itt is Gauss-görbéket fektettünk át az adatokon, a szemléltetés céljára). Az f_ν keverék nem árul el semmit a komponensekről, az $F_\nu\left(\frac{10}{6}\right)$ értékek viszont már élesen különváló két eloszlást mutatnak, bár még nem járunk az extrémális esetnél.

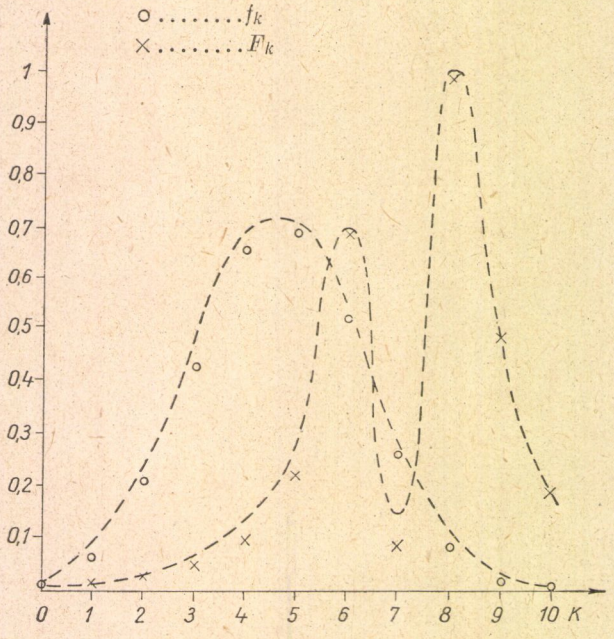
2. §

Az előző §-ban tárgyalt eljárás gyakorlati végrehajtásakor több nehézségre bukkanunk. Az egyik: a kísérleti adatok vagy 0, vagy $\max M_k = M$ szomszédságában (esetleg mindkét helyen is) sokszor hibahatáron belül kicsik, és ezért a kutatók zérusnak veszik őket, hiszen nincs alap más érték hozzárendelésére.

Ha csak a bal-, illetve jobbszélen ilyenek az adatok, a nehézség még áthidalható, mert az $F_\nu(\mu)$-k szempontjából csak a ν-nél *nagyobb* indexű f_ϱ-értékek fontosak (vesd össze: (5)), vagyis hibás *balszéli* adatok legfeljebb az első $F_\nu(\mu)$ értékeket rontják el, ami a felbontást alig zavarja, hibás *jobbszéli* adatoknál pedig egyszerűen tükrözzük az egész eloszlást a 0, ..., max M_k intervallum felező



2. a. ábra



2. b. ábra

merőlegese körül, mikoris a keverék lényegileg változatlan marad, a használhatatlan adatok pedig a balszélre kerülnek, s így lényegtelenné válnak. Kritikus tehát csak az az eset marad, amidőn mindkét szélen használhatatlanok az adatok. Számunkra ez azt jelenti, hogy valamilyen l-nél és j-nél az $f_0, f_1, \ldots, f_l$ és $f_M, f_{M-1}, \ldots, f_{M-j}$ értékeket is 0-nak kell vennünk, holott értékük véges, bár igen kicsiny. Ezzel azonban (5) alkalmazásakor igen nagy hibákat követhetünk el : pl. az $F_M(\mu) = \mu^M f_M$ értéket 0-nak kell tekintenünk, holott nagy M esetén $\mu > 1$ folytán μ^M nagy szám és így $F_M(\mu)$ kis f_M-nél is elég nagy marad. Egyenlő M_k értékeknél éppen ebben a pontban várhatnánk a kiugró eloszlás-értéket, az említett elhanyagolás azonban épp ezt (meg a szomszédos) adatokat teszi használhatatlanná.

Olyan eljárást kell tehát keresnünk, amely szélső adatokra *nem* támaszkodva szolgáltatja a felbontást vagy mindjárt az egyes adatokat.

Bebizonyítható, hogy a fentebb tárgyalt szétválasztási eljárás nem alkalmazható, ha nem használhatjuk fel az összes f_ν-értékeket. Közelfekvő gondolat megpróbálni az f_ν értékek olyan $\Phi_s(f_1, \ldots, f_M)$ $(s = 0, 1, \ldots)$ függvényeit előállítani, amelyeknek értékét alig befolyásolják a szélső f_ν-k értékei, amellett a $\Phi_s(f_1, \ldots, f_M)$ értékekből vagy következtetni tudunk az ismeretlen paraméterekre, vagy visszavezethetjük a problémát a 2. §-ban tárgyalt módszer alkalmazására. (Itt tulajdonképpen globálisan kezelnénk az f_ν-ket, ami a kísérleti adatok feldolgozásánál azért előnyös, mert a hibák kevésbé befolyásolják a végeredményt.) Végeredményben az f_ν-k sorozatát kellene áttranszformálnunk a $\Phi_s(f_1, \ldots, f_M)$ sorozatba egy $\Phi_s(f_1, \ldots, f_M) = \sum_{k=1}^{M} B_k(s) \cdot f_k$ típusú transzformáció segítségével, ahol azonban a $B_k(s)$ transzformáló sorozatot feltételeinknek megfelelően kell meghatároznunk. Sajnos, minden kikötésünknek eleget tevő $B_k(s)$ sorozatot eddig nem sikerült találnunk.

Ha tudjuk, hány komponenst várhatunk (a gyakorlatban sokszor ez az eset) és az összes M_k-k egyenlőek, sikerrel alkalmazható a következő, algebrai eljárás :

Tekintsük ismét a keverék (3) alatti alakját $(M_k = n)$:

$$f_\nu = \sum_{k=1}^{N} A_k \binom{n}{\nu} p_k^\nu (1 - p_k)^{n-\nu}, \quad (\nu = 0, 1, \ldots, n). \tag{6}$$

Ez felfogható magasabbfokú egyenletrendszernek az A_k és p_k-k értékek meghatározására, amely $2N \leq n$ esetén meg is oldható. A megoldást egy már régebben kidolgozott módszer[2]) általánosítása adja meg : Vezessük be az új D_ν, B_k, x_k paramétereket, amelyeket az

$$\frac{f_\nu}{\binom{n}{\nu}} = D_\nu \quad , \quad p_k = \frac{x_k}{1 + x_k} \quad , \quad B_k = \frac{A_k}{(1 + x_k)^n}$$

összefüggések definiálnak. Ekkor (6) így alakul :

$$D_\nu = \sum_{k=1}^{N} B_k \cdot x_k^\nu, \quad (\nu = 0, 1, \ldots, n), \tag{7}$$

[2]) Lényegében szerepel már *Gauss*-nál : [4]. Lásd még : [5], p. 275 — Alkalmazását illetőleg lásd : [6], főleg p. 713.

ahol a B_k, x_k mennyiségek az ismeretlenek. Mivel a B_k, x_k mennyiségekből az A_k és p_k értékek egyértelműen kiszámíthatók, elég a (7) egyenlettel foglalkoznunk. Olyan megoldási módszert keresünk, amely nem használja fel a szélső f_ν értékeket, illetve a megfelelő indexű D_ν értékeket.

Tekintsünk pozitív egész számokból álló olyan

$$\begin{gathered} \alpha_1 < \alpha_2 < \ldots < \alpha_{N+1} \\ \beta_1 < \beta_2 < \ldots < \beta_{N+1} \\ \gamma_1 < \gamma_2 < \ldots < \gamma_{N+1} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \tau_1 < \tau_2 < \ldots < \tau_{N+1} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{gathered}$$

rendszert, amelyre

$$\begin{gathered} \alpha_2 - \alpha_1 = \beta_2 - \beta_1 = \gamma_2 - \gamma_1 = \ldots = \tau_2 - \tau_1 = \ldots \\ \alpha_3 - \alpha_1 = \beta_3 - \beta_1 = \gamma_3 - \gamma_1 = \ldots = \tau_3 - \tau_1 = \ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{N+1} - \alpha_1 = \beta_{N+1} - \beta_1 = \gamma_{N+1} - \gamma_1 = \ldots = \tau_{N+1} - \tau_1 = \ldots, \end{gathered} \tag{8}$$

és tegyük fel, hogy

$$\alpha_{N+1} \leqq n, \quad \beta_{N+1} \leqq n, \quad \gamma_{N+1} \leqq n, \quad \ldots \quad , \quad \tau_{N+1} \leqq n, \quad \ldots . \tag{9}$$

Szemeljük ki a (7) rendszerből először a $\nu = \alpha_1, \alpha_2, \ldots$ indexű egyenleteket:

$$\begin{gathered} B_1 x_1^{\alpha_1} + \ldots + B_N x_N^{\alpha_1} = D_{\alpha_1} \\ B_1 x_1^{\alpha_2} + \ldots + B_N x_N^{\alpha_2} = D_{\alpha_2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ B_1 x_1^{\alpha_{N+1}} + \ldots + B_N x_N^{\alpha_{N+1}} = D_{\alpha_{N+1}} \end{gathered} \tag{10}$$

Szorozzuk ezeket sorban az egyelőre ismeretlen $c_1, c_2, \ldots, c_N$ értékkel, az utolsót 1-gyel s adjuk össze a jobb- és baloldalakat:

$$B_1 x_1^{\alpha_1}(c_1 + c_2 x_1^{\alpha_2-\alpha_1} + \ldots + c_N x_1^{\alpha_N-\alpha_1} + x_1^{\alpha_{N+1}-\alpha_1}) +$$

$$+ B_2 x_2^{\alpha_1}(c_1 + c_2 x_2^{\alpha_2-\alpha_1} + \ldots + c_N x_2^{\alpha_N-\alpha_1} + x_2^{\alpha_{N+1}-\alpha_1}) + \ldots$$

(11) $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$$\ldots + B_N x_N^{\alpha_1}(c_1 + c_2 x_N^{\alpha_2-\alpha_1} + \ldots + c_N x_N^{\alpha_N-\alpha_1} + x_N^{\alpha_{N+1}-\alpha_1}) =$$

$$= c_1 D_{\alpha_1} + c_2 D_{\alpha_2} + \ldots + c_N D_{\alpha_N} + D_{\alpha_{N+1}}$$

Válasszuk meg $c_1, \ldots, c_N$-t úgy, hogy az $x_1, x_2, \ldots, x_N$ értékek épp a

(12) $$c_1 + c_2 x^{\alpha_2-\alpha_1} + \ldots + c_N x_1^{\alpha_N-\alpha_1} + x^{\alpha_{N+1}-\alpha_1} = 0$$

egyenlet gyökei legyenek. Ekkor (11) baloldala zérus, vagyis a c_i értékek közt fenn kell állnia a

(13) $$D_{\alpha_1} c_1 + D_{\alpha_2} c_2 + \ldots + D_{\alpha_N} c_N = - D_{\alpha_{N+1}}$$

összefüggésnek.

Vegyük most a (7) rendszerből a $\nu = \beta_1, \beta_2, \ldots$ indexű egyenleteket, szorozzuk meg őket ismét az előbbi $c_1, c_2, \ldots$ értékekkel és adjuk össze:

$$B_1 \cdot x_1^{\beta_1}(c_1 + c_2 x_1^{\beta_2-\beta_1} + \ldots + c_N x_1^{\beta_N-\beta_1} + x_1^{\beta_{N+1}-\beta_1}) + \ldots$$

(14) $\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$$= c_1 D_{\beta_1} + c_2 D_{\beta_2} + \ldots + c_N D_{\beta_N} + D_{\beta_{N+1}}.$$

Mivel azonban (8) fennállásából indultunk ki,

$$c_1 + c_2 x_1^{\beta_2-\beta_1} + \ldots + x_1^{\beta_{N+1}-\beta_1} \equiv c_1 + c_2 x_1^{\alpha_2-\alpha_1} + \ldots + x_1^{\alpha_{N+1}-\alpha_1}$$

(és így tovább a többi x_i változóra).

Ezek azonban feltevés szerint zérusok, következésképp (12) baloldala is zérus és a c_i értékekre a következő újabb összefüggést kapjuk:

(15) $$D_{\beta_1} c_1 + D_{\beta_2} c_2 + \ldots + D_{\beta_N} c_N = - D_{\beta_{N+1}}.$$

Hasonlóan járunk el a $\gamma_1, \gamma_2, \ldots$ indexű egyenletekkel, végeredményül a

(16) $$D_{\gamma_1} c_1 + D_{\gamma_2} c_2 + \ldots + D_{\gamma_N} c_N = - D_{\gamma_{N+1}}$$

összefüggést kapjuk. Újabb, a (8) és (9)-nek megfelelő $\tau_1, \tau_2, \ldots$ indexrendszerhez tartozó egyenleteket választva ki (7)-ből, ismét más összefüggést nyerünk a c_i-k közt. Az eljárást mindaddig folytatjuk, míg (12), (14), (15)

és a többi egyenlet N számú egyenletet nem szolgáltat. Ezek együtt lineáris egyenletrendszert adnak a c_i értékekre.

Tegyük fel, hogy ez az egyenletrendszer megoldható (ennek a vizsgálatára nem térünk ki ; ha a gyakorlatban valamilyen α_i, β_i, ... értékrendszernél esetleg zérus volna az egyenletrendszer determinánsa, más értékrendszerrel próbálkozunk). — A megoldással nyért c_i értékeket beírjuk a (12) egyenletbe s az így kapott $(\alpha_{N+1} - \alpha_1)$-edfokú egyismeretlenes egyenletet pontosan vagy közelítő módszerrel megoldjuk. A gyökök szolgáltatják az x_i értékeket, ezeket a (7) alatti egyenletrendszer N darab egyenletébe beírjuk és az ilymódon a B_i-értékekre kapott lineáris egyenletrendszert megoldjuk. Az x_i és B_i értékekből a keresett p_k, A_k értékek könnyen meghatározhatók.

Az ismeretlenek meghatározását tehát visszavezettük két N-ismeretlenes lineáris egyenletrendszer és egy $(\alpha_{N+1} - \alpha_1)$-edfokú egyismeretlenes egyenlet megoldására, mindehhez a szükséges kísérleti adatokat elég tág határok között választhattuk ki. A kísérleti adatokból akkor kell a legkevesebb, ha $\beta_1 = \alpha_2, \beta_2 = \alpha_3, \ldots, \gamma_1 = \alpha_3, \gamma_2 = \alpha_4, \ldots$, és így tovább, vagyis ha az egyes adatok egyenlő távolságú pontokból valók. Ez azonban meg is köti a pontok választását ; az eredeti (8) értékrendszer jobban biztosítja, hogy megbízható adatokat használhassunk fel. A (8) rendszert mindig az adatok megbízhatóságát figyelembevéve állítsuk össze. Természetesen csak addig használható a leírt eljárás, amíg (9) fennáll, ez azonban a gyakorlatban legtöbbször teljesül.

Általában ajánlatos többféleképpen megválasztott adatseregből többször határozni meg az ismeretlen paramétereket, s azután közepelni az eredményeket.

2—3 komponensnél ez az eljárás jól alkalmazható. Legfőbb haszna, hogy akkor is alkalmazható, ha az f_ν-értékeknek csak egy része ismeretes ; ez gyakran előfordul, mert a gyorsaság kedvéért nem mindig mérik ki az összes f_ν értéket.[3])

E paragrafus záradékául még a következőket említjük meg :

Ha $\max M_k$ illetve n nagy (esetleg több száz), a szóráscsökkentő eljárást meglassítja a nagy számolási munka. Az algebrai módszer ekkor is használható, feltételezve, hogy ismerjük a komponensek számát. Gondolhatnánk arra is, hogy az eloszlást lépcsősfüggvénnyé alakítjuk át (egységnyi vízszintes szakaszokat illesztünk az egyes pontokhoz), s ezt a függvényt valamilyen ortogonális függvényrendszer szerint kifejtjük, majd összefüggést keresünk a kiindulási és a csökkentett szórású keverék sorfejtésének együtthatói között. Sajnos, Fourier-sor alkalmazásakor ez az ötlet keresztülvihetetlen. Másirányú vizsgálataink még folyamatban vannak, s ezért nem időzünk tovább ennél a kérdésnél, csak azt említjük meg, hogy nagy $\max M_k$ illetve n esetén a binomiális eloszlások már jól közelíthetők a normális sűrűségfüggvénnyel, vagyis az eloszlás pontjait összekötő folytonos görbe jó közelítéssel Gauss-függvények keverékének tekinthető, erre pedig alkalmazhatjuk az [1]-ben leírt eljárásokat.

[3]) A (7) magasabbfokú egyenletrendszer egyike azoknak, amelyeket »ad hoc« módszerrel meg lehet oldani, illetve egyszerűbb egyenletekre vissza lehet vezetni. Az irodalomban nem sok ilyen speciális, megoldható típus ismeretes; érdemes volna az ilyeneket és megoldási módszerüket összegyűjteni.

3. §.

A következőkben célkitűzésünknek megfelelően Poisson-eloszlások konstans súlyokkal vett

$$f_\nu = \sum_{k=1}^{N} A_k \cdot e^{-\lambda_k} \frac{\lambda_k^\nu}{\nu!} \qquad (\nu = 0, 1, \ldots) \tag{17}$$

keverékének $\left(A_k > 0,\ \sum_{k=1}^{N} A_k = 1,\ \lambda_k > 0,\ \text{a } \lambda\text{-értékek különbözők} \right)$ felbontásával foglalkozunk: az f_ν értékek ismeretében meg akarjuk határozni az ismeretlen A_k és λ_k paramétereket.

A binomiális eloszlásnál bemutatott szóráscsökkentéses, illetve algebrai módszerek lényegében itt is alkalmazhatók. Itt is elég különböző λ_k értékek esetével foglalkoznunk.

A szóráscsökkentéses módszer ebben az esetben azon az észrevételen alapszik, hogy a $\mu < \min \lambda_k$ paraméterrel képezett

$$F_\nu(\mu) = \sum_{k=1}^{N} A_k \cdot e^{-(\lambda_k - \mu)} \frac{(\lambda_k - \mu)^\nu}{\nu!} \qquad (\nu = 0, 1, \ldots) \tag{18}$$

keverék a (17) alattitól csak abban különbözik, hogy a komponensek maximális eloszlásértéke ebben nagyobb, szórása kisebb, és $\mu \to \min \lambda_k = \lambda_\tau$ esetén a τ indexű komponens a $\nu = 0$-nál A_τ magasságban jelentkező egyetlen ponttá fajul el, ami paraméterek megállapítását, komponens-leválasztást, további kezelést épp oly módon tesz lehetővé, mint ahogy azt a binomiális eloszlás esetében már részletesen letárgyaltuk.

Tekintsük problémánkat itt is megoldottnak, ha az f_ν és általunk választott, monoton növekvő μ-értékek segítségével elő tudjuk állítani az $F_\nu(\mu)$ kifejezéseket.

Erről az előállításról szól a következő

Tétel: Poisson-eloszlások

$$f_\nu = \sum_{k=1}^{N} A_k\, e^{-\lambda_k} \frac{\lambda_k^\nu}{\nu!} \tag{19}$$

keveréke és az ismeretlen A_k, λ_k paraméterek meghatározására szolgáló

$$F_\nu(\mu) = \sum_{k=1}^{N} A_k \cdot e^{-(\lambda_k - \mu)} \frac{(\lambda_k - \mu)^\nu}{\nu!} \tag{20}$$

($\mu < \min \lambda_k$ általunk választott valós szám) *keverék közt a következő kapcsolat áll fenn:*

$$F_\nu(\mu) = e^\mu \sum_{\varrho=0}^{\nu} \frac{(-\mu)^{\nu-\varrho}}{(\nu-\varrho)!} f_\varrho\,. \tag{21}$$

Bizonyítás: Nyilván

$$F_\nu(\mu) = e^\mu \sum_{\varrho=0}^{\nu} \binom{\nu}{\varrho} \frac{1}{\nu!} (-\mu)^{\nu-\varrho} \sum_{k=1}^{N} A_k \cdot e^{-\lambda_k} \lambda_k =$$

$$= e^\mu \sum_{\varrho=0}^{\nu} \frac{(-\mu)^{\nu-\sigma}}{(\nu-\varrho)!} \sum_{k=1}^{N} A_k \cdot e^{-\lambda_k} \frac{\lambda_k^\varrho}{\varrho!} = e^\mu \sum_{\varrho=0}^{\nu} \frac{(-\mu)^{\nu-\varrho}}{(\nu-\varrho)!} f_\varrho$$

mivel a szummációkat felcserélhetjük, q. e. d.

Vegyük észre, hogy $F_\nu(\mu)$ meghatározásához az *első* f_ν-kre van szükség. Ha a legelső f_ν-k megbízhatatlanok, vagy nem mindegyiket ismerjük — viszont ismerjük a komponensek számát —, 2—3 komponensnél itt is alkalmazható algebrai eljárás, épp úgy, mint a binomiális eloszlásoknál láttuk. Az

$$f_\nu = \sum_{k=0}^{N} A_k \cdot e^{-\lambda_k} \frac{\lambda_k^\nu}{\nu!} \quad (\nu = 0, 1, \ldots) \tag{22}$$

értékeket az $A_k \cdot e^{-\lambda_k} = G_k$, $\nu!\, f_\nu = H_\nu$ relációkkal bevezetett új G_k, H_ν paraméterek segítségével átvisszük a

$$\sum_{k=1}^{N} G_k \cdot \lambda_k^\nu = H_\nu \tag{23}$$

egyenletrendszerbe. Ez azonban ugyanúgy kezelhető, mint a (7) egyenletrendszer. Gyakorlati felhasználására is az ott mondottak vonatkoznak, s így itt mindezt felesleges lenne megismételnünk.

IRODALOM

[1] Medgyessy P.: »Valószínűség-eloszlásfüggvények keverékének felbontása összetevőikre.« *A Magyar Tudományos Akadémia Alkalmazott Matematikai Intézetének Közleményei* 2 (1953) 165—177.

[2] Medgyessy P.: »Újabb eredmények valószínűség-eloszlásfüggvények keverékének összetevőire bontásával kapcsolatban.« *A Magyar Tudományos Akadémia Alkalmazott Matematikai Intézetének Közleményei* 3 (1954) 155—169 (jelen kötet).

[3] Medgyessy P.—Rényi A.—Tettamanti K.—Vincze I.: »A frakcionáló megosztás matematikai tárgyalása nem-teljes diffúzió esetében.« *A Magyar Tudományos Akadémia Alkalmazott Matematikai Intézetének Közleményei* 3 (1954) 81—97 (jelen kötet).

[4] C. F. Gauss: »Methodus nova integralium valores per approximationem inveniendi.« — *Werke*, Königliche Gesellschaft der Wissenschaften, Göttingen, 1866, Band 3. 165—196.

[5] C. Runge—R. König: *Vorlesungen über numerisches Rechnen*. Springer, Berlin, 1924.

[6] G. Doetsch: »Die Elimination des Dopplereffekts bei spektroskopischen Feinstrukturen und exakte Bestimmung der Komponenten.« *Zeitschrift für Physik* 49 (1928) 705—730.

РАЗЛОЖЕНИЕ НА КОМПОНЕНТЫ СМЕСИ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ

П. Меддеши

Резюме

Настоящая работа посвящается следующим проблемам:

a) Известное дискретное распределение вероятностей является наложением биномиальных распределений с различными параметрами, (то есть »смесью с постоянными весами«), то есть оно имеет вид (3). Зная f_ν, следует определить неизвестные параметры A_k, M_k, p_k компонентов.

b) Известное дискретное распределение является наложением (17) распределений Пуассона с различными параметрами. Следует определить неизвестные параметры A_k, λ_k.

Такие смеси часто встречаются при различных исследованиях. Проблема *a)* возникает, например, при т. н. фракционирующем разделении в химической промышленности (см. список литературы в (3)).

При решении обеих проблем используем, главным образом, т. н. »метод уменьшения дисперсии«, примененный в (1) и (2). В случае *a)* это приводит к следующему:

Зная значения f_ν из (3), мы в состоянии представлять распределение $F_\nu(\mu)$ (4). Оно отличается от (3) только тем, что все значения p_k умножены на выбранный нами параметр μ. Пусть распределения μ представляются при помощи возрастающих $F_\nu(\mu)$. Если параметром μ достигается значение $1/\max p_k$, то компоненты $F_\nu(\mu)$ с параметром вырождаются в единственную выступающую точку, из данных которой параметры соответствующих компонентов определяются хорошим приближением. В то же время уменьшается и дисперсия других компонент. После отделения вырожденных компонент, предлагаемый прием повторяется.

Величны $F_\nu(\mu)$, представляющие решение проблемы, суть линейные выражения, составленные из f_ϱ формулой (5).

В качестве иллюстрации приводим два примера (рис. 1 и 2). На рис. 1 исходная смесь распределение изображается полными точками (это распределение характеризует результат фракционирующего разделения). Результат разложения, соответствующего значению $\mu = 1{,}073$, представляется неполными точками. Видно, что мы имеем не 2, а 5 компонент. Это подтверждается и разложением, проведенном после отщепления сильно выступающей последней компоненты (рис. 1.b.). — На рис. 2.a. и 2.b. показывается разделение фиктивной смеси. Смесь имеет единственный максимум; после разделения отдельные компоненты отчетливо проявляются.

Описанный прием может найти применение лишь при известных f_ν. В случае, если известна лишь часть этих данных, можно применять следующий способ: Пусть нам известно число N компонент и $M_k = n$ ($k = 1,2,\ldots, N$). Запишем для достаточного числа точек сложного распределения соотношение (6). Таким образом мы приходим к системе уравнений для A_k и p_k, которую преобразуем к форме (7).

Предположим, что известные нам f_ν делают возможным записать (7) для систем $\nu = \alpha_k, \beta_k, \gamma_k, \ldots$ соответствующим (8) [(10, (14)]. Решение полученной таким образом системы уравнений приводится при помощи известного приема [5] к решению уравнения высшей степени и двух систем линейных уравнений. Из решений уже можно вычислить искомые A_k и p_k.

Следует заметить, что для больших n, смесь (6) с достаточной точностью представляется как смесь нормальных функций плотности. В этом случае применимы методы, изложенные в [1] и [2].

В случае разложения смеси (17) распределений Пуассона (случай *b*) мы также применяем изложенный метод. Основой разложения служит представление смеси (18) при помощи (17). Здесь параметр μ следует вычитать из λ_k. Компоненты (18) разделяются в случае $\mu = \min \lambda_k$ таким же образом, как и компоненты смеси *a*). Связь между (17) и (18) устанавливается формулой (21). При надобности можно прибегнуть и в данном случае к алгебраическим методам.

DECOMPOSITION OF DISCRETE COMPOUND PROBABILITY DISTRIBUTIONS

P. MEDGYESSY

Summary

This paper deals with the following problems :

a) A given discrete probability distribution is the superposition (a compound) of binomial distributions with different parameters, — i. e. it is of the form (3). The probabilities f_r being given, we have to determine the unknown parameters A_k, M_k, p_k of the components.

b) A given discrete probability distribution is a superposition (17) of Poisson-distributions with different parameters ; the unknown parameters A_k, λ_k are to be determined.

We often meet such compound distributions in various scientific investigations. E. g. the problem *a*) arises in connection with the procedure called »counter-current distribution« used in chemical industry (see the bibliography of [3]).

For the solution of both problems we primarily apply the so called »variance reduction method« used in [1] and [2]. In the case of *a*) the application of this method can be done as follows : From the knowledge of the probabilities f_ν (see (3)) we can construct the distribution $F_\nu(\mu)$ given by (4). This differs from (3) so far as all the p_k-s in it are multiplied with the parameter μ . μ will be chosen arbitrarily (but so that $\mu p_k \leqq 1$). When increasing μ we reach $1/\max\ p_k$, those components of $F_\nu(\mu)$ which have the parameter $\max\ p_k$ degenerate into a single point (see fig. 1. a.). If these points appear separately their coordinates furnish approximate values of the parameters A_k, M_k. — Simultaneously, the variances of all the other components will also diminish. After the separation of the degenerated components the procedure can be repeated.

The values $F_\nu(\mu)$ yielding the solution are given by (5) as linear forms of the probabilities f_ϱ.

The method is illustrated by two examples (Fig. 1. and 2.) On Fig. 1. the full points show the original compound distribution (it is the result of some counter-current distribution). The result of a transformation with $\mu = 1{,}073$ is marked by circles. It is easy to see that there are not 2 but 5 components. This is also justified by a second decomposition performed after the separation of the last component. (Fig. 1. b.)

Fig. 1. a. and b. show the decomposition of an artificial compound distribution. There is only one maximum ; after the decomposition the separate components present themselves very markedly.

The described method uses all the probabilities f_ν. If they are not all available, then instead of the preceding procedure we may apply the following method [inasmuch the number N of the components is known and $M_k = n$ $(k = 1, 2, \ldots, N)$]. We write (6) for an adequate set of the points of the compound distribution. Then we obtain a system of equations for the A_k and p_k-s, which can be brought to the form (7). Suppose that the known f_ν-s make possible to write (7) for the systems of values $\nu = \alpha_k$, β_k, γ_k, ... corresponding to (8) (see (10), (14)). By the aid of a well-known technique [5] the solution of the system of equations thus obtained can be reduced to that of an equation of higher degree with an only unknown and of two systems of linear equations. From this solution, the A_k and p_k-s can already be determined.

We remark that in the case of larger n, the compound distribution (6) can very well be approximated by a compound of normal frequency functions ; then the methods given in [1] and [2] may be applied.

Similarly, for the decomposition of a compound (17) of Poisson-distributions (case b)), the above method can be applied. The basis of a decomposition is the construction of the compound distribution (18) by the aid of (17). Here the parameter μ will be subtracted from the λ_k. In the case of $\mu = \min\ \lambda_k$ the components of (18) will generally separate from each other like those of the compound a). The relation between (17) and (18) is given by (21). If necessary an algebraic procedure may be applied in this case also.