

# Optimum methods in statistics

(Optimum-módszerek a statisztikában)

Szerkesztő: STEINER Ferenc

Akadémiai Kiadó, 1997, 370 oldal

Ez a könyv több szempontból is különleges. Egyik különlegessége az, hogy a kézirat elkészülte és a könyv megjelenése között mindössze 2 hónap, vagy még annyi sem múlt el. A jelen ismertetés írója augusztus 20-i határidővel nézte át a végső kéziratot, és az októberi frankfurti könyvvásárra a kötet kész volt — ez a magyar tudományos könyvkiadásban nagyon szokatlan eredmény. A gyors megjelentetés részben a *Művelődési és Közoktatási Minisztérium*, valamint szakmai szervezetek, így az *Eötvös Loránd Geofizikai Alapítvány*, a *GES Geofizikai Szolgáltató Kft.*, a *Magyar Geofizikusokért Alapítvány*, a *Miskolci Egyetem Bányamérnöki Kara* és a *MOL Rt.* anyagi támogatásának is köszönhető. A másik különlegesség: a „piros” kötet több korábbi társa mellé sorakozik fel, jelezve az adott területen a Miskolci Egyetemen STEINER Ferenc vezetésével, külső munkatársak bevonásával folyó évtizedes kutatómunka eredményeit. Az egykori „legnagyobb reciprokok módszere” a kezdeti, meglehetősen nehézkes és éppen ezért kevés helyen használt eljárásból az évtizedek során matematika-ilag is jól megalapozott, nagyon sok részletkérdésben is tisztázott módszerre fejlődött. Ennek a fejlődésnek a legújabb eredményeit foglalja össze a mostani könyv.

A könyv folytatja a (földtudományokban) alkalmazott, illetve a szerkesztő és csoportja által helyettük ajánlott statisztikai módszerek elemzését, ez alkalommal elsősorban az adott szakterületen előforduló eloszlástípusok vizsgálatát és a hozzájuk legcélszerűbben alkalmazható statisztikai módszerek kiválasztását hangsúlyozva. A csoport által bevezetett eloszlástípus-családok széles választékot tartalmaznak a szimmetrikus eloszlások körében, kezdve a viszonylag rövid szárnyú, tehát kevés, a várt értéktől nagymértékben eltérő adatot tartalmazó típusoktól a csoport által sokkal gyakoribbnak gondolt, nagyszámú ilyen adatot magukban foglaló eloszlásokig.

A megszokott szerzőgárda, geofizikusok, matematikusok összefogásában a gyakorlatban jól alkalmazható recepteket ad a felhasználóknak. Kiindulva az elméleti alapokból, foglalkozik az eloszlástípusokkal, az aktuális adatsorhoz illeszkedő eloszlástípus kiválasztásával, a szórás jellemzésével, a típus ismeretében az egyes statisztikai eljárás-

sok hatékonyságával, a robusztusság és a rezisztencia fogalmával, azok meghatározásával az egyes módszerek és eloszlástípusok esetében. Ezek a kérdések azért is fontosak, mert a gyakorlati adatrendszerek sok esetben kisebbek annál, mint amik a szokásos statisztikai módszerek alkalmazásához szükségesek, emiatt az eloszlástípus sem határozható meg megbízhatóan belőlük. Ezért viszont fontossá válik az a kérdés, hogy választásunk olyan statisztikai módszerre essék, amely akkor is nagy hatásfokkal működik, ha a feltételezett és a valóságos (de meg nem határozható) eloszlástípus nem egyezik meg.

Kétségtelen az, hogy kezdetben a „legnagyobb reciprokok módszere” jelentős többletmunkát igényelt, és a tapasztalatlan felhasználó könnyen elkövethette azt a hibát, hogy a függvénynek nem az abszolút minimumát találta meg, hanem egy helyi minimumot — aminek azután nagyon kellemetlen következményei lehettek. Ma már ilyen probléma nincsen, a „leggyakoribb érték” és a megfelelő norma megtalálása rutinfeladat. A könyv utolsó fejezetében számos példát ad meg az alkalmazás lehetőségeiről úgy, hogy ezek a példák lehetőleg széles választékot nyújtsanak, mind a statisztikai módszerek, mind az alkalmazás geofizikai jellegét illetően. Ennek megfelelően az első példa egy vető helyzetének meghatározása bányában, ami régebben nem volt éppen könnyű feladat, különösen azért, mert a mérések elrendezése nem szabadon választható. Evvel kapcsolatban alapos összehasonlításokat adnak a szerzők az egyes eljárások eredményeiről, pontosságáról. A második példa gravitációs módszerrel történő üregkutatás, ahol a kedvezőtlen jel/zaj viszony szintén nagyon megnehezíti az inverziót. A harmadik példa fűrólyukszelvények adataiból történő rétegmeghatározás; ebben a tényleges felhasználás szintjéig kidolgozott módszerrel ismerkedhet meg az olvasó. Végül az utolsó, a negyedik példa a korrelációs számításra való alkalmazást mutatja be, a korrelációs faktor meghatározásának problémáit a legkisebb négyzetek és a leggyakoribb érték alapján összehasonlítva. Ezek és a korábbi könyvekben közölt példák sok felhasználót fognak meggyőzni a leggyakoribb érték módszerének előnyeiről a többi statisztikai módszerrel szemben.

Végül kilenc függelék egyes részkérdéseket tárgyal, ezek között például a típus meghatározását a minta terjedelme alapján, vagy a geofizikai műszerekbe egyre gyakrabban beépítésre kerülő előfeldolgozó programok megválasztásának problémáját. Ezek a függelékek, bár nem kapcsolódnak szorosan a könyv gerincéhez, mégis sok kérdésben tájékoztatják az olvasót fontos problémák felől.

A könyv kiegyensúlyozottan tartalmaz elméleti és gyakorlati, a felhasználókat tájékoztató részeket, ügyelve arra, hogy az elméleti kérdésekben járhatóbb és azok iránt kevésbé érdeklődő olvasó is használhassa az ajánlott módszereket. A könyvön végigvonul a vita a hagyományos, egyszerűsége miatt közkedvelt legkisebb négyzetes módszerrel,

avval a dogmával, hogy a hibák normális eloszlásúak (egyébként ez a címe a függelékek közül az egyiknek is).

Úgy gondolom, hogy mindenki, nemcsak a földtudományok művelői, hanem mások is, haszonnal forgatják majd ezt a könyvet, ha nagyobb — esetenként kisebb — adatmennyiségekből kell statisztikai alapon következtetéseket levonniuk. Ha szélesebb körben elterjednek a könyvben ismertetett módszerek, akkor a statisztikai módszerek jelentős hatékonyság növekedésére számíthatunk.

Az alábbiakban mellékeljük a könyv tartalomjegyzékét.

*Verő József*

<b>Preface</b> .....	9	<b>4. Different characteristics for the dispersion of data</b> .....	87
<b>Introduction</b> .....	25	4.1. Various measures of the uncertainty (L. Csernyák–B. Hajagos–F. Steiner) .....	87
<b>1. Type-distance of probability distribution pairs</b> .....	43	4.2. Comparison of minimum norms and semi-intersextile ranges for the type interval Cauchy–Gaussian (F. Steiner) .....	91
1.1. Definition of type-distance (F. Steiner–B. Hajagos) .....	43	4.3. Asymptotic scatter of some dispersion characteristics (F. Steiner) .....	92
1.2. Study of the distance of types of the $f_a(x)$ -supermodel from that of the Gaussian (L. Csernyák) .....	46	4.4. Error of indirectly measured quantities (F. Steiner) .....	97
<b>2. Determination of probability distribution types</b> .....	51	<b>5. Asymptotic scatter characterising the accuracy of the determinations of the location parameter</b> .....	101
2.1. Choice of a type from an adequate supermodel. Sample sizes needed for distinction between similar types (B. Hajagos–F. Steiner) ..	51	5.1. Some general formulae of the robust statistics for calculating the asymptotic scatter of the location parameter (F. Steiner) .....	101
2.2. Type-determinations in the neighbourhood of the Gaussian on the basis of statistical moments (F. Steiner–B. Hajagos) .....	52	5.2. Contribution to the so-called CML-estimate. Comments on similar cases (i.e., if the asymptotic scatter is infinite) and some other related topics (L. Csernyák–F. Steiner) .....	105
2.3. Estimation of type using $F^{-1}$ -scaled ordinates (F. Steiner) .....	54	5.3. Asymptotic variance of the most frequent value and of the dihesion (B. Hajagos).....	114
2.4. Determination of types minimising the type distance between the given empirical distribution function and the theoretical $F_a(x)$ (F. Steiner–B. Hajagos) .....	56	<b>6. Statistical efficiencies</b> .....	119
<b>3. Norms of deviations and residuals</b> .....	67	6.1. Practical importance of statistical efficiencies (F. Steiner) .....	119
3.1. The integral expressions of the $P_k$ - and $L_p$ -norms and their connections with the types of the $f_a(x)$ - and $f_p(x)$ supermodels (F. Steiner) .....	67	6.2. Asymptotic behaviour of statistical efficiencies if the flanks are larger and larger (F. Steiner) .....	122
3.2. Some mathematical aspects concerning norms .....	70	6.3. Statistical efficiency in function of the type distance of $f_a(x)$ -types from the Gaussian (F. Steiner–B. Hajagos–G. Hursán) .....	134
3.2.1. Theoretical results related to the $P_k$ norms (L. Csernyák) .....	70	<b>7. Indices of robustness for characterising the weak or strong dependency of efficiencies upon the error distribution types</b> .....	141
3.2.2. Connections between the norms $P_k$ and $L_2$ used in statistics (B. Hajagos).....	77		
3.3. The $P_k^*$ -norm (B. Hajagos–F. Steiner) .....	81		

7.1. Practical definition of robustness also for error distribution types having small flanks (F. Steiner–B. Hajagos) .....	141
7.2. Indices of the general robustness (F. Steiner–B. Hajagos) .....	153
<b>8. Resistance against outliers. Breakdown bounds for the practice</b> .....	157
8.1. Methods to increase the resistance of the computation of most frequent values (B. Hajagos–F. Steiner) .....	157
8.2. Distortion of error characteristics by outliers (F. Steiner) .....	165
8.3. Investigations concerning resistance and breakdown bounds (B. Hajagos–F. Steiner)....	174
8.4. Comparison of the resistances of the minimum $P_k$ - and $P_k^*$ -norms (i.e., of the $U_k$ - and $U_k^*$ -uncertainties) for the example given in the Section 10.1. (B. Hajagos–F. Steiner) .....	187
<b>9. Generalised and robustified covariance and correlation matrix</b> .....	193
9.1. Measure of the linear dependence (B. Hajagos–F. Steiner) .....	93
9.2. Generalisation of the covariance and correlation matrix (B. Hajagos–F. Steiner) .....	206
9.3. Robustification of the correlation and covariance matrix (B. Hajagos–F. Steiner) .....	229
<b>10. Examples to show the application of modern optimum methods</b> .....	235
10.1. Determination of fault's position in mines. Comparison of results obtained by ten norms (B. Hajagos–F. Steiner) .....	235
10.2. Gravimetrical example from the environmental geophysics. The determination method of parameter errors if an arbitrary norm for inversion is minimised (F. Steiner–B. Hajagos). .....	252
10.3. Joint inversion of well log data minimising the $P$ -norm (P. Szűcs) .....	257
10.4. Estimation of the correlation coefficient of a parameter-pair simultaneously with the parameter errors (shown on an electromagnetic example) (E. Turai) .....	275
<b>APPENDICES</b> .....	285
App. I. Basic difference between the determination method of the scale parameter between the maximum likelihood principle (applied as $\partial L/\partial S=0$ ) and the MFV-procedures (minimising the information loss) (L. Csernyák) .....	287
App. II. Determination of type using sample range (L. Csernyák) .....	290
App. III. Comment on an old dogma: "The data are normally distributed" (P. Szűcs) .....	294
App. IV. Opposite behaviour of the $P$ -norms compared to that of the $L_2$ -norm with respect to the simultaneously achieved accuracy in the space- and frequency-domain (F. Steiner–B. Hajagos) .....	299
App. V. Theoretical and practical consequences of the global optimisation methods (P. Szűcs) .....	303
App. VI. MFV-filtering to suppress errors. — A comparison with median filters (B. Hajagos–F. Steiner) .....	312
App. VII. "Built-in" statistics in geophysical instruments (B. Hajagos–F. Steiner) .....	318
App. VIII. Symmetrical stable probability distributions nearest lying to the types of the supermodel $f_a(x)$ (B. Hajagos–F. Steiner) .....	325
App. IX. MFV-corrected variances (F. Steiner–B. Hajagos–G. Hursán) .....	329
<b>Acknowledgement — Concluding meditation of the editor</b> .....	347
<b>References</b> .....	351
<b>Bibliography</b> .....	355
<b>Concise table of some statistical principles and norms</b> .....	365