MASARYKOVA UNIVERZITA

Přírodovědecká fakulta

DISERTAČNÍ PRÁCE

Brno, 2009 **Hana Kotoučková**

MASARYKOVA UNIVERZITA

Přírodovědecká fakulta

Hana Kotoučková

Historie robustních matematicko-statistických metod

Disertační práce

Školitel: prof. RNDr. Jana Jurečková, DrSc.

Brno, 2009

Bibliografická identifikace

Jméno a příjmení autora: **Hana Kotoučková**

Název disertační práce: **Historie robustních matematicko-statistických metod**

Název disertační práce anglicky: **History of Robust Statistical Methods**

Studijní program: Matematika

Studijní obor (směr), kombinace oborů: Obecné otázky matematiky

Školitel: prof. RNDr. Jana Jurečková, DrSc.

Rok obhajoby: 2009

Klíčová slova v češtině: Robustní metody, metoda nejmenších čtverců, small sample asymptotics, Mayer, Boscovich, Laplace, Gauss, Newcomb, Hulst, Princetonská studie, Huber, Hampel, mnohorozměrné modely.

Klíčová slova v angličtině: Robust methods, least squares method, small sample asymptotics, Mayer, Boscovich, Laplace, Gauss, Newcomb, Hulst, Princeton study, Huber, Hampel, multivariate model.

Prohlášení:

Prohlašuji, že jsem dizertační práci zpracovala samostatně a s použitím uvedené literatury.

V Brně dne 22. 3. 2009 Hana Kotoučková

Poděkování

Děkuji svojí školitelce prof. RNDr. Janě Jurečkové, DrSc. za její pomoc, cenné rady a čas, který mi věnovala při přípravě a psaní této práce.

Abstrakt

Cílem práce bylo stručně popsat, jak se začaly objevovat statistické postupy, které bychom dnes označili přízviskem robustní. Především šlo o to popsat, proč vůbec vyvstala potřeba používat nové postupy v situacích, kdy klasické statistické metody selhávaly. Nejprve jsou krátce popsány počátky samotné teorie pravděpodobnosti, kombinatoriky, demografie a pojistné matematiky. Samostatná kapitola je věnována metodě nejmenších čtverců. Metoda nejmenších čtverců nebyla jedinou možností, jak kombinovat nekonzistentní rovnice, nicméně byla nejúspěšnější. Problémem kombinování jednotlivých pozorování se zabýval kartograf Tobias Mayer, matematik Leonard Euler nebo Pierre Simon Laplace. Poněkud méně známou osobností je jezuitský kněz Roger Joseph Boscovich, který dal slovní a geometrický popis metody, která může být považována za alternativu k metodě nejmenších čtverců. Samotnou metodu nejmenších čtverců pak uveřejňuje roku 1805 Adrie Marie Legendre. Nicméně již o čtyři roky později si objev této metody přisvojuje Carl Friedrich Gauss. Gauss je spojen i se zavedením normálního rozdělení. Dlouhou dobu se předpokládalo, že většina náhodných veličin se řídí právě tímto rozdělením a již se neověřovalo, jestli se skutečně zkoumaná veličina řídí normálním rozdělením. Pokud se nejedná o normální rozdělení, jsou např. přítomny odlehlé hodnoty, klasické statistické procedury (včetně metody nejmenších čtverců) mohou selhat. Mezi jedny z prvních, kteří si tohoto faktu všimli, patří americký astronom Simon Newcomb, matematik Percy John Daniell nebo nizozemský astronom Hendrik van de Hulst. Moderní robustní metody se objevují až v druhé polovině dvacátého století. Zásadním byl z tohoto pohledu článek Petera Hubera z roku 1964 *Robust Estimation of a Location Parameter*. Huber zde mj. zavádí *M*-odhady. V letech 1970–1971 se v Princetonu konal seminář o robustních odhadech. Poznatky byly shrnuty v knize *Robust Estimates of Location: Survey and Advances*. Autoři zde prezentují nové robustní odhady a tyto odhady pak porovnávají mezi sebou pomocí influenční funkce a bodu selhání, které navrhl Frank Hampel. Přirozeně bylo žádoucí u robustních odhadů zjistit alespoň přibližné rozdělení pravděpodobností. Touto problematikou se zabýval např. britský statistik Henry Ellis Daniels, švýcarský statistik Frank Hampel a Kanaďan Christopher Field. Ricardo Antonio Maronna (Argentina) přišel již v roce 1976 s robustními metodami v mnohorozměrných modelech, které jsou trendem i v současnosti. Práce je zakončena pohledem na problémy a směry současné robustní statistiky.

Abstract

The intention of this work is to give a brief description of the beginning of robust methods. My primary goal was to describe the situations, where the classical statistical procedures did not work, what was a reason for finding new methods. The first part of the Thesis briefly describes the foundation of probability theory, combinatorics, insurance mathematics and demography. The next chapter deals with the least squares method, which was the most succesful among the early methods of combining inconsistent equations. The problem of combining inconsistent equations was studied by cartographer Tobias Mayer and by mathematicians Leonard Euler and Pierre Simon Laplace. The Jesuit Roger Joseph Boscovich gave a geometric description of the method that was a predecessor of the least squares. The first publication on the least squares method was that by Adrien Marie Legendre in 1805. However, Carl Friedrich Gauss denoted himself as the author of the least squares in his book, published four years later. The name of Gauss is also connected with the introduction of the normal distribution. The normal distribution was for a long time considered as the probability distribution of most random variables. However, it turns out that the classical statistical methods (including the least squares) can fail if the random variable doesn`t come from the normal distribution, as was observed by American astronomer Simon Newcomb, English mathematician Percy John Daniell or Dutch astronomer Hendrik van de Hulst. Modern robust methods appeared in the second half of the twentieth century. In 1964 Peter Huber published the significant article *Robust Estimation of a Location Parameter,* introducing the *M*-estimators. In 1970–1971, John Wilder Tukey organized seminars on robust estimators in Princeton. Their results and conclusions the Princeton group published in the book *Robust Estimates of Location: Survey and Advances*. The authors studied new robust estimators and compared them by means of the influence function and the breakdown point, the concepts introduced by Frank Hampel. An important problem was to derive the probability distribution of robust estimators, at least approximatively. This problem have been solved by British statistician Henry Ellis Daniels, Swiss statistician Frank Hampel and Canadian Christopher Field. The first extension of the robust methods to multivariate models was made by Ricardo Antonia Maronna (Argentina) in 1976. The robust methods in multivariate models are the main problem of interest up to now. The final part of the Thesis briefly describes the recent trends in robust statistics.

Obsah

Úvod

Disertační práce se věnuje vzniku teorie pravděpodobnosti, metodě nejmenších čtverců a především počátkům robustních matematicko-statistických metod. Výraz „robustní“ se ve statistice používá poměrně krátkou dobu. Statistický význam tomuto slovu dal až G. E. P. Box v polovině dvacátého století. V dnešní době existují různé definice, více či méně matematicky přesné, ale obecně robustní znamená necitlivý na malé odchylky z idealizovaných předpokladů, pro které je odhad optimalizován. Mne ale nezajímaly pouze moderní robustní metody, jak je známe dnes a jak se vyučují na vysokých školách. Zajímala jsem se spíše o to, jak robustní postupy vznikly, co robustním metodám předcházelo, v jakých souvislostech se objevila potřeba používat jiné, jak klasické statistické metody.

Aby bylo možné dostat se ke vzniku robustních metod, podívala jsem se nejprve na to, jak vznikla pravděpodobnost a statistika jako taková. Často bývá za počátek teorie pravděpodobnosti považována korespondence z roku 1654 mezi Pascalem (Blaise Pascal, 1623–1662) a Fermatem (Pierre de Fermat, 1601–1665). Nicméně počátky musíme hledat mnohem dříve. Teorie pravděpodobnosti se využívala i při různých hrách. Jednou z nejstarších her byla nejspíš hra s kostkami. Kostky mohly být nahrazovány i jinými předměty, jako jsou kůstky nebo tyčinky. Tyto hry se objevují již na starověkých egyptských malbách (kolem roku 3500 před n. l.).

Počátky teorie pravděpodobnosti stručně popisuje první kapitola této disertační práce. Zabývám se zde mj. úlohou o rozdělení sázky, první publikovanou prací o počtu pravděpodobnosti *De ratiociniis in ludo aleae*, kterou napsal Christiaan Huygens (1629–1695). Huygensův spis je přetištěn, rozebírán a úlohy zobecňovány v díle Jakoba Bernoulliho (1654–1705) *Ars conjectandi*. Kombinatorické úlohy se sice objevují již mnohem dříve, ale kombinatorika jako samostatná část matematiky se začíná vyčleňovat až v polovině 17. století. Jedno z prvních děl zabývajících se kombinatorikou *Traité du triangle arithemétique* pochází od Blaise Pascala (1623–1662). Pascal sám však dílo nevydal, jedná se spíše o soubor textů, který byl vydán až po Pascalově smrti roku 1665. V druhé polovině 17. století vzniká infinitesimální počet. Tyto nové metody používají v teorii pravděpodobnosti Pierre-Rémond de Montmort (1642–1727) a Abraham De Moivre (1667–1754). V souvislosti s rozvojem teorie pravděpodobnosti jsem se také krátce zmínila o demografii a pojistné matematice. Nejstarší známé úmrtnostní tabulky pocházejí již z roku 211 a jejich autorem byl římský právník Domitius Ulpianus (zemřel r. 228). V Evropě jsou známy první seznamy zemřelých

až z roku 1532 v souvislosti s morovou epidemií v Anglii. V roce 1592 začínají v Londýně vycházet pravidelné týdenní seznamy úmrtí. Z těchto studií vycházela i první publikovaná práce z pojistné matematiky z roku 1662, jejímž autorem byl John Graunt (1620–1674). Teorií doživotních důchodů se zabýval dokonce i anglický astronom Edmond Halley (1656–1742).

Druhá kapitola se věnuje historickým kořenům metody nejmenších čtverců. Metoda nejmenších čtverců byla nejúspěšnější z raných metod kombinování nekonzistentních rovnic. Důvodem obliby této metody byl fakt, že byla založena na snadno pochopitelných kritériích. Je zajímavé podívat se na pozadí jejího vzniku v souvislosti s geofyzikálními a astronomickými problémy 18. a 19. století. Mezi hlavní problémy patří např. určení a matematický popis pohybů Měsíce, vysvětlení neperiodických odchylek v pohybu planet Jupiteru a Saturnu, stanovení přesného tvaru zeměkoule nebo stanovení střední hustoty Země a gravitační konstanty.

Tobias Mayer (1723–1762) pracoval jako kartograf v Norimberku. V letech 1748–1749 provedl mnoho pozorování Měsíce a jeho pohyb popsal pomocí 27 nekonzistentních rovnic o třech neznámých. Byl přesvědčen o tom, že přesnost výsledku lze zlepšit vhodnou kombinací jednotlivých pozorování a navrhl statistické řešení problému, jak pozorování vhodně kombinovat. Výsledkem toho byla soustavy tří rovnic pro tři neznámé.

Leonard Euler (1707–1783) se zabýval výpočtem drah planet, variačním počtem, pohyby Měsíce, ale také aplikací matematiky např. ve stavbě lodí a ve fyzice. Zaměřila jsem se na jeho studium nepravidelností ve vzájemném pohybu Saturnu a Jupitera. Euler po ověření na empirických datech získal soustavu 75 rovnic s osmi neznámými. Protože byl ale exaktním matematikem, nepoužil pro zlepšení přesnosti kombinaci pozorování jako Mayer.

Méně známou postavou je jezuitský kněz Roger Joseph Boscovich (1711–1787). Zabýval se měřením poledníkového úhlu. Při analýzách těchto dat dospěl k prvnímu úspěšnému řešení nekonzistence různých obloukových měření. Svoji metodu však neformuluje analyticky, dává pouze slovní a geometrický popis.

Na Eulera a Mayera navázal ve své analýze pohybů Saturnu a Jupitera Pierre Simon Laplace (1749–1827). Pomocí kombinace nekonzistentních lineárních rovnic odhalil periodicitu pohybů těchto dvou planet. V roce 1787 své závěry publikoval v knize *Théorie de Jupiter et de Saturne* [55].

V roce 1805 Adrien Marie Legendre (1752–1833) publikoval knihu o určování dráhy komet *Nouvelles méthodes pour la déterminations des orbits des cometes* [56]. V apendixu

své knihy navrhl metodu nejmenších čtverců, kterou aplikoval na naměřená pozorování drah komet. V roce 1809 Carl Friedrich Gauss publikoval svoji metodu nejmenších čtverců a směle připisoval objevení této metody sobě se slovy, že uvedenou metodu používá již od roku 1795. Toto tvrzení ovšem nemohl doložit žádnou konkrétní publikací a jeho nárokování prvenství je tak založeno pouze na nepřímých důkazech, které jsou v druhé kapitole zmíněny. Brzy po publikaci Legendreovy práce se metoda nejmenších čtverců stala standardní metodou astronomie a geodézie.

Carl Friedrich Gauss (1777–1855) se zabýval podobnými problémy jako Mayer, Euler, Laplace a Legendre. Gauss ale hledal rozdělení pravděpodobností chyb měření, pro které je odhad metodou nejmenší čtverců optimálním odhadem měřené konstanty. Dospěl tak k hustotě normálního rozdělení. Toto rozdělení má v matematické statistice velice významnou roli. Od roku 1810 se dokonce předpokládalo, že většina náhodných veličin se řídí právě tímto rozdělením.

Normální rozdělení bylo dlouhou dobu považováno za rozdělení, kterým se řídí většina náhodných veličin. I při používání metody nejmenších čtverců se často bez dalšího ověřování předpokládalo, že velikosti chyb pochází z normálního rozdělení. Na odlehlé hodnoty jsou citlivé i klasické statistické procedury, jako je průměr a rozptyl. Právě o citlivosti klasických statistických charakteristik a procedur na odchylky od normality pojednává třetí kapitola. V této souvislosti zmiňuji amerického astronoma Simona Newcomba (1835–1909), objevitele směsi normálních rozdělení. S průkopnickou prací *Observations Weighted According to Order* přichází v roce 1920 Percy John Daniell (1889–1946). Zavádí zde statistiky, kde jsou váhy přiřazeny měřením podle jejich pořadí. Nicméně Daniellovo dílo zůstalo nepovšimnuto a znovuobjevení jeho výsledků trvalo dalších dvacet let. Robustní statistiky bychom mohli najít v práci nizozemského astronoma Hendrika Ch. van de Hulsta (1918–2000). Ale i jeho práce zůstala dlouhou dobu nepovšimnuta. Hulst se zabýval robustní statistikou pouze krátce. Tím hlavním důvodem bylo nejspíš to, že v roce 1944 začal slavil úspěchy na poli teoretické astronomie.

Poslední kapitola se zabývá moderními robustními odhady. Hned na začátek jsem zařadila něco málo z teorie robustních metod kvůli častému používání těchto pojmů ve zbytku kapitoly. Zásadním zlomem pro moderní robustní statistické metody byl článek Petera J. Hubera (*1934) z roku 1964 *Robust Estimation of a Location Parameter*. Huber zde mj. pojednává o asymptotické teorii odhadu parametru polohy pro kontaminovaná normální

rozdělení. Zavádí zde také tzv. *M*-odhady. Další významnou událostí pro rozvoj robustních metod byl seminář v Princetonu, který se konal v akademickém roce 1970–1971. V té době zde působil jako profesor statistiky John Wilder Tukey (1915–2000). Semináře se účastnili nejen zaměstnanci a studenti Princetonu, ale i pozvaní hosté. Základem této skupiny byli David F. Andrews , Peter J. Bickel, Frank R. Hampel, Peter J. Huber, W. H. Rogers a John W. Tukey. Princetonská skupina upozornila na nedostatky klasických odhadů a nabídla jako alternativu robustní odhady. Příspěvky byly posouzeny a následně doplněny. Výsledkem tohoto snažení pak byla kniha *Robust Estimates of Location: Survey and Advances*, která vyšla v roce 1972. Autoři nejen navrhli nové robustní odhady, ale navíc se snažili určit kvality i případné nedostatky těchto odhadů pomocí Hampelova bodu selhání a influeční funkce.

A právě o influeční funkci a bodu selhání, které studoval Frank Hampel, je další část čtvrté kapitoly. Stručně bychom mohli říct, že bod selhání je nejmenší podíl pozorování, která po nahrazení libovolnými hodnotami (co nejnepříznivějšími) mohou vést k hodnotám odhadu nekonečno.

Po Princetonské studii nastal poměrně překotný rozvoj robustních odhadů. Stalo se trendem uvažovat pouze odhady, které jsou „dostatečně dobré“. Tato kvalita odhadu byla měřena pomocí bodu selhání. Odhady, které měly bod selhání menší než jedna polovina, byly považovány za nedostatečné. Nad tím, zda je nutné dodržovat toto striktní omezení, zda jsou skutečná data opravdu tak pesimistická, se zamýšlel Stephen Stigler v roce 1977. Jeho závěry jsem zařadila do části *Povaha reálných dat*.

U robustních odhadů by bylo logicky žádoucí zjistit jejich alespoň přibližné rozdělení pravděpodobností. V roce 1954 používá britský statistik Henry Ellis Daniels (1912–2000) metodu sedlového bodu k odvození velice přesné aproximace rozdělení aritmetického průměru. Obdobnou problematikou se zabýval také Frank Hampel. Vhodně zde použil označení „small sample asymptotics“, který vystihuje podstatu těchto metod. V části *Small sample asymptotics* je dále pojednáno i o práci Christophera A. Fielda.

V současnosti jsou trendem robustní metody v mnohorozměrných modelech. Nicméně nejedná se o novinku několika posledních let. Na pole robustních odhadů přišel s mnohorozměrnými modely Ricardo Antonio Maronna již v roce 1976. Disertační práci uzavírá krátké pojednání o současných problémech a směrech robustní statistiky a o jejích přínosech pro obecný rozvoj moderní statistiky.

Kromě uvedené literatury jsem ještě používala dva cenné internetové zdroje. Tím prvním byl The Mathematics Genealogy Project, dostupný na http://www.genealogy.ams.org/.

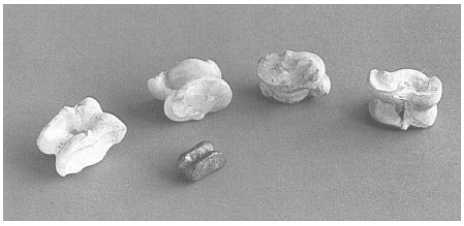
Jedná se o službu, kterou poskytuje Oddělení matematiky na North Dakota State University společně s Americkou matematickou společností. Druhým zdrojem pak byl The MacTutor History of Mathematics archive (http://www.history.mcs.st-andrews.ac.uk/history/), vytvořený Johnem J. O'Connorem a Edmundem F. Robertsonem z University of St. Andrews.

Kapitola 1:

Historický rozvoj pravděpodobnostního a statistického způsobu vnímání

1.1 Počátky teorie pravděpodobnosti

Za počátek teorie pravděpodobnosti je obecně považována korespondence mezi Pascalem[1] a Fermatem[2], kterou spolu vedli v létě a na podzim roku 1654. Nicméně Pascal s Fermatem měli velké množství předchůdců. Počátky teorie pravděpodobnosti lze hledat v hrách. Nejstarší z těchto her byla nejspíš hra v kostky. Ovšem v počátcích se neobjevuje hra s kostkami, jak je známe dnes, ale s jinými předměty, které je nahrazovaly. Nejstarší z nich jsou patrně hlezenní kosti kopytnatců (viz obr. 1.1.). Tato „kostka" může padnout čtyřmi způsoby. Hlezenní kosti v lidských sídlištích v nápadně velkém množství jsou nacházeny v sídlištích starých až 40 000 let. Archeologové se domnívají, že tyto kůstky byly používány právě k hrám.



Obrázek 1.1. *Hlezenní kůstky kopytnatců*

Na starověkých egyptských malbách z I. dynastie (kolem roku 3500 před n. l.) lze nalézt znázornění her, při kterých se používala kostka. V čem přesně hra spočívala a jaká byla její pravidla, známo není. Nicméně ví se alespoň to, že hra spočívala v posouvání figurek a k určení počtu kroků se používaly buď hlezenní kosti nebo tyčinky.

Nejspíš první hry o peníze se objevují ve starověkém Řecku. Řekové házeli hlezenními kůstkami, které měly očíslované strany 1, 3, 4 a 6. V Řecku byly kostky oblíbenou hrou, o čemž svědčí fakt, že byly častým motivem uměleckých děl. Kostky se objevovaly mj. na keramice, kresbách, mincích (viz obr. 1.2.).

[1] **Blaise Pascal** (1623–1662) – narodil se v Clermon – Ferrandu, jeho otec byl soudcem, ale také se zabýval matematikou. Ve dvaceti letech Blaise Pascal sestrojil počítací strojek, který pomáhal jeho otci při výpočtu daní. Zabýval se pravděpodobností, projektivní geometrií, ale i fyzikou, filozofií a teologií. Blaise Pascal umírá ve věku 39 let v Paříži v důsledku těžké nemoci.

[2] **Pierre de Fermat** (1601–1665) – vystudoval právo na univerzitě v Orleans. Pracoval jako člen parlamentu v Toulouse. Zabýval se matematickou analýzou, teorií čísel, ale i optikou. Je považován za zakladatele teorie čísel (Velká Fermatova věta).

I Římané se ve starověku věnovali hře v kostky. Používali kostky kamenné. Ve starověké Indii zase využívali na hry oříšky vibhidaka. Nikde se ale doposud neobjevuje známka o počítání relativních četností jednotlivých hodů. Někdy se jako důvod uvádí nedokonalost kostek a tudíž nemožnost počítat četnosti vrhů.

Obrázek 1.2. Keramika *znázorňující ženy při hře v kostky (kolem r. 300 před n. l.)*

Kombinatorika se objevuje nejdříve v Asii. *Bhagabati Sútra* přibližně z roku 300 před n. l. obsahuje počty kombinací a permutací k prvků z n pro $k = 1, 2, 3$. Tyto počty se využívají při řešení problémů typu „jaké podsoubory lze vytvořit z daného počtu mužů a žen“ apod. I v dalších sútrách lze nalézt kombinatorické otázky.

S pravděpodobností se setkáváme i v rabínské literatuře. U Židů byla hra v kostky zakázána, nicméně los se používal jak v liturgii, tak pro rozhodování sporů. Nejčastějším způsobem, jak dát průchod náhodě, bylo losování z urny. V rabínské literatuře se objevuje základní počítání s pravděpodobnostmi poměrně často, avšak především ve smyslu řešení praktických příkladů. Příklady na řešení situací založených na náhodě můžeme nalézt v *Talmudu*[3] . Jsou zde úlohy např. na násobení pravděpodobností.

V Evropě se objevuje výpočet z kombinatoriky u biskupa Wibolda z Cambray. Ten pro mnichy vymyslel hru (kolem r. 965), kde mniši házeli třemi kostkami a tím dostali jednu z možných 56 kombinací. Mnich si tak vylosoval jednu z 56 ctností, kterou musel po následujících dvacet čtyři hodin praktikovat. Některé čtenáře možná překvapí fakt, že hra pochází z církevního prostředí. V Evropě totiž byla hráčská vášeň kritizována a potlačována

[3] **Talmud** je jedním z hlavní děl rabínské literatury. Objevuje se ve dvou variantách – jeruzalémské (okolo r. 400 a babylónské (427–560). Jedná se o soubor předpisů nejen náboženských, ale i právních z oblasti trestního a občanského práva.

nejen ze strany státu, ale i ze strany církve. Nicméně zákaz se vztahoval pouze na hry o peníze. Proto i v době III. křížové výpravy měli vojáci jasně vymezená pravidla pro hraní kostek. Kombinatorická čísla se objevují v básni *De vetula* (*O stařence*). Autorem této básně byl nejspíš Richard de Fournival[4]. Uvádí zde návod, jak počítat kombinatorická čísla pro počty možností jednotlivých součtů při hodu třemi kostkami.

Ve srovnání s jinými matematickými disciplínami se teorie pravděpodobnosti začala rozvíjet až poměrně pozdě. Podívejme se tedy nejprve na možné příčiny tohoto pozdního vývoje. Jednou z nich je i to, že nikdo neviděl žádnou souvislost mezi náhodnými jevy a matematikou. Náhoda byla buď zbožštěna (nic se neděje náhodně, vše je řízeno vyšší mocí), nebo byla náhoda považována za pouhou neznalost všech vztahů a příčin. Další příčinou bylo i to, že zkoumání náhody dříve nebylo nutné. Potřeba popsat náhodné jevy se objevuje v souvislosti s rozvojem demografie, pojišťovnictví, astronomie.

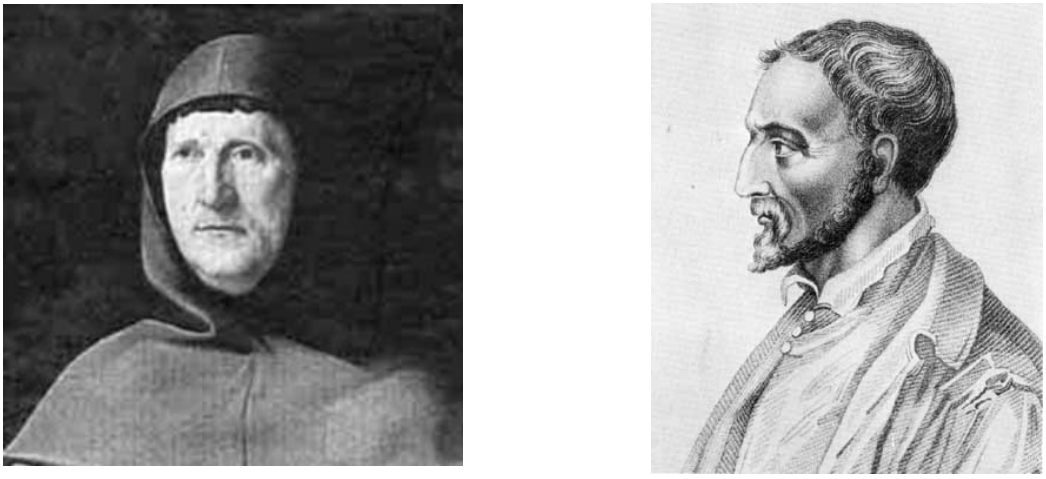
Obrázek 1.3. *Pierre de Fermat a Blaise Pascal*

Mohli bychom říci, že impulsem pro zkoumání náhodných jevů byly hazardní hry. Hry v kostky začínají ve čtrnáctém století doplňovat karetní hry. V počátcích byly ale karty poměrně nákladnou záležitostí, a proto se hrály především mezi lidmi z vyšších vrstev. Hazardní hry se objevují i v již zmiňované korespondenci mezi Pascalem a Fermatem z roku 1654. Více o této korespondenci lze nalézt např. v [17]. V této souvislosti by ale bylo vhodnější mluvit spíše o počtu pravděpodobnosti. Nejednalo se totiž o budování teorie, šlo o řešení konkrétních příkladů. Většina uvedené korespondence se zabývá úlohou o rozdělení

[4] **Richard de Fournival** (1190–1260) – kancléř katedrály v Amiensu.

sázky. Tato úloha se objevuje i u dalších autorů. Ve své nejjednodušší podobě se dá formulovat následovně:

Dva „stejně dobří“ hráči (pravděpodobnost výhry v každé hře je pro oba hráče stejná) hrají určitou sérii her o částku *C*. Vyhrává ten, kdo jako první vyhraje *k* her. Tato série her může být předčasně ukončena, a to ve chvíli, kdy prvnímu hráči chybí do výhry *m* her a druhému hráči *n* her. Úkolem je rozdělit spravedlivě částku *C* mezi tyto dva hráče.



Obrázek 1.4. *Luca Pacioli (vlevo) a Girolamo Cardano*

Poprvé se však úloha o rozdělení sázky neobjevuje v uvedené korespondenci. Její původ je mnohem starší. Až donedávna se předpokládalo, že nejstarší řešení této úlohy podal italský matematik Pacioli[5]. Nicméně O. Ore našel zmínku o úloze v rukopise z roku 1380. Pacioli uvažuje příklad se dvěma hráči, kteří ve stavu 5:3 ukončují sérii a potřebují mezi sebe rozdělit sázku. Pacioli navrhuje, že to mají udělat v poměru k již dosaženým bodům. Na toto chybné řešení zareagoval ve své knize *General trattoto di numeri et misura* Tartaglia[6]. Ovšem ani on se nedobral ke správnému řešení. Ani jeden z nich se totiž nezabývá herní budoucností, tj. tím, kolik her ještě zbývá k vítězství. To odvodil až Cardano[7]. Cardano si povšiml toho, že rozdělení sázky nezávisí na odehraných hrách, ale na počtu her, které chybí každému z hráčů k vítěznému konci. Sázka má být rozdělena úměrně k počtu způsobů, kterými lze vyhrát.

[5] **Luca Pacioli** (1445–1517) – matematik, učil na univerzitách v Pise, Bologni, Florencii.

[6] **Niccolo Fontana Tartaglia** (1499–1557) – italský matematik, známý především díky algebraickému řešení kubických rovnic. Autor několika matematických knih, včetně prvních italských překladů Archimeda a Euclida.

[7] **Girolamo Cardano** (1501–1576) – italský matematik. Cardano vystudoval medicínu. Od r. 1526 pracoval jako lékař, nicméně v r. 1632 se vrátil do Milána a byl jmenován učitelem matematiky. Stal se nejvýznamnějším lékařem v Miláně, jeho služby byly žádané po celé Evropě. Současně vyučoval matematiku. Napsal několik knih z oblasti medicíny, matematiky, astronomie, náhodných her. Jeho nejznámější knihou je nejspíš *Ars Magna*. Od r. 1562 byl profesorem medicíny v Boloně. V r. 1570 byl obviněn z kacířství a dostal zákaz přednášet veřejně a vydávat knihy. Teorií pravděpodobnosti se zabývá se své knize *Liber de Ludo Aleæ*, která ale byla publikována až v roce 1663.

V Pacioliho příkladu jen to pak poměr rozdělení sázky 7:1. Cardano byl podle Halda [28] notorickým hráčem. Tuto svoji hráčskou zkušenost využil ve své knize *Liber de Ludo Aleae*. Kniha byla ovšem publikována až v roce 1663. Není známo, kdy byl rukopis dokončen. Nicméně ve 20. kapitole se objevuje datace 1564. *Liber de Ludo Aleae* je pojednáním o teoretických, praktických a morálních aspektech her založených na náhodě. Většina teorie je podána formou příkladů. Některé případy řeší metodou pokus – omyl a v knize uvádí jak správné, tak špatné řešení. Díky pozdní publikaci tato kniha neovlivnila přímo další vývoj. Nicméně se dá předpokládat, že výsledky Cardana byly koncem šestnáctého století známé matematické komunitě. Více o knize např. v Haldovi [28].

V této době se také objevovaly úlohy, které bychom dnes označili jako kombinatorické, např. kolika způsoby může při současném hodu několika kostkami padnout jistý počet ok. Při prvních pokusech o vyřešení těchto úloh se však nikde nepíše o „pravděpodobnosti“. Tento pojem je nahrazen výrazy „dělení částky“, „šance na výhru“ apod. Otázku, zda se při hodu třemi kostkami vyskytuje součet 9 stejně často jako součet 10, řeší ve svém krátkém pojednání *Sopra le Scoperte dei Dadi* Galileo Galilei[8]. Anglický překlad této práce je možné najít v dodatku knihy [18], kde David datuje Galileovu práci do období 1613–1623.

Obrázek 1.5. *Christiaan Huygens*

[8] **Galileo Galilei** (1564–1642) – astronom, matematik, fyzik, filozof. Od roku 1589 působil na katedře matematiky na univerzitě v Pise, v roce 1592 odchází na univerzitu v Padově. Galileo se proslavil také jako astronom. Sestrojil vylepšené dalekohledy, objevil čtyři měsíce Jupiteru, fáze Venuše, ukázal, že na povrchu Měsíce jsou krátery a pohoří. V roce 1632 vydává spis *Dialogo*, který je vlastně dialogem nezi Aristotelovským a Koperníkovským pohledem na vesmír. O mechanice je dílo *Discorsi e dimostrazioni matematiche intorno a due nuove scienze attinenti alla mecanica ed i movimenti locali* z roku 1638. Kvůli svým názorům se dostává Galileo do sporu s církví.

První publikovanou prací o počtu pravděpodobnosti byl Huygensův[9] spis *De ratiociniis in ludo aleæ* (*O výpočtech v hazardní hře*), který vyšel v roce 1657. Christian Huygens při studiu práv navštívil Paříž a zde se dozvěděl o korespondenci Pascala s Fermatem. Ve své práci Huygens publikoval i správné řešení úlohy o rozdělení sázky.

Svůj spis *De ratiociniis in ludo aleæ* napsal Huygens holandsky, do latiny ho přeložil Franciscus van Schooten – ten ho také vydal jako přílohu se své práci *Exercitationum mathematicarum libri quinque*.

Huygensovo dílo je rozčleněno do čtrnácti témat, tzv. *Propositio*. Zabývá se v nich hrou o nějakou sázku, úlohou o rozdělení sázky a dalšími příklady s herní motivací. V dodatku je připojeno pět neřešených úloh, které ponechává Huygens na procvičení čtenářům. Ani Huygens zde nezavádí pojem „pravděpodobnost“, mluví o tzv. „očekávané výhře“, popř. používá formulaci, že „očekávané výsledky mohu získat stejně snadno“. Za hlavní přínos bychom mohli označit zavedení střední hodnoty diskrétní náhodné veličiny, kterou ovšem nazývá „očekávaná výhra“. V obecných příkladech již využívá pravidla o sčítání a násobení relativních četností – „pravděpodobností“ – nezávislých jevů. Oproti korespondenci Pascala s Fermatem je zde ale podstatný rozdíl v tom, že zatímco Pascal s Fermatem řešili pouze konkrétní úlohy, Huygens se pokouší o zavádění obecných pojmů a postupů. V propositiones I–III uvádí tři tvrzení, které bychom dnes mohli označit jako definice, a z nich potom vychází při řešení svých příkladů. Vždy pod tvrzením je vysvětlení, proč vztah platí. Z našeho hlediska je nejdůležitější třetí tvrzení, které je zobecněním prvních dvou. Zájemci mohou najít otištěnou Huygensovu práci v původní latinské verzi i v českém překladu v [61], odkud jsou i následující překlady do češtiny.

TVRZENÍ III.

Jestliže by počet případů, v nichž mi připadne a, byl roven p, ale počet případů, v nichž mi připadne b, byl roven q, pak za předpokladu, že všechny případy jsou stejně možné, mé očekávání bude mít hodnotu $\frac{pa+qb}{p+q}$.[10]

[9] **Christiaan Huygens** (1629–1695) – narodil se i zemřel v Haagu v Holandsku; zajímal se o broušení optických čoček a konstrukci dalekohledu, pomocí něhož v r. 1655 objevil první měsíc planety Saturn; patentoval první kyvadlové hodiny v r. 1656, které podstatně zlepšily přesnost měření času; sestrojil několik kyvadlových hodin, pomocí nichž určoval zeměpisnou délku při svých plavbách po moři; ve své práci *Horologium Oscillatorium sive de motu pendulorum* z r. 1673 popsal teorii pohybu kyvadla; v r. 1663 byl přijat do Královské společnosti v Londýně; byl předním členem francouzské Akademie věd.

[10] *PROPOSITION III.*

V následujících tvrzeních IV–IX Huygens řeší úlohu o rozdělení sázky. Začíná nejprve řešením případů pro dva hráče. Poté se dostává k řešení v případě tří hráčů. Svoje pojednání o rozdělení sázky uzavírá v tvrzení IX, kde dává návod, jak nalézt obecné řešení úlohy o rozdělení sázky pro libovolný počet stejně dobrých hráčů. Tuto formulaci bychom dnes mohli označit za matematickou větu. Udává rekurentní postup, jak se dostat k řešení úlohy.

TVRZENÍ IX.

Abychom mohli vypočítat podíl každého hráče při libovolně mnoha hráčích, z nichž některému chybí více a jinému méně her, je třeba uvážit, co náleží hráči, jehož podíl má být stanoven, když on sám nebo nějaký jiný hráč vyhraje následující hru. Sečtou-li se takto získané části dohromady a dělí-li se tento součet počtem hráčů, obdrží se hledaný podíl dotyčného hráče.[11]

V následujících tvrzení X–XII se Huygens zabývá hrou v kostky. Nejprve počítá pravděpodobnost hození šestky v prvním hodu. Opět ale nepoužívá slova pravděpodobnost. Ptá se, kolik by musel vsadit ten, kdo se pokouší hodit šestku v prvním hodu a kolik musí vsadit jeho protivník, aby byl poměr sázek spravedlivý. Stejným způsobem uvádí i další příklad – jaká je šance, že hráči v prvním hodu dvěma kostkami padnou na obou kostkách šestky. V poslední úloze z této části řeší otázku, kolika kostkami musí hráč házet, aby pravděpodobnost, že hodí dvě šestky najednou v prvním hodu, byla alespoň jedna polovina.

Poslední dvě tvrzení se týkají jisté hry v kostky pro dva hráče. První z nich hodí současně dvěma kostkami. Vyhrává, pokud padne sedm bodů. Druhý vyhraje, pokud padne deset bodů. Pokud je součet ok na kostkách jiný, hráči si rozdělí výhru rovným dílem. Ve druhé hře je princip obdobný jen s tím rozdílem, že protihráč má první tah a počet ok potřebný k vítězství je jiný.

Celý spis je zakončen pěti neřešenými úlohami, které jsou označeny jako problemata. U první, třetí a páté úlohy jsou uvedeny výsledky.

Si numerus casuum, quibus mihi eveniet a, sit p, numerus autem casuum quibus mihi eveniet b sit q, sumendo omnes casus æquè in proclivi esse: expectatio mea valebit $\frac{pa+qb}{p+q}$. [61, str. 46]

[11] *PROPOSITIO IX.*
Ut tot collusorum, quot quis voluerit, ex quibus uni plures ξ alii pauiores lusu deficiunt, conjusque pars inveniatur, considerandum est, quid illi, cujus partem invenire volumus, degeretur, si vel ipse, vel quislibet reliquorum primum sequentem ludum vinceret. Horum autem partes si in unam summam colligantur,ξ aggregatum per numerum collusorum dividatur, quotiens ostendet unius quæsitam partem. [61, str. 52]

Co se týče samotné teorie pravděpodobnosti, jak uvádí Mačák [61] nebo Stigler [85], o jejím počátku lze mluvit až v souvislosti se spisem Jakoba Bernoulliho *Ars conjectandi*. Tato práce byla napsána již kolem roku 1685. Podle korespondence Leibnize s Jakobem Bernoulli na ní však pracoval ještě před smrtí, tj. v prvních letech osmnáctého století. *Ars conjectandi* bylo vydáno až roku 1713 synovcem Jakoba Bernoulliho Nicolausem (1685–1759). Rod Bernoulliů pocházel z Antverp, kvůli náboženským nepokojům se ale během šestnáctého století několikrát stěhoval, roku 1620 zakotvil jeden z jeho příslušníků v Basileji. V rodu Bernoulliů bylo mnoho činných matematiků. Aby je historikové mezi s sebou odlišili, označují je pořadovými číslicemi. V teorii pravděpodobnosti byli činní Jakob I. (1654–1705), Nicolaus I. (1687–1759) a Daniel I. (1700–1782). Jakob Bernoulli se narodil v Basileji 27. prosince 1654. Fontenelle[12] uvádí, že byl svým otcem předurčen pro duchovní úřad, duchovní studia dokončil a v mládí působil jako duchovní ve Švýcarsku a ve Francii. Od roku 1684 se Jakob Bernoulli zabýval spolu se svým bratrem Johannem diferenciálním a integrálním počtem. Navazovali tak na práci Leibnize. O jejich vztahu ale Stigler říká [85, str. 64], že spíše než spolupracovníky byli soupeři. Když Jakob Bernoulli 16. srpna 1705 zemřel, zůstalo po něm několik nepublikovaných (a v některých případech i nedokončených) matematických prací. Jednou z nich bylo i jeho dílo z teorie pravděpodobnosti *Ars conjectandi*.

Spis *Ars conjectandi* je rozčleněn do čtyř částí. V první z nich uvádí Bernoulli plné znění Huygensova spisu *De ratiociniis in ludo aleæ* se svými komentáři a doplňky. Druhá část je přehledem kombinatoriky, ve třetí části se zahrnuje úlohy z oblasti her. Za nejdůležitější pro rozvoj teorie pravděpodobnosti je považována čtvrtá část, kde Bernoulli poprvé formuluje a uvádí důkaz zákona velkých čísel. V této práci se také poprvé objevuje pokus o definování pojmu „pravděpodobnost“.

V první části Bernoulli v několika případech uvádí u úloh své vlastní řešení naprosto odlišné od toho, které využívá Huygens. V části zabývající se hrou v kostky Bernoulli zobecnil úlohu, kterou řeší Huygens pouze pro současný hod dvěma a třemi kostkami. Ptá se, kolika způsoby může padnout součet 2, 3,..., 12, při hodu třemi kostkami pak 3, 4,..., 18. Bernoulli uvádí rekurentní řešení:

$$Z(n,k) = \sum_{i=1}^{6} Z(n-1,k-i) .$$

[12] **Bernard le Bovier de Fontenelle** (1657–1757) – od r. 1697 sekretář Francouzské Akademie.

Tento vzorec nám říká, že když na n kostkách má padnout součet rovný k, pak na $n - 1$ kostkách musí padnout součet $k - 1$, $k - 2$, ..., $k - 6$. V tabulce 1.1. je uvedena zkrácená verze Bernoulliovy tabulky. Políčka tabulky obsahují hodnoty $Z(n,k)$.

	k													
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	...
1	1	1	1	1	1	1								...
2		1	2	3	4	5	6	5	4	3	2	1		...
3			1	3	6	10	15	21	25	27	27	25	21	...
4				1	4	10	20	35	56	80	104	125	140	...
5					1	5	15	35	70	126	205	305	420	...
6						1	6	21	56	126	252	456	756	...
...							...	...	...	...	...	...	...	

Tabulka 1.1. *Bernoulliho tabulka četností padnutí součtu k ok při hodu n kostkami*

První řádek tvoří samé jedničky, protože při hodu jednou kostkou je pouze jedna možnost, jak může padnout 1, 2, ..., 6. Číslo v k-tém sloupci a n-tém řádku získáme tak, že vezmeme řádek nad, tj. ($n - 1$)-ní řádek, a posčítáme šest hodnot nalevo od k-tého sloupce. Pokud se stane, že některá z těchto šesti hodnot neexistuje, nebo je políčko nevyplněné, bereme místo ní hodnotu nula.

Další zobecnění Huygensovy úlohy můžeme nalézt v tvrzeních X–XII. Bernoulli hledá pravděpodobnost toho, že např. jednička padne v n hodech právě k-krát. Zavádí tak to, co dnes známe pod názvem binomický zákon rozdělení pravděpodobnosti náhodné veličiny. V současném značení zapisujeme pravděpodobnost, že v n nezávislých pokusech nastane nějaký jev právě k-krát:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

kde p značí pravděpodobnost jevu v jednom pokusu.

Jak již bylo řečeno, nejpodstatnější je ale čtvrtá část spisu, ve které se objevuje první formulace a důkaz zákona velkých čísel. Důkaz zde uvádět nebudu, zájemci ho mohou nalézt ve Stiglerově knize [85, od str. 66], pro české čtenáře bude nejspíš přístupnější důkaz v Mačákově knize [61, od str. 83]. Uvedu pouze Bernoulliho formulaci zákona velkých čísel, ovšem v moderní notaci, na kterou jsme zvyklí. Tento zákon lze slovně vyjádřit tak, že se zvyšujícím se počtem nezávislých pokusů se relativní četnosti blíží jejich pravděpodobnosti. Označme X počet pozorovaných úspěchů nebo příznivých případů z celkového počtu N

pozorování a nechť p je neznámý poměr (procento). Pak novodobý zápis Bernoulliho závěru je, že pro libovolné malé kladné číslo ε a dané velké kladné číslo c, může být N určeno tak, že platí:

$$\mathrm{P}\left\{\left|\frac{X}{N}-p\right| \leq \varepsilon\right\} > c\mathrm{P}\left\{\left|\frac{X}{N}-p\right| > \varepsilon\right\}.$$

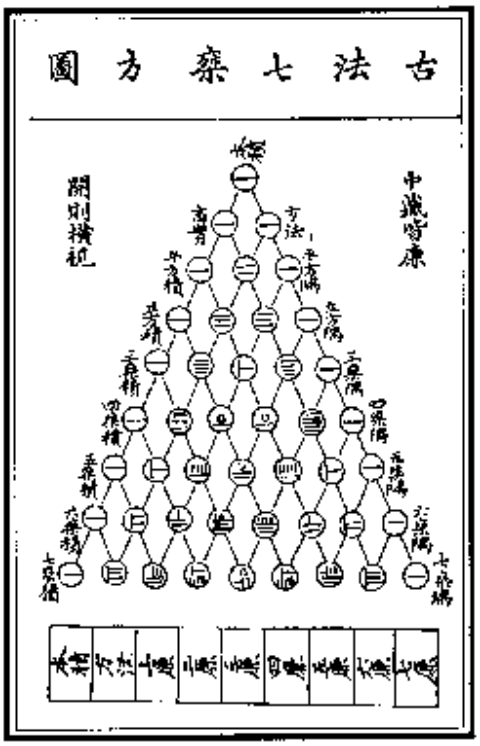
Toto tvrzení může být snadno upraveno do tvaru nyní známého jako Bernoulli slabý zákon velkých čísel:

$$\mathrm{P}\left\{\left|\frac{X}{N}-p\right| > \varepsilon\right\} < \frac{1}{c+}.$$

1.2 Kombinatorika

O prvních kombinatorických úlohách jsem se zmínila již dříve. Kombinatorika jako samostatná část se začíná z matematiky vyčleňovat zhruba v polovině 17. století. Jednou z prvních knih zabývající se kombinatorikou byl Pascalův spis *Traité du triangle arithemétique* (*Pojednání o aritmetickém trojúhelníku*). Tato práce vznikala v období korespondence mezi Pascalem a Fermatem, tedy v roce 1654. Nicméně Pascal sám nikdy dílo nevydal a tak byl spis uveřejněn až o jedenáct let později (roku 1665) po Pascalově smrti. Nejedná se vlastně ani o ucelené dílo, jde spíše o soubor textů. Některé části jsou psané latinsky, jiné francouzsky. V úvodu vydavatel uvádí, že některé texty byly nalezeny mezi Pascalovými listinami již vytištěné, což svědčí o tom, že je zamýšlel vydat. Vypadá to ale, že na nich chtěl ještě pracovat. Možným důvodem jejich nevydání byl nejspíš fakt, že se Pascal posledních osm let svého života zabýval pouze náboženskými otázkami.

Z hlediska teorie pravděpodobnosti je pro nás zajímavá část, kde Pascal řeší úlohu o rozdělení sázky pomocí aritmetického trojúhelníku. Zajímavý je také fakt, že Pascal v důkazu jednoho ze svých tvrzení používá metodu úplné indukce. Díky tomu je považován za objevitele této metody.



Obrázek 1.6. *Pascalův trojúhelník v knize Jaspisové zrcadlo čtyř prvků*

Aritmetický trojúhelník (Pascalův trojúhelník), jak ho známe z dnešní doby má dlouhou historii a Pascal rozhodně nebyl jeho objevitelem. Pascalův trojúhelník (až do řádu $n = 8$) se objevuje v arabské matematice již u al-Karádži[13]. V Číně se Pascalův trojúhelník nejspíš poprvé objevuje v knize Ču Š`t`ia (žil na přelomu třináctého a čtrnáctého století) *Jaspisové zrcadlo čtyř prvků* (viz obr. 1.6.). Pascalův trojúhelník, jak ho známe dnes, se poněkud liší od toho, který uvažuje Pascal ve své práci. Dnes uvažujeme trojúhelník vytvořený z kombinačních čísel. Pascal ale pracuje s obecnějším schématem, které je vytvořeno pomocí určitých pravidel z jednoho daného čísla (generátoru). My dnes uvažujeme pouze aritmetický trojúhelník, kde je generátorem číslo 1. Nicméně Pascal stejně pracuje téměř výhradně s aritmetickým trojúhelníkem, jak ho známe nyní, i když jeho tvrzení platí také pro ostatní trojúhelníky vytvořené jinými generátory. V porovnání s dnešním schématem Pascal používá pro označení jednotlivých prvků trojúhelníku písmena řecké nebo latinské abecedy, což trochu komplikuje pochopení jeho tvrzení. V tabulce 1.2. je ukázka části Pascalova trojúhelníku, kde generátorem je číslo 1 (převzato z [61]).

[13] **Abu Bekr ibn Muhammad ibn al-Husajn Al-Karadží** (953–kolem 1029) – matematik. Znám především díky svým výsledkům v oblasti algebry.

	1	2	3	4	5	6	7	8	9	10
1	G 1	σ 1	Π 1	λ 1	μ 1	δ 1	ζ 1	1	1	1
2	Φ 1	ψ 2	Θ 3	R 4	S 5	N 6	7	8	9	
3	A 1	B 3	C 6	ω 10	ξ 15	21	28	36		
4	D 1	E 4	F 10	ρ 20	Y 35	56	84			
5	H 1	M 5	K 15	35	70	126				
6	P 1	Q 6	21	56	126					
7	V 1	7	28	84						
8	1	8	36							
9	1	9								
10	1									

Tabulka 1.2. *Pascalův aritmetický trojúhelník*

Z našeho pohledu je zajímavé, jak Pascal řešil úlohu o rozdělení sázky pomocí tohoto trojúhelníku. Pascal nejprve formuluje problém a uvádí základní princip řešení (ještě bez použití aritmetického trojúhelníku). Dále ukazuje, jak se úloha řeší pro konkrétní případy. Poté se zabývá zobecněním úlohy. K tomu ale ještě musí uvést několik pomocných tvrzení o vztazích mezi prvky v aritmetickém trojúhelníku. Následuje poslední část, ve které je uveden obecný princip řešení úlohy o rozdělení sázky mezi dva stejně dobré hráče. Pascalův výsledek bychom mohli v dnešní terminologii přepsat následovně: Pokud jednomu hráči chybí do výhry *m* her a druhému *n* her, pak sázku je třeba rozdělit v poměru

$$\sum_{i=m}^{m+n-1}\binom{m+n-1}{i} : \sum_{j=0}^{m-1}\binom{m+n-1}{j}.$$

1.3 Teorie pravděpodobnosti v 18. století

V letech 1660–1680 vzniká díky Newtonovi[14] a Leibnizovi[15] infinitesimální počet. Tyto nové matematické metody využívá již Jakob Bernoulli, především ale Montmort[16] a Moivre[17].

[14] **Isaac Newton** (1642–1727) – anlický matematik, fyzik, astronom a filozof, zakladatel klasické mechaniky a objevitel gravitačního zákona. Vystudoval Trinity College v Cambridge, kde v roce 1668 nastupuje na místo

Montmort navazuje na Huygense. Své poznatky z teorie pravděpodobnosti publikoval v knize *Essay d'Analyse sur les Jeux de Hazard*, která vyšla v Paříži v r. 1708. Obecně lze říci, že jeho kombinatorické úvahy převyšují úroveň kombinatoriky v *Ars Conjectandi*. Montmort používá mj. podmíněné pravděpodobnosti. Nicméně není jisté, zda Montmort v době, kdy svoji knihu publikoval, znal něco z dosud nevydaného *Ars Conjectandi*. Montmortova kniha byla poprvé vydána v roce 1708. Ars Conjectandi bylo poprvé vydáno až v roce 1713 Nicolausem Bernoullim. Je ale možné, že Montmort se dozvěděl něco z dosud nevydaného díla např. od některého z žáků Jakoba Bernoulliho, respektive od Nicolause Bernoulliho, který ze strýcova díla čerpal ve svých pracích z oblasti práva.

První vydání jeho knihy začíná hledáním šancí na výhru v různých karetních hrách. Spis se poměrně dobře čte, protože Montmort u každé hry uvádí její pravidla. Po stanovení pravidel řeší jednoduché případy metodou podobnou Huygensovu řešení a poté postupuje k obecnému řešení, které je sice správné, ale ne vždy je tak zřejmé. Některá Montmortova řešení jsou uvedena ve formě nekonečných řad. A nejspíš poprvé se zde objevuje při řešení problémů z teorie pravděpodobnosti jako limitní případ exponenciální funkce.

Montmortova kniha vyšla r. 1713 ve druhém vydání, kde byly úlohy a problémy z prvního vydání podstatně rozvinuty. Oproti prvnímu vydání je rozsah knihy více než

---

profesora matematiky. V roce 1705 je královnou Annou povýšen do šlechtického stavu. Newton vytvořil základ klasické mechaniky, je autorem tzv. Newtonových pohybových zákonů. V roce 1687 vydává *Philosiphiae naturalis principia mathematica*. Teorie gravitačního zákona je rozvinuta v jeho spisu *De motu Corporum* z roku 1685. Společně s Leibnitzem vytváří diferenciální a integrální počet. Newton se věnuje také optice.

[15] **Gottfried Wilhelm Leibnitz** (1646–1716) – studoval práva. Vstoupil do diplomatických služeb mohučského kurfiřta, díky tomu působil čtyři roky (1672–1676) v Paříži. Později působil jako knihovník a dvorní rada v Hannoveru. Do dějin matematiky se zapsal jako jeden ze zakladatelů infinitesimálního počtu.

[16] **Pierre-Rémond de Montmort** (1678–1719) – narodil se v Paříži, jeho otec chtěl, aby studoval práva. Nicméně jeho syn raději odešel do zahraničí. Když bylo Montmortovi 22 let, jeho otec zemřel a zanechal mu velké bohatství. Montmort studoval u Malebranche filozofii a u Descartesa fyziku. Studoval také matematiku, především algebru a geometrii. Do r. 1706 byl kanovníkem Notre-Dame. Roku 1708 publikuje výsledky svých výzkumů z oblasti pravděpodnosti v *Essay d`Analyse sur les Jeux de Hazard*. Druhého vydání se kniha dočkala roku 1713. Montmort velkou část svého života strávil na Château de Montmort. V roce 1715 byl Montmort zvolen členem Royal Society a o rok později členem Académie Royal des Sciences.

[17] **Abraham De Moivre** (1667–1754) – narozen ve Vitry (poblíž Paříže). Studoval na College de Harcourt v Paříži. Roku 1685 byl ve Francii zrušen edikt nanteský. De Moivre jako protestant odmítl přestoupit na katolickou víru. Kvůli tomu byl tři roky vězněn, po propuštění odešel do Anglie. Jako cizinec však nezískal místo na univerzitě, proto vyučoval matematiku soukromě. De Moivre si přivydělával tím, že radil hráčům hazardních her a soukromým pojišťovatelům. Zabýval se algebrou, geometrií, teorií nekonečných řad a teorií pravděpodobnosti. V r. 1718 publikoval práci *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play*. Kniha vyšla ještě dvakrát v rozšířené podobě v roce 1738 a 1756. V r. 1730 vydává knihu *Miscellanea Analytica*. Z jeho výsledků je znám např. Stirlingův vzorec pro odhad hodnoty $n!$ pro velká n. Aproximoval binomické rozdělení limitním rozdělením - dnes známým pod názvem Gaussovo normální rozdělení. De Moivre je dnes znám díky vztahu $(\cos\alpha + \sin\alpha)^n = \cos n\alpha + \sin n\alpha$, který je po něm nazván. V tomto případě však De Moivre pouze uvedený vztah používal, nicméně nikdy ho v této podobě nezapsal, ani nebyl jeho objevitelem. Slovní formulaci poprvé uvedl François Viète (1540–1603). Současná podoba vztahu je připisována Rogeru Cotesovi (1682–1716). De Moivre také aproximoval binomické rozdělení Poissonovým rozdělením.

dvojnásobný a odráží jistý pokrok v myšlení autora v této oblasti. Druhé vydání vyšlo s pomocí Montmortova přítele Nicolause Bernoulliho. Nicolaus Bernoulli se setkal s Montmortem, když navštívil Paříž. Od té doby spolu udržovali čilou korespondenci. Velkou část knihy (přes sto stran) tvoří korespondence mezi Montmortem a Johannem a Nicolausem Bernoullim. Montmort zřejmě publikoval takové množství korespondence, protože chtěl ukázat na výsledky, jejichž autorem byl Nicolaus. Nicolaus se podílel hlavně na zobecnění problémů, které navrhl buď Montmort nebo Nicolausův strýc Johann. Nicolaus také předložil problém, který se stal známý pod názvem Petrohradský problém[18].

Infinitesimální počet používal v pravděpodobnostních problémech i Abraham De Moivre. Moivre se zabýval teorií nekonečných řad, geometrií a algebrou. Protože se mu v Anglii jako cizinci nepodařilo získat místo na univerzitě, vyučoval matematiku soukromě. Tato práce však byla časově náročná a špatně placená. V pozdějších letech si začal přilepšovat vydáváním knih a poskytováním rad hazardním hráčům a pojišťovatelům. Teorií pravděpodobnosti se tak nejspíš začal zabývat z praktických důvodů.

1.4 Demografie a pojistná matematika

Již na začátku této kapitoly jsem uvedla možné důvody pro rozvoj teorie pravděpodobnosti. Jedním z nich byl i rozvoj demografie [53] a pojišťovnictví. Počátky pojistné matematiky lze najít v Římě. Nejstarší známé úmrtnostní tabulky pochází z roku 211 a jejich autorem byl Ulpianus[19].

V Evropě se první seznamy zemřelých pořizovaly v Anglii v r. 1532 kvůli morové epidemii. O více jak půl století později v r. 1592 začínají v Londýně vycházet pravidelné týdenní seznamy úmrtí. Seznamy obsahovaly příčinu úmrtí – mor nebo jiné důvody. Nejspíš tak měly posloužit jako ukazatel příchodu případné morové epidemie.

První publikovanou prací z pojistné matematiky byla kniha Johna Graunta[20] *Natural and political observations mentioned in a following index and made upon the bills of*

[18] **Petrohradský problém (paradox)** – hráč hází mincí tak dlouho, dokud nepadne první orel. Počáteční výhra a se po každém hodu zvyšuje a-krát, tzn. po n hodech je výhra rovna a^n. Při použití Huygensova vzorce pro výpočet střední hodnoty dostaneme očekávanou střední hodnotu výhry jako součet všech možných výher vynásobených jejich pravděpodobnostmi, tj. $\sum_{i=1}^{\infty} a/2 \rightarrow$ pro každé $a \geq$!. Problémem je výše poplatku A za vstup hráče do hry. Výška tohoto vstupního poplatku závisí na ochotě hráče riskovat. Paradoxem je to, že i když je možné získat neomezeně vysokou částku, hráči většinou nejsou ochotni platit více než $A = 20$ za hru.

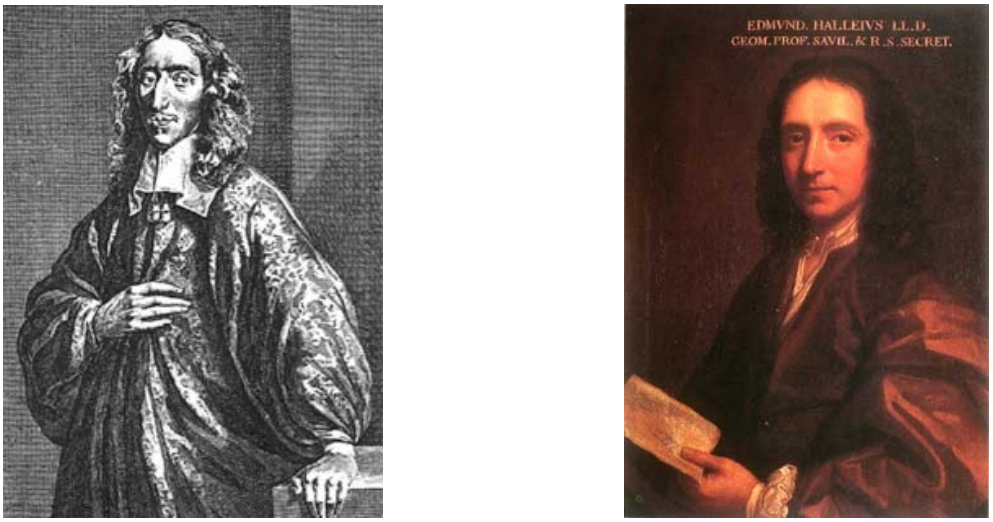
[19] **Domitius Ulpianus** (zemřel r. 228) – římský právník za doby vlády císaře Alexandra Severa.

[20] **John Graunt** (1620–1674) – obchodník žijící v Londýně. Proslavil se až svojí knihou *Natural and political observations mentioned in a following index and made upon the bills of mortality*, která vyšla v několika vydáních (1662, 1663, 1665, 1676). Byl zvolen členem Royal Society.

mortality, která vyšla v Anglii v roce 1662. Tato studie je založená právě na týdenních seznamech úmrtí v Londýně. V jeho knize se objevuje např. poměr mužů a žen, poměr bojeschopných mužů, poměr žen v plodném věku k celkovému počtu apod. Grauntovy úmrtnostní tabulky např. ukazují, že 16 let se dožívá pouze čtyřicet procent pokřtěných dětí. Významným výsledkem Grantovy knihy je podstatně nižší, než se obecně předpokládalo, a patrně správný odhad počtu obyvatel Londýna.

Další významnou postavou v této oblasti byl politický vládce Nizozemska Johann de Witt[21]. Jeho spis *Waarde van lyf-renten naer proportie van Losrenten* (překládaný do angličtiny jako *Treatise on Life Annuities*) vyšel v roce 1671. De Witt pomocí teoretických úmrtnostních tabulek vypočítává hodnotu doživotní renty poskytované státem.

Teorií doživotních důchodů se zabýval i anglický astronom Edmond Halley[22]. V jeho článku *An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain to price of annuities upon lives*, který vyšel v roce 1693, je mj. použito geometrického pohledu na pravděpodobnostní úlohu. Halley zvolil Vratislav kvůli tomu, že oproti Londýnu zde byl stabilnější počet obyvatel. Na základě vratislavských tabulek sestavil úmrtnostní tabulky a z nich pak vypočítal průměrnou cenu životní renty pro celou populaci.



Obrázek 1.7. *Johann de Witt (vlevo) a Edmond Halley*

[21] **Johann de Witt** (1625–1672) – státník, matematik a právník. Byl Velkým pensionářem Hollandu.
[22] **Edmond Halley** (1656–1742) – anglický astronom a fyzik. Objevil planety, které dnes známe pod jeho jménem. Od r. 1720 byl královským astronomem v Greenwichi.

Kapitola 2:

Historické kořeny metody nejmenších čtverců

2.1 Metoda nejmenších čtverců

Metoda nejmenších čtverců byla nejúspěšnější z raných metod kombinování nekonzistentních rovnic. Důvodem úspěchu této metody byl fakt, že byla založena na snadno pochopitelných objektivních kritériích. Pro kombinování nezávislých pozorování jedné veličiny se na konci sedmnáctého století začíná používat aritmetický průměr. Na začátku osmnáctého století se objevuje zobecněná verze aritmetického průměru, kde jsou jednotlivým pozorováním přiřazeny váhy. V roce 1757 se objevuje tzv. „Boscovichova metoda“, která však po uvedení metody nejmenších čtverců (1805) upadla v zapomnění. Abychom mohli lépe pochopit, co vedlo k objevu metody nejmenších čtverců, je potřeba se podívat, jaké problémy řešili vědci ve století, které objevu předcházelo.

Jako hlavní vědecké problémy osmnáctého století bychom mohli označit:

(i) Určení a matematický popis pohybů Měsíce.

(ii) Vysvětlení odvěké nerovnosti, která je pozorována v pohybech Jupiteru a Saturnu.

(iii) Určení tvaru Země.

První narážka na to, že Země není perfektní koule se objevila již v roce 1672 u Jeana Richera[23]. Zjistil, že kyvadlo blízko rovníku je méně ovlivněno gravitační přitažlivostí, než stejné kyvadlo v Paříži. Isaac Newton v *Principech* (1678) ukázal, jak rotace Země může způsobovat zplošťování Země na pólech. Země má tak tvar rotačního elipsoidu. Newton navíc odhadl její zploštění a eliptičnost, tj. podíl, kterým poloměr na rovníku překračuje poloměr na pólu, jako 1/230. Oproti tomu ředitel Royal Observatory v Paříži Domenico Cassini si myslel, že Země je protáhlý sféroid, zploštělá na rovníku, ne na pólech.

Dvěma hlavními metodami k určení tvaru Země bylo pozorování pohybu kyvadla a měření délky oblouku stejného úhlu na tomtéž poledníku na různých vzdálených místech. Určení tvaru Země z měření oblouku ale vyžadovalo spolupráci mnoha vědců a náročnou několikaměsíční práci. Myšlenkou bylo změřit délku stupně zeměpisné šířky[24] na dvou nebo

[23] **Jean Richer** (1630–1696) – francouzský astronom. V roce 1666 byl jmenován členem akademie věd. V roce 1671 byl vyslán, aby v Cayenne konal astronomická pozorování. A právě zde pozoroval, že kyvadlo je méně ovlivněno gravitací, než stejné kyvadlo v Paříži. Výsledky svých tříletých pozorování v Cayenne shrnul v práci *Observations astronomiques et physiques faites en l´isle de Cayenne*.

[24] **Zeměpisná šířka** určuje polohu na povrchu Země směrem k severu nebo k jihu od rovníku. Jedná se o úhel, který svírá rovina rovníku s přímkou, procházející středem Země a příslušným bodem na povrchu Země. Měří se ve stupních. Body se stejnou zeměpisnou šířkou se nazývají rovnoběžky.

více dostatečně vzdálených zeměpisných šířkách. Pokud by délka stupně blízkého rovníku byla kratší než délka stupně u pólu, byl by tvar Země zploštělý. Rozdíl mezi těmito dvěma délkami by mohl být použit k vypočítání zploštění.

Vztah mezi délkou oblouku a zeměpisnou šířkou lze odvodit z geometrie. My budeme uvažovat pouze krátké oblouky (ty jediné je možné změřit). Pro tyto oblouky pak platí jednoduchá aproximace – pokud a je délka jednoho stupně zeměpisné šířky, mající střed v zeměpisné šířce α a naměřená podél poledníku, pak pro aproximaci platí

$$a = z + y \sin^2 \alpha \ ,$$

kde z je délka stupně na rovníku, y přebytek (nebo naopak nedostatek) stupně na Severním pólu oproti délce stupně na rovníku. Domenico Cassini se svým synem Jacquesem měřili francouzský oblouk před rokem 1720. Na základě svých měření se přiklonili k tvrzení, že Země je protáhlá. Postavili se proti výsledkům Newtona. Nicméně omezený rozsah zeměpisné šířky (pouhých 9°) a možná omezená přesnost měření byly příčinou, proč tento závěr nebyl všeobecně přijat. V roce 1735 následovala výprava francouzské akademie do Peru a do Laponska. Šlo o to změřit oblouky blízko rovníku a kolem 66° zeměpisné šířky a srovnat je s měřeními blízko Paříže. Protože místa se nacházejí na dostatečně vzdálených zeměpisných šířkách, bylo důvodné očekávat vyšší vypovídací sílu těchto měření. Vědci francouzské akademie na základě svých měření vyvrátili Cassiniho hypotézu a přiklonili se na stranu Newtona, že Země je zploštělá. Nicméně zůstal úkol určit eliptičnost nebo velikost zploštění. Různé dvojice oblouků totiž dávaly rozdílné hodnoty.

Podívejme se teď na další ze zmiňovaných problémů, a to pozorování neperiodických odchylek v pohybu planet Jupiteru a Saturnu. V roce 1676 astronom Halley potvrdil podezření, že Jupiter a Saturn mají sklon k nepatrným dlouhodobým nerovnoměrnostem ve svých pohybech. Po porovnání aktuálních pozic Jupiteru a Saturnu s tabulkovými hodnotami získanými v několika staletích bylo zjištěno, že průměrný pohyb Jupiteru je zrychlující se, zatímco u Saturnu zpomalující se. Nicméně Halley nebyl schopen toto své tvrzení podepřít matematickou teorií.

2.2 Johann Tobias Mayer

Johann Tobias Mayer[25] pracoval jako kartograf a astronom. V letech 1748–1749 provedl velké množství pozorování Měsíce. Mayer odhalil kývavý pohyb Měsíce, který dnes

[25] **Johann Tobias Mayer** (1723–1762) – matematik samouk a astronom. Jeho první důležitá astronomická práce se zabývala librací Měsíce. Proslavil se především lunárními a solárními tabulkami, publikovanými v *Transactions*. Publikoval také dvě geometrické práce v roce 1746. Působil na univerzitě v Göttingenu. Od roku

nazýváme librace. Vyvrací tím ve své době převládající názor, že ze Země vidíme stále stejnou polovinu Měsíčního povrchu. Díky libraci však můžeme pozorovat i část odvrácené strany Měsíce. Ze Země tak vidíme 59% povrchu Měsíce. Více o libraci v [5] a [50].

Mayer si všímal několika významných lunárních vlastností a v roce 1750 ukázal [63], jak tato data mohou být použita k určení různých charakteristik oběžné dráhy Měsíce. Jako význačný bod na Měsíci si Mayer vybral kráter Manilius. Uvažoval pak následující lineární závislost

$$\beta - (90° - i) = \alpha \sin\theta \cos(g - \ell),$$

kde α je úhel mezi pólem Země a pólem Měsíce, β označuje selenografickou[26] šířku vybraného útvaru na Měsíci (Mayer použil kráter Manilius), k ekliptikální délku výstupního uzlu oběžné dráhy Měsíce, $k + \theta$ ekliptikální délku vzestupného ekvinokcia, g ekliptikální délku a 90º - h ekliptikální šířku vybraného útvaru na Měsíci. Hodnoty g a h jsou pozorovány a k lze nalézt v lunárních tabulkách.

Mayer provedl 27 pozorování, a tím získal 27 rovnic pro tři neznámé α, β a $\alpha\sin\theta$. Průkopnické je to, jak se Mayer s přeurčenou soustavou rovnic vypořádal. Rozdělil 27 rovnic do tří skupin po devíti pozorováních. V každé skupině sečetl všech devět rovnic. Tím získal pouhé tři rovnice pro tři neznámé. Jak ale probíhalo rozdělování rovnic do skupin? Mayer si všímal koeficientů u neznámých α. První skupiny tvořilo devět rovnic s největšími kladnými hodnotami neznámé α. Ve druhé skupině pak bylo devět rovnic s nejmenšími hodnotami (zápornými) koeficientu α. A konečně ve třetí skupině měly být rovnice s největšími hodnotami u neznámé $\alpha\sin\theta$. Tento popis třetí skupiny není úplně přesný, protože Mayer do třetí skupiny dává devět zbylých, dosud nepoužitých rovnic. A toto přesně neodpovídá rovnicím, které mají největší hodnoty koeficientu $\alpha\sin\theta$. Nicméně i tak je tento způsob kombinování rovnic průkopnický a zaslouží si naši pozornost. Rozdělení tímto způsobem maximalizuje rozdílnost koeficientu α a tím jsou sumace rovnic vzhledem k α tak velké, jak je to jen možné.

Mayer tedy ze tří rovnic může určit hodnotu neznámých α, β a θ. Pokračoval dále a uvažoval, jak to bude s přesností těchto hodnot. Nešlo mu o žádnou obecnou analýzu chyby, problém řešil empirickým stanovením přesnosti. Mayer ukázal, jak by vypadalo řešení pro tři neznámé, kdyby použil pouze tři rovnice. V tomto případě mu vychází výsledek α = 1º40´. Když však použil všech 27 rovnic, dopočítal α = 1º30´. Z toho, že měl k dispozici devětkrát

---

1754 až do své smrti byl ředitel observatoře. Jeho rukopisy publikoval roku 1775 G. C. Lichtenberg pod názvem *Opera inedita*.

[26] **Selenografické souřadnice** jsou obdobou geografických souřadnic pro Měsíc.

tolik pozorování, usuzuje, že výsledek by měl být devětkrát přesnější. Správnou hodnotu označuje jako α = 1º30´ ± x, kde x označuje chybu (odchylku skutečné hodnoty od hodnoty určené pomocí 27 rovnic). Chyba při použití tří rovnic je tak 10´ ± x. Z rovnice

$$\pm x : 1/27 = (10' \pm x) : 1/3$$

dostává Mayer hodnotu ± x = 1,25´. Mayer tak na závěr konstatuje, že skutečná hodnota α může být o 1´ nebo 2´ menší nebo větší než hodnota 1º30´ .

Obrázek 2.1. *Tobias Mayer a Leonard Euler*

2.3 Leonard Euler a zkoumání neperiodických odchylek v pohybech Saturnu a Jupitera

Leonard Euler (1707–1783) provedl matematickou analýzu pohybu Saturnu a Jupiteru [19]. Euler se primárně zaměřil na planetu Saturn. Připustil, že oběžná dráha Saturnu i Jupiteru je eliptická, nicméně že roviny oběžných drah nejsou stejné.

Po dokončení matematické analýzy si chtěl Euler své výsledky ověřit empiricky. Měl k dispozici 75 kompletních skupin pozorování z let 1582 až 1745. Jeho formule obsahovala celkem osm neznámých. Z pozorování tedy sestavil celkem 75 rovnic pro osm neznámých. Euler byl především exaktní matematik, a proto nepřijal myšlenku, že by kombinací rovnic mohl získat přesnější výsledek. Euler pracoval s malými skupinami rovnic (většinou tolika skupinami, kolik bylo neznámých). A uznával numerické výsledky pouze v tom případě, že různé skupiny rovnic dávaly téměř totožné výsledky.

Když porovnáme řešení Mayera a Eulera, můžeme konstatovat, že Mayer přistupoval k problému jako praktický astronom. Rozdělil pozorování do skupin tak, aby v jedné skupině byla pozorování, vytvořená v podstatě za stejných podmínek. Euler byl však především matematik. Vycházel z předpokladu, že pozorování byla pořízena za různých, neznámých

podmínek. Mayer nahlížel na chyby jako na náhodné. Byl přesvědčen, že kombinací jednotlivých pozorování zvýší přesnost výsledku v poměru k počtu kombinovaných rovnic. Euler však nepřijal tento statistický pohled na věc, že náhodné chyby mají tendenci se navzájem vyrušit.

2.4 Roger Joseph Boscovich

V roce 1755 Boscovich[27] spolu a Christopherem Mairem (1697–1767) publikoval výsledky měření poledníkového úhlu pod názvem *De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus* [7]. V následných Boscovichových analýzách těchto dat nacházíme první úspěšné řešení nekonzistence různých obloukových měření. Jeho metoda byla založena na minimalizování váženého součtu absolutních hodnot odchylek měření od hledané hodnoty. Vážení používá proto, že jednotlivá měření se lišila svou přesností. Tím Boscovich pokládá základ robustních statistických metod. My se nyní na jeho metodu podíváme podrobněji.

Obrázek 2.2. *Podobizna R. G. Boscoviche na jugoslávské poštovní známce (vlevo) a chorvatské bankovce*

Když se Boscovich pustil do problému určení eliptičnosti Země, setkal se jen s omezeným úspěchem. K podstatnému zlepšení se dopracoval až později. Boscovich si byl vědom toho, že potřebuje měření na dostatečně vzdálených místech. Protože jinak velice malé chyby v určení oblouku by se mohly značně zvětšit při kombinování jejich dvojic. Proto použil pět měření, která byla pořízena ve vzdálených lokalitách, a dalo se u nich předpokládat,

[27] **Roger Joseph Boscovich** (1711–1787) – jezuitský kněz. Narodil se roku v 1711 v Raguse, dnešním Dubrovníku. Studoval v Římě a většinu svého dospělého života prožil v Itálii nebo Paříži. V roce 1740 byl v Římě jmenován profesorem matematiky. O dvacet let později odjel do Londýna, kde zůstal asi šest měsíců. Setkal se zde s několika významnými anglickými vědci, mj. s Benjaminem Franklinem. Boscovich se zabýval astronomií, optikou, gravitací a trigonometrií.

že jsou přesná (viz tab. 2.1.). Délka je v originále uváděna v jednotkách toise, kde 1 toise (do češtiny bychom mohli přeložit jako jeden sáh, dnes už nepoužívaná míra) je přibližně 6,39 stopy, tj. 1,947 metru.

Místo	Zeměpisná šířka (α)	Délka oblouku (v sázích)	$\sin^2\alpha \quad 0^4$
(1) Quito	0°0′	56,751	0
(2) Mys Dobré Naděje	33°18′	57,037	2,987
(3) Řím	42°59′	56,979	4,648
(4) Paříž	49°23′	57,074	5,762
(5) Laponsko	66°19′	57,422	8,386

Tabulka 2.1. *Boscovichovy podklady k poledníkovým obloukům, zdroj: Boscovich a Maire [8, str. 482]*

Boscovich se tak dostává k pěti rovnicím $a_i = z + y\sin^2\alpha$, kde a_i jsou délky příslušných oblouků (v jednotce sáh na stupeň), α jsou zeměpisné šířky bodu ve středu oblouku. Neznámé proměnné *y* a *z* vyjadřují přebytek jednoho stupně oblouku na pólu oproti jednomu stupni na rovníku a délku stupně na rovníku. Boscovich věděl, že každé dvě z těchto rovnic mohou být použity k vypočtení eliptičnosti a polárního přebytku. To také udělal. Pro každou z deseti dvojic vypočítal polární přebytek *y* a eliptičnost. Pro výpočet eliptičnosti použil Boscovich přibližný vzorec 1/eliptičnost$= \cdot z / y$. Výsledky jsou shrnuty v tabulce 2.2.

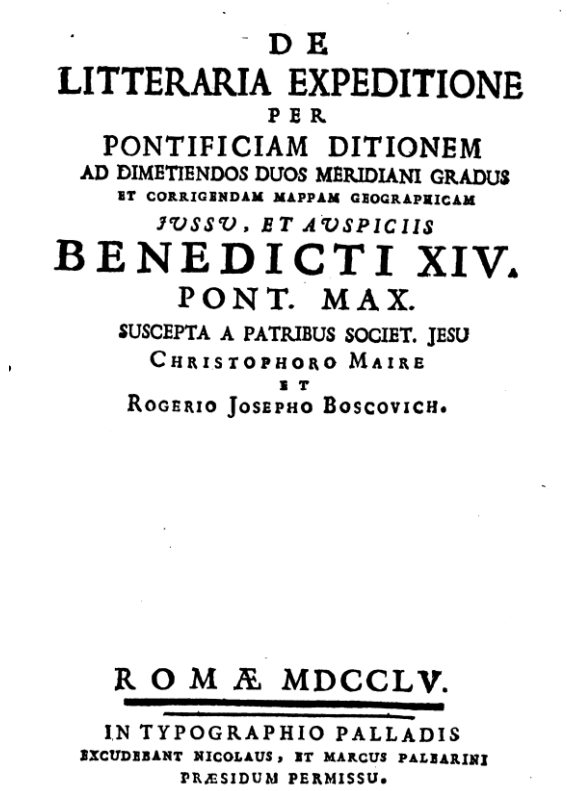
Dvojice	Polární přebytek (y, v sázích)	Eliptičnost	Dvojice	Polární přebytek (y, v sázích)	Eliptičnost
1, 5	800	1/213	2, 4	133	1/128
2, 5	713	1/239	3, 4	853	1/200
3, 5	1,185	1/144	1, 3	491	1/347
4, 5	1,327	1/128	2, 3	-350	-1/486
1, 4	542	1/314	1, 2	957	1/78

Tabulka 2.2. *Boscovichovy výsledky výpočtů eliptičnosti a polárního přebytku pro všech deset dvojic pozorování, zdroj: Boscovich a Maire [7, str. 501]*

Tímto způsobem ale dostal Boscovich deset různých hodnot eliptičnosti. Protože neznal žádný lepší způsob, jak se s takovým přebytkem hodnot vypořádat, zprůměroval tyto hodnoty a dostal se k eliptičnosti 1/155 [28]. Nicméně tato hodnota se mu zdála příliš velká, a

[28] Hodnota zploštění Země udávaná dnes je 1/298,26.

proto se rozhodl zamítnout dvojice (1, 2) a (2, 3), neboť byly příliš odlišné od ostatních. Znovu spočítal průměr a tentokrát došel k výsledku 1/198. Ani to se ale Boscovichovi nezdálo uspokojivé. Místo, aby vzal jednu z těchto hodnot, popř. nějaké jejich zprůměrování za výslednou hodnotu eliptičnosti, zaměřil se na odchylky mezi zjištěnými hodnotami eliptičnosti mezi jednotlivými dvojicemi. Zamítá hypotézu o tom, že Země je elipsoid. Kdyby totiž Země byla elipsoid, všechna měření musí docházet ke stejné eliptičnosti. Ale v tomto případě mu připadají rozpory mezi jednotlivými měřeními natolik významné, že podle něj nepřipadá v úvahu, aby Země byla elipsovitého tvaru.


DE
LITTERARIA EXPEDITIONE
PER
PONTIFICIAM DITIONEM
AD DIMETIENDOS DUOS MERIDIANI GRADUS
ET CORRIGENDAM MAPPAM GEOGRAPHICAM
JUSSU, ET AUSPICIIS
BENEDICTI XIV.
PONT. MAX.
SUSCEPTA A PATRIBUS SOCIET. JESU
CHRISTOPHORO MAIRE
ET
ROGERIO JOSEPHO BOSCOVICH.

ROMÆ MDCCLV.
IN TYPOGRAPHIO PALLADIS
EXCUDEBANT NICOLAUS, ET MARCUS PALEARINI
PRÆSIDUM PERMISSU.

Obrázek 2.3. *Titulní strana práce z roku 1755*

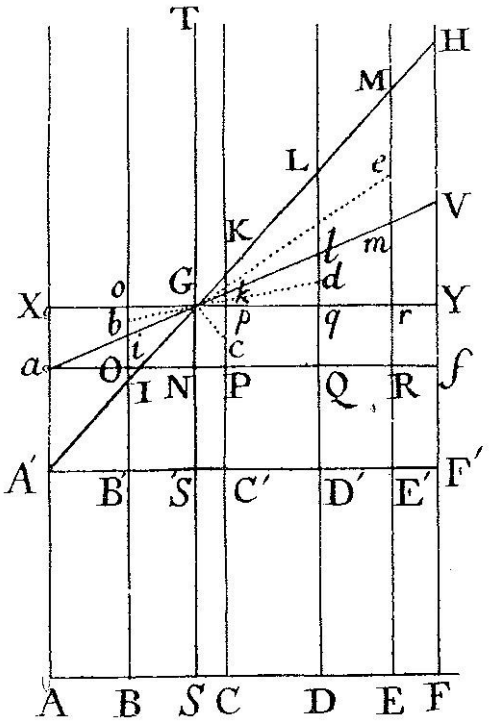
Ani s tímto svým závěrem se ale Boscovich nesmířil a rozhodl se dále pokračovat v práci. V roce 1757 (tedy o dva roky později) publikoval poměrně stručné pojednání o zcela novém způsobu kombinování nekonzistentních obloukových měření. A v roce 1760 dává Boscovich úplný popis tohoto principu společně s návodem, jak ho využít v praxi. Vše dokumentuje na příkladu pěti měření poledníkového oblouku z roku 1755. Rozdíl oproti přístupu Mayera je v tom, že tam, kde Mayer postupuje ad hoc, Boscovich používá určitý obecný postup. Formuluje pravidla, která by měl mít průměr založený na kombinaci obloukových měření (ne obyčejný aritmetický průměr) mít. Pro každou naměřenou délku oblouku udělal korekci. Tyto korekce musely splňovat tři podmínky:

(i) Jejich rozdíly musí být úměrné rozdílům mezi převráceným sinům dvojnásobků jejich zeměpisných šířek. Tato podmínka je nazývána zákonem rovnováhy, což vyžaduje eliptický tvar.

(ii) Součet pozitivních korekcí musí být stejný jako součet negativních korekcí. Tato podmínka je odůvodněna tím, že kladné i záporné odchylky od skutečné hodnoty jsou stejně možné.

(iii) Součet všech korekcí (pozitivních i negativních) musí být co nejmenší možný při splnění dvou předchozích podmínek.

Boscovich však neformuluje tyto podmínky analyticky, ale pouze slovně a ve své práci dává pouze geometrický nebo mechanický popis, jak problém řešit. Nicméně velkou výhodou navrženého řešení je jeho obecnost. Diskuzi nad řešením Boscovich doprovází diagramem, ve kterém popisuje svůj algoritmus (obr. 2.4.).



Obrázek 2.4. *Boscovichův algoritmus, převzato z [8]*

Vodorovná osa AF udává $\sin^2\theta$, kde θ je zeměpisná šířka středu oblouku, vertikální osa AX udává délku oblouku v sázích na stupeň. Pět oblouků je označeno jako *a*, *b*, *c*, *d*, *e*; G je těžiště. Délku úsečky AF považujeme za jednotku. A označuje počátek, A , B, C, D, a E jsou hodnoty $\sin^2\theta$ vyznačené na jednotkovém intervalu. Délky (v sázích na stupeň) příslušných naměřených oblouků jsou v diagramu vyznačeny úsečkami A*a*, B*b*, C*c*, D*d* a E*e*.

Boscovich řeší problém, jak naleznout přímku A´H takovou, že korekce *a*A´, *b*O, *c*K, *d*L a *e*M budou splňovat všechny tři požadované podmínky. První podmínka je splněna tím, jak jsme zachycovali do grafu jednotlivé vzdálenosti. Druhá podmínka (podmínka rovnováhy) vyžaduje, aby přímka, kterou hledáme, procházela těžištěm bodů *a*, *b*, *c*, *d*, *e*. V diagramu je těžiště zakresleno jako bod G. Stačí tedy splnit poslední třetí podmínku. Boscovich si do grafu zakreslil přímku SGT. Tu „ukotvil" v bodě G a otáčel s ní po směru hodinových ručiček. Teď šlo jen o to, při kterém otočení dostáváme hledanou přímku. Jak přímka rotovala, suma korekcí (součet všech korekcí bez znaménka, tj. součet jejich absolutních hodnot) bude klesat, až dosáhne svého minima, a poté začne zase růst. Boscovich říká, jak dlouho je potřeba s přímkou otáčet. Korekce pro jednotlivé oblouky se mění při otáčení přímky v poměru k vzdálenostem AS, BS, SC, SD a SE. Je potřeba v rotaci pokračovat, dokud přímka neprotne pět bodů tak, že alespoň polovina součtu těchto pěti vzdáleností AS + BS + SC + SD + SE byla dosažena vzdálenostmi, které odpovídají protnutým bodům. Předtím, než bude tato přímka nalezena, součet korekcí bude při rotování klesat. Po dosažení hledané přímky začne suma korekcí růst.

Vše, co Boscovich uveřejnil o své metodě, je slovní popis a příklad, který jsem uvedla (ve zjednodušené podobě) v minulých odstavcích. Přidal ještě diskuzi s Mairem. V této debatě používá metodu na devět naměřených oblouků a poté aplikování metody po zamítnutí tří nejvíce „podezřelých" měření, která se nejvíce lišila od ostatních. Boscovich tak dochází k závěru, že Země je nepravidelného tvaru, ale podobá se elipsoidu. Boscovich dále neuvádí žádné vlastnosti své metody, neuveřejňuje její další rozvoj, ani se nepokouší o analytickou formulaci. Zabývá se pouze problémem tvaru Země, svoji metodu neaplikuje na další problémy. Uvádí sice, že metoda je obecná a dá se použít i v jiných situacích, nicméně neuvádí žádné příklady. Boscovichovou metodou se zabýval např. i Gilbert Bassett a Roger Koenker [3].

2.5 Pierre Simon de Laplace

Na nerovnosti v pohybech planet Saturnu a Jupiteru naráží v kurzu historie v roce 1787 Laplace[29]. Navrhuje rozšíření Mayerovy metody zacházení s nekonzistentními lineárními rovnicemi. Navíc se Laplaceovi podařilo ukázat, že pohyby Saturnu a Jupiteru jsou

[29] **Pierre Simon de Laplace** (1749-1827) – narodil se v Normandii a jeho život zahrnoval Napoleonskou éru. Laplace byl jedním z nejvýznamnějších francouzských vědců. Byl členem Akademie věd, profesorem na École Militaire a École Normale, ministrem vnitra. Ve své vědecké práci se zabýval mechanikou nebeských těles, matematikou, pravděpodobností a fyikou. Mezi jeho nejdůležitější spisy patří *Traité de mécanique Céleste* (1799–1805) a *Théorie analytique des probabilités* (1812).

ve skutečnosti periodické – s velmi dlouhou periodou. Své výsledky shrnuje ve spisu *Théorie de Jupiter et de Saturne*, vydaném Akademií věd v roce 1787.

Teoretické výsledky svého zkoumání Laplace srovnává s naměřenými daty. Laplace vybral 24 měření Saturnu z období 200 let. Jeho teorie však obsahovala čtyři neznámé, které nebylo možné získat s dostatečnou přesností. Proto se rozhodl, že tyto neznámé určí ze samotných pozorování. Čtyři neznámé označovaly určité korekce pro střední zeměpisnou délku Saturnu v roce 1750, jeho průměrný roční pohyb, jeho excentricitu a pozici afélia[30]. Laplace tak získal 24 nekonzistentních rovnic. Tyto rovnice byly lineární ve všech svých čtyřech proměnných. Dostal se tak vlastně téměř ke stejnému problému, jako před ním Euler a Mayer. Mayer v této situaci seskupil všechny rovnice do disjunktních skupin. Laplace postupuje trochu odlišně. Rozhodl se zredukovat 24 rovnic na čtyři následující rovnice:

(i) součet všech rovnic,

(ii) rozdíl součtu rovnic 1–12 a součtu rovnic 13–24,

(iii) lineární kombinace rovnic $-\ +\ +\ -\ +\ 0+\ 1-\ 4+\ 7+\ 8-\ 0+\ 3+\ 4$,

(iv) lineární kombinace rovnic $2-\ -\ +\ +\ -\ 2-\ 3+\ 5+\ 6-\ 9+\ 1+\ 2$.

Nyní již Laplace dostal čtyři rovnice pro čtyři neznámé, které vyřešil. Navíc pomocí reziduí ověřil, jak dobře výsledné rovnice vystihují naměřená pozorování. Laplace nicméně nevysvětluje, proč se rozhodl právě pro tyto lineární kombinace původních rovnic. Oproti Mayerovi je tu ten rozdíl, že Mayer každé pozorování použil pouze jednou. Mayer rozděloval rovnice do skupin podle hodnot koeficientů jediné neznámé. Laplace stejné rovnice používá vícekrát.

Laplace se zabýval i problémem tvaru Země. K tomotu tématu se vrací v roce 1789, tentokrát však bere na zřetel práci Boscoviche. Vychází také z měření oblouků na zemském povrchu. Zjišťuje, že Boscovichova metoda je sice dobrá, ve své originální podobě však příliš složitá. Proto převádí tuto metodu na analytickou formu. To by sice nemuselo být považováno za žádný velký pokrok, pouze o jistý přepis jedné metody, nicméně Laplace o deset let později Boscovichovu metodu ještě poněkud modifikuje. Počítá již pouze se sedmi oblouky. Laplace upravuje původní Boscovichovy podmínky na nové:

(i) suma chyb způsobených při měření celých oblouků musí být nula,

(ii) suma všech chyb braných v absolutních hodnotách musí být minimální.

[30] **Afélium** – česky odsluní. Nejvzdálenější místo od Slunce (od ohniska dráhy), jímž prochází těleso, které se pohybuje kolem Slunce po elipse.

Navíc ve své analýze používal vážení pozorování. Ovšem naprosto odlišné od svých předchůdců. Laplace přiřazoval váhy jednotlivým měřením podle délek jejich oblouků. Delší oblouky tak měly větší váhu.

Laplace poskytl analytický důkaz, že toto je skutečně řešení daného problému tvaru Země. Ze sedmi obloukových měření vypočetl hodnotu eliptičnosti 1/132.

Obrázek 2.5. *Pierre Simon de Laplace a Adrien Marie Legendre*

2.6 Legendre a uveřejnění metody nejmenších čtverců

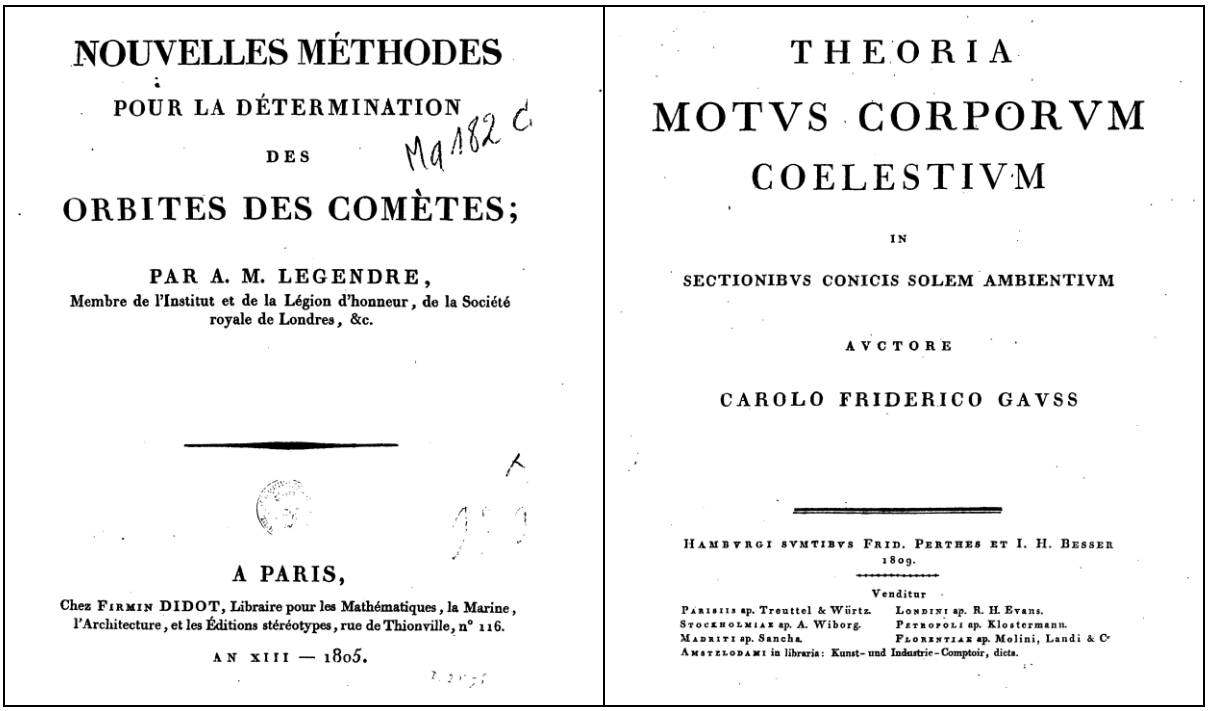
Francouzská revoluce přinesla do své země mnoho změn. Jednou z nich byl i požadavek změnit dávný systém měr a vytvořit nový metrický systém. Základem tohoto nového systému měl být jeden metr, definovaný jako 1/10 000 000 poledníkového kvadrantu[31]. Na francouzské vědě bylo určení délky tohoto oblouku. Do tohoto projektu byl zapojen i Legendre[32]. Francouzi ale nepočítali s tím, že by změřili celý kvadrant. Svůj výpočet délky nové jednotky založili na změření 10° tohoto oblouku. Celá tato část se nalézala na francouzském území – od Montjouy do Dunkirque. Měření probíhalo v roce 1795. Do roku 1799 byl proveden převod velkého množství úhlových měření na délky oblouků. Oficiální

[31] **Poledníkový kvadrant** – vzdálenost z rovníku na pól, v tomto případě se měla na mysli vzdálenost od rovníku na Severní pól.

[32] **Adrien Marie Legendre** (1752–1833) – pocházel z bohaté rodiny, díky tomu získal vynikající vzdělání v matematice a fyzice na Collége Mazarin v Paříži. V letech 1775–1780 přednášel na École Militaire. Poté odešel na Berlínskou akademii. Zabýval se elipsoidy, definoval Legendreovy funkce. Ve svém článku *Recherches sur la figure des planetes* z roku 1784 používá Legendreovy polynomy. Dále se zabývá teorií čísel. V roce 1785 se Legendre stává řádným členem Académie des Sciences. Roku 1791 byl zapojen do výboru Académie des Sciences pro zavedení metrického systému. Rok poté pracuje Legendre na přípravě logaritmických a trigonometrických tabulek. Legendre se zabýval i geometrií. V roce 1794 vydává své dílo *Eléments de géométrie*. V roce 1805 publikuje *Nouvelles méthodes pour la détermination des orbites des cométes*. V této práci poprvé uveřejňuje metodu nejmenších čtverců.

zprávy o tomto výzkumu byly zveřejněny až v roce 1805. Nicméně již kolem roku 1799 byl k dispozici souhrn naměřených dat.

Legendre se k datům vrací při přípravě spisu o určení oběžných drah komet. Tato práce vychází v roce 1805 pod názvem *Nouvelles méthodes pour la détermination des orbites des cométes*. Následovala vydání v roce 1806 a 1820, rozšířená o dodatek. V této práci poprvé publikuje tuto metodu, jak pomocí minimalizování čtverců chyb získat požadované hodnoty pozorovaných veličin. Pravidlo, jak vytvořit normální rovnice je odvozeno a aplikováno na praktických příkladech. O prvních důkazech metody nejmenších čtverců pojednává v [64] Mansfield Merriman.



NOUVELLES MÉTHODES

POUR LA DÉTERMINATION

DES

ORBITES DES COMÈTES;

PAR A. M. LEGENDRE,

Membre de l'Institut et de la Légion d'honneur, de la Société royale de Londres, &c.

A PARIS,

Chez FIRMIN DIDOT, Libraire pour les Mathématiques, la Marine, l'Architecture, et les Éditions stéréotypes, rue de Thionville, n° 116.

AN XIII — 1805.

THEORIA

MOTVS CORPORVM

COELESTIVM

IN

SECTIONIBVS CONICIS SOLEM AMBIENTIVM

AVCTORE

CAROLO FRIDERICO GAVSS

HAMBVRGI SVMTIBVS FRID. PERTHES ET I. H. BESSER

1809.

Venditur

PARISIIS ap. Treuttel & Würtz. LONDINI ap. R. H. Evans.
STOCKHOLMIAE ap. A. Wiborg. PETROPOLI ap. Klostermann.
MADRITI ap. Sancha. FLORENTIAE ap. Molini, Landi & C.
AMSTELODAMI in libraria: Kunst- und Industrie-Comptoir, dicta.

Obrázek 2.6. *Titulní strany vydání knihy Legendrea z roku 1805 a Gausse z roku 1809*

2.7 Gauss a spor o prvenství v objevení metody nejmenších čtverců

Jak již bylo řečeno, metoda nejmenších čtverců byla poprvé publikována v roce 1805 Legendrem. V roce 1809 vychází Gaussova[33] kniha *Theoria motus corporum coelestium*, ve

[33] **Carl Friedrich Gauss** (1777–1855) – se narodil v Brunswicku v Německu. Díky svému mimořádnému talentu a výborným výsledkům na místním gymnáziu získal stipendium, díky kterému mohl studovat na Collegium Carolinum. V roce 1795 odešel na univerzitu v Göttingenu, kde pokračoval ve studiu. Později (1807) se na této univerzitě stal profesorem astronomie a ředitelem hvězdárny v Göttingenu, kde setrval po zbytek svého života. Během svého života se věnoval mj. algebře, teorii čísel, analýze, diferenciálním rovnicím. Ve své doktorské práci v roce 1799 dokázal základní větu algebry a ukázal, že libovolný reálný polynom může být zapsán jako součin lineárních a kvadratických členů s reálnými koeficienty. Jako teoretický matematik pracoval

které Gauss publikuje své metody k výpočtům oběžných drah několika planet. Zde také poprvé uvádí metodu nejmenších čtverců. Gauss tvrdí, že prvenství objevení metody nejmenších čtverců náleží jemu a že zmiňovanou metodu používal již od roku 1795. Měl vůbec Gauss na toto tvrzení nárok? A na jakých důkazech se jeho prohlášení zakládalo?

Theoria motus byla původně napsána v němčině již na podzim roku 1806. Anders Hald uvádí [29], že v červenci 1806 měl Gauss k dispozici kopii Legendreovy knihy *Nouvelles méthodes pour la détermination des orbites des cométes*, než byla poslána Olbersovi[34] k revizi. Až v roce 1807 našel Gauss vydavatele své knihy, který nicméně požadoval překlad do latiny. *Theoria motus* tak byla publikována až v roce 1809.

Gaussovo přisvojení metody nejmenších čtverců roztrpčilo Legendrea, který poslal Gaussovi v květnu 1809 dopis, v němž ho mj. upozorňuje, že není možné nárokovat si objev jen pouhými slovy, že metodu používal již dříve.

Stephen Stigler [83] uvádí čtyři hlavní důkazy, které uvedl Gauss a jeho následovníci na obhajobu Gaussova prvenství:

(i) Gaussova slova z roku 1809, že používal tuto metodu již od roku 1795.

(ii) Tajemný zápis v Gaussově matematickém deníku s datem 17. června 1798: „Calculus probabilitatis contra La Place defensus“.

(iii) Gaussovo tvrzení, že o metodě řekl dalším astronomům (Olbers, Lindenau[35], Zach[36]) před rokem 1805.

(iv) Gaussův dopis, který byl v roce 1799 publikován v Allgemeine Geographische Ephemeriden, v němž zmiňuje „meine Methode“. Píše v něm o tiskařské chybě v popisu oblouku mezi Panthéonem a Evaux. Místo 76 545,74 zde bylo 76 154,74. Objevil chybu po aplikaci svojí metody na čtyři naměřené oblouky. S chybou vycházela eliptičnost 1/150 a po odstranění této chyby 1/187. Nicméně poznamenal, že chyba není v tomto případě podstatná, protože koncové body leží příliš blízko u sebe.

Gauss sám. Nicméně v aplikované matematice pracoval společně s astronomy, později s geodety a fyziky. Mezi jeho spolupracovníky z oblasti astronomie patřili Friedrich Wilhelm Bessel (1784–1846) a Heinrich Wilhelm Matthäus Olbers (1758–1840). Spolupracoval také s profesorem fyziky v Göttingenu Wilhelmem Eduardem Weberem (1804–1891). Více o Gaussovi např. v [22].

[34] **Heinrich Wilhelm Matthäus Olbers** (1758–1840) – německý astronom, lékař a fyzik. Vystudoval lékařství v Göttingenu. Po absolvování se věnoval medicíně v Brémách. Zabýval se také astronomií. Byla po něm mj. pojmenována kometa, kterou objevil v roce 1815.

[35] **Bernhard August von Lindenau** (1779–1854) – německý právník, astronom a ministr.

[36] **Franz Xaver von Zach** (1754–1832) – astronom a matematik. Studoval nejprve ve Vídni, poté v Anglii. Od roku 1788 vydával časopis *Allgemeine Geographische Ephemeriden*, ve kterém sám publikoval svoje vědecké práce z astronomie. O pohybech Slunce napsal práci *Novae et correctae tabulae motuum solis*, která vyšla v roce 1792. Od roku 1800 byl editorem měsíčníku se širším zaměřením *Monatliche Correspondenz zur Beförderung der Erd- und Himmelskunde*. Působil v Anglii, Švýcarsku, Francii.

Ani jeden z těchto důkazů však není dostatečně přesvědčivý. Podívejme se na každý z nich podrobněji. Gaussovo prohlášení, že používal metodu již od roku 1795, je podporováno Gaussovými matematickými schopnostmi. Nepotřeboval dělat falešná prohlášení. Na druhou stranu Gaussovo tvrzení se zakládá na dodatečné vzpomínce. Proč neuveřejnil tuto metodu jako první, když ji, jak prohlašuje, objevil? Pokud připustíme, že opravdu metodu znal, pak je možné, že jí nepřikládal velkou důležitost a neuvědomoval si její praktický význam. A z tohoto důvodu nepovažoval za nezbytné její uveřejnění. Nebo se mu metoda zdála příliš zřejmá a jednoduchá.

Pokud se podíváme na druhý bod – záznam v deníku, pak ani toto není jednoznačný důkaz. Záznam nás pouze ujišťuje o tom, že se Gauss zabýval otázkami souvisejícími s pravděpodobností v červnu 1798.

Také třetí bod postrádá přesvědčivost. Astronom Olbers sice podpořil Gaussovo prohlášení v poznámce pod čarou v roce 1816 s tím, že si vzpomíná, že mu Gauss zmínil základní princip metody již v roce 1803. Nicméně tak učinil až po sedmi letech opakovaného pobízení Gaussem. A co další dva astronomové, o kterých se Gauss zmiňoval? Zach v letech 1800 až 1813 vydával astronomický časopis *Monatliche Correspondenz*, který se skládal především z recenzí a dopisů. Lindenau mu asistoval ve vydávání tohoto periodika od roku 1807. Lindenau v recenzi článku o geodézii ze srpna 1806 detailně rozebírá metodu nejmenších čtverců. Není zde ale žádná zmínka o Gaussovi. Metoda je připisována Legendreovi. Je zde i citována jeho práce *Nouvelles méthodes pour la détermination des orbites des cométes*. V listopadovém vydání *Monatliche Correspondenz* z roku 1807 se objevují dvě poznámky o metodě nejmenších čtverců. Ani jedna z nich není podepsaná, pravděpodobně byly napsány vydavateli. V obou se mluví o Legendreově metodě a v obou je metoda označována francouzským názvem Méthode des moindres quarrés. O Gaussovi není nikde ani zmínka. Přitom Gauss tvrdil, že o svojí metodě řekl oběma astronomům. Nicméně ani toto není jednoznačný důkaz toho, že Gauss jim o nejmenších čtvercích nic neřekl před rokem 1805. Možná se jen cítili omezeni vědeckou normou, že první publikování určuje prvenství. Mohlo se stát i to, že Gauss jim sice metodu popsal, nicméně oni mu správně neporozuměli. Výtah z Gaussových dopisů Olbersovi a von Zachovi je možné nalézt v [71].

Dostáváme se k poslednímu nepřímému důkazu – k výsledkům Gaussova počítání s použitím „meine Methode“. To nás přivádí k zajímavé otázce: mohou být stejné výsledky odvozeny z originálních dat použitím metody nejmenších čtverců? Data, která měl Gauss k dispozici jsou uvedena v tabulce 2.3. Jak poznamenal Gauss, místo 76 545,74 má být hodnota 76 145,74. Údaje pocházejí z geodetického měření, které bylo základem pro určení

nového metrického systému. V roce 1793 bylo ve Francii rozhodnuto, že základní jednotkou nového metrického systému bude jeden metr – délka rovná jedné desetimiliontině poledníkového kvadrantu, což je vzdálenost ze severního pólu na rovník podél odpovídající zeměpisné šířky procházející Paříží. Pro každou část délka oblouku je udávána v modulech, a to je délka odpovídající přibližně 12,78 stopy. Na základě těchto dat Gauss spočítal „svojí metodou“ eliptičnost Země 1/150. Poté však nalezl tiskovou chybu a přepočítal eliptičnost na 1/187. Mohlo by se zdát, že na základě těchto znalostí bude jednoduché jednoznačně zjistit, zda Gauss opravdu použil metodu nejmenších čtverců. Stačí vzít odpovídající hodnoty, použít metodu nejmenších čtverců a zjistit, zda se námi vypočtené hodnoty shodují s těmi Gaussovými. Bohužel problém je složitější. Problém je v tom, že vztahy mezi délkou oblouku, zeměpisnou šířkou, poledníkovým kvadrantem a eliptičností nejsou lineární. Takže existuje více způsobů, jak úlohu převést na problém lineárních nejmenších čtverců. Navíc se mohou objevit odchylky způsobené zaokrouhlováním. Obvyklá lineární formulace problému, kterou použil mj. Boscovich (1755) nebo Legendre (1805) je za předpokladu, že je Země je elipsoid následující:

$$a = z + y \sin^2 L ,$$

kde $a = S/d$ (délka oblouku v modulech dělené stupněm zeměpisné šířky), z je délka stupně na rovníku a y je rozdíl (přebytek) stupně na pólu v porovnání s jedním stupněm na rovníku.

	Moduly *S*	Stupně *D*	Střed oblouku *L*
Dunkirk – Pantheon	62 472,59	2,189 10	49° 56′ 30″
Pantheon – Evaux	76 545,74	2,668 68	47° 30′ 46″
Evaux – Carcassonne	84 424,55	2,963 36	44° 41′ 48″
Carcassonne – Barcelona	52 749,48	1,852 66	42° 17′ 20″
Součet	275 792,36	9,673 80	-

Tabulka 2.3. *Francouzské měření oblouku. Tabulka udává délku čtyř na sebe navazujících částí poledníkového oblouku procházejícího Paříží, jak v modulech S (1 modul je přibližně 3,8953 metru), tak ve stupních d zeměpisné šířky (určených pomocí astronomických pozorování). L značí zeměpisnou šířku středu každé části oblouku.*

Pokud budeme *S* brát jako závisle proměnnou a vyřešíme rovnici pro *z* a *y* za použití metody nejmenších čtverců, Stigler [83] se dostává k výsledkům:

z = 28 227,162 05

y = 541,263 935 3

1/eliptičnost = 157,95

poledníkový kvadrant = 2 564 801,46

Pokud budeme pracovat s tiskovou chybou, pak dostáváme:

z = 28 074,826 97

y = 906,790 131 4

1/eliptičnost = 94,38

poledníkový kvadrant = 2 567 539,98

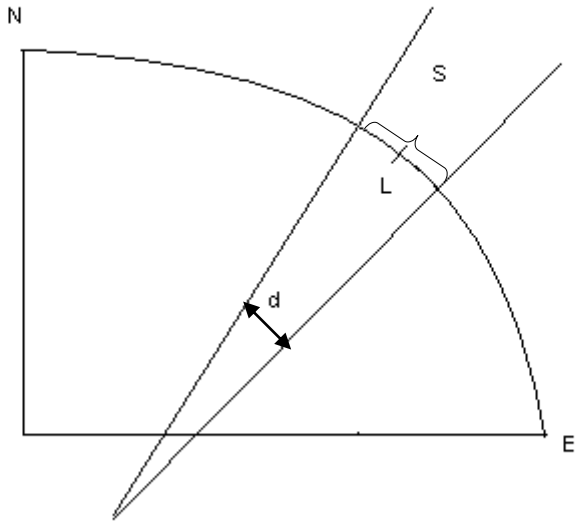
Gaussovi však vycházejí značně odlišné hodnoty:

z = 28 271,456 5

y = 457,220 3

1/eliptičnost = 187

poledníkový kvadrant = 2 565 006



Obrázek 2.7. *Vztah mezi délkou oblouku, zeměpisnou šířkou, poledníkovým kvadrantem a eliptičností*

Proto se Stigler pokusil použít vážené nejmenší čtverce, ale ani v tomto případě nebyl úspěšný. Ani položení d za závisle proměnnou nepřineslo kýžené výsledky. Zbývají tedy dvě možnosti. Buď Gauss aplikoval metodu nejmenších čtverců na vztah $a = z + v\sin^2 L$ a dopustil se při výpočtu chyby, nebo nejmenší čtverce nepoužil. První varianta se zdá být vysoce nepravděpodobná. Gauss byl vynikajícím počtářem a v tomto případě se jednalo pouze o krátkou posloupnost výpočtů.

Stigler se dále zabývá otázkou, jak tedy Gauss přišel ke svým výsledkům. Gauss byl v první řadě matematik, ne statistik. Stigler tedy vznesl zajímavou otázku. Co když se Gauss nespokojil s první aproximací $a = \ + \ \sin^2 L$? Boscovich a Laplace věděli, že chyby v určení *S*, *d* a *L* byly poměrně velké ve srovnání s chybami způsobenými aproximací. Neměli tak důvod pokračovat dále. Neočekávalo se žádné podstatné zlepšení v přesnosti. Sám předpoklad, že Země je elipsoidem, je sám o sobě aproximací. Gauss se ale s takovou odpovědí nemusel spokojit. Zdá se tedy možné, že Gauss mohl použít nejmenší čtverce, ale na aproximaci, řekněme, druhého stupně. Problém je v tom, že čím větší třídu vztahů použijeme, tím složitější je situace. Určitou expanzi druhého stupně použil Bowditch (1832) a Bessel (1837):

$$S = \ d + \ \sin d \cos 2L + \ \sin 2d \cos 4L \, .$$

Na *x*, *y* a *z* může být pohlíženo jako na nelineární funkce eliptičnosti a délky stupně na rovníku. Stigler uvádí dva důvody pro podporu tvrzení, že Gauss by mohl použít metodu nejmenších čtverců na rozvoj druhého řádu. Prvním z nich je výrazný rozpor mezi eliptičností, kterou našel před a po odstranění typografické chyby. Tohoto velkého rozdílu nedosáhneme při aplikování nejmenších čtverců nebo Boscovichovy metody na rozvoj prvního stupně. Jako druhý důvod uvádí Stigler fakt, že zkoušel určit z druhého rozvoje poledníkový oblouk s využitím nejmenších čtverců a jeho výsledky byly velmi povzbudivé. Např. pokus řešit vztah $S = \ d + \ \sin d \cos 2L + \ \sin 2d \cos 4L$ pomocí nevážených nejmeších čtverců dal poledníkový kvadrant 2 565 012 a jiný přístup používající Besselův přístup přes nelineární nejmenší čtverce dal přesně Gaussovu hodnotu 2 565 006. Nicméně žádný z těchto Stiglerových postupů nebyl úspěšný v simulaci určení Gaussových hodnot pro eliptičnost.

Poté, co Stigler v roce 1981 publikoval svoji domněnku [83, str. 320–321, rozšířená verze článku Stigler S.: *Gauss and the Invention of Lest Squares*. Annals of Statistics 9(1981), 465–474], několik schopných analytiků se pokusilo napodobit Gaussovy výsledky s použitím rozvoje druhého stupně nebo jiných obměn metody nejmenších čtverců. Ovšem nikdo z nich nebyl úspěšný. Gilstein a Leamer (1983) ukázali, že Gauss nemohl své výsledky najít pomocí vážených nejmenších čtverců z formulace prvního stupně použitím všech možných vah a Celmins (1998) došel ještě dále [11] a demonstroval, že Gaussovy výsledky nemohly být získány žádnou ze široké třídy přístupů vycházejících z nejmenších čtverců, včetně požadovaného vztahu a mnoha přibližných vyšších řádů expanze. Stigler tak dochází k závěru, že nejpravděpodobnější je varianta, že Gauss použil nějaký odlišný přístup

k problému, který dosud nebyl znovu objeven. Následné statistické přepočty ukazují, že nejpozději v díle z roku 1799 měl Gauss k dispozici nějakou metodu sloučení nekonzistentních rovnic, získaných pozorováním. Pokud to tedy nebyla metoda nejmenších čtverců, co to bylo?

Pokud bychom připustili, že Gauss objevil metodu nejmenších čtverců nezávisle na Legendreovi a používal ji již před rokem 1799, pak stále zůstává otázkou, proč on sám tuto metodu nepublikoval? Jakou jí vlastně přisuzoval důležitost? Je tak možné, že se sice o svojí metodě zmínil jiným astronomům před rokem 1805, ale nejasný Gaussův výklad nebo nedostatek možností aplikace této metody mohly být příčinou nepochopení nejmenších čtverců. O to větší obdiv si ale zaslouží Legendre, který uveřejněním nejmenších čtverců v roce 1805 dosáhl okamžitého a všeobecného úspěchu. Další příčinou toho, proč Gauss nově nalezenou metodu nepublikoval, může být i fakt, že Gaussovi metoda nejmenších čtverců připadala příliš jasná a zřejmá. Mohl proto nabýt dojmu, že není potřeba něco tak jednoduchého publikovat. Gauss byl tedy možná tím, kdo objevil metodu nejmenších čtverců jako první, nicméně Legendre byl první, kdo metodu uveřejnil a zpřístupnil ji široké veřejnosti.

Kapitola 3:

Citlivost odhadu metodou nejmenších čtverců k odchylkám od normálního rozdělení a první alternativní odhady

3.1 Dogma normality

Normální rozdělení bylo dlouho považováno za rozdělení, kterým se řídí většina náhodných veličin. Často uváděný citát z roku 1912, který Poincaré přisuzuje Lippmannovi by se dal přeložit následovně: *Všichni věří v normální rozdělení chyb: experimentátoři, protože je pokládají za matematický teorém, a matematikové, protože je pokládají za experimentální fakt.*[37]

Dogma normality bylo možné popřít až v době výkonných počítačů. Ale již v roce 1965 Abram M. Kagan[38], Yuri V. Linnik[39] a Calyampudi R. Rao[40] dokázali, že odhad metodou nejmenších čtverců je optimální pro normální rozdělení chyb a v jiných případech může úplně selhat. Když se podíváme na Gaussovo zavedení normálního rozdělení, zjistíme, že Gauss vlastně zavádí normální rozdělení tak, aby vyhovovalo aritmetickému průměru. K téměř nedotknutelnosti dogmatu normality nejspíš přispěla i Gauss-Markovova věta (nejlepší lineární nestranný odhad očekávané hodnoty je aritmetický průměr) a Centrální limitní teorém (součet mnoha malých na sobě nezávislých chyb je přibližně normální).

V předchozí kapitole jsme mluvili o metodě nejmenších čtverců. Lidé, kteří tuto metodu používají, někdy mlčky předpokládají bez dalšího ověřování, že velikosti chyb se řídí normálním rozdělením. Už v devatenáctém století si Legendre nejspíš uvědomoval, že metoda nejmenších čtverců není vždy optimální. Ve své práci o metodě nejmenších čtverců z roku 1805 doporučuje nejprve zamítnout měření, která jsou příliš velká na to, aby mohla být považována za přípustná.

Nejspíš poprvé si všimli citlivosti klasických statistických charakteristik, jako je průměr a rozptyl, k odlehlým hodnotám astronomové a fyzikové při určování různých

---

[37] *… car les expérimentateurs s`imaginent que c`est un théoréme de mathématiques, et les mathématiciens que c`est un théoréme de mathématiques, et les mathématiciens que c`est un fait expérimental.* [72, str. 149]

[38] **Abram Meerovich Kagan** (*1936) – působil na státní univerzitě v Leningradu (nyní Petrohrad) a na univerzitě v Taškentu v Uzbekistánu. Od roku 1988 působí na University of Maryland. Mezi oblasti jeho vědeckého zájmu patří parametrické odhady, zobecněné lineární modely, Fisherova informace.

[39] **Yuri Vladimirovich Linnik** (1915–1972) – ruský matematik, zabýval se teorií čísel, teorií pravděpodobnosti a matematickou statistikou. Působil na univerzitě v Leningradu (nyní Petrohrad).

[40] **Calyampudi Radhakrishna Rao** (*1920) – indický matematik. Mezi jeho nejznámější objevy patří Cramér-Raova mez a Rao-Blackwellova věta. Zabýval se teorií odhadu, lineárními modely, mnohorozměrnou analýzou, biometrikou, funkcionálními rovnicemi.

fyzikálních, geofyzikálních a astronomických konstant. Jedním z nich byl i James Short[41], který v roce 1763 odhadoval paralaxu Slunce pozorováním oběžné dráhy Venuše [76]. Nespokojil se s pouhým aritmetickým průměrem, ale zprůměrňoval tři průměry: prostý průměr, průměr všech pozorování s rezidui menšími jak jedna sekunda a průměr pozorování s rezidui menšími jak půl sekundy. V této souvislosti bych si dovolila udělat malou odbočku a zmínit se o tom, proč vůbec Short pozoroval oběžnou dráhu Venuše.

Již od počátku osmnáctého století uměli astronomové docela dobře určit relativní vzdálenosti ve sluneční soustavě, tedy vzájemné vzdálenosti mezi oběžnými dráhami planet, mezi planetami a Sluncem. Nicméně neznali absolutní vzdálenosti. Pokud by se jim podařilo určit jednu takovou vzdálenost, pak by od této vzdálenosti mohli odvodit všechny ostatní. Středem jejich zájmu byla vzdálenost Země od Slunce. V osmnáctém století se tak vědci rozhodli určit paralaxu Slunce. Nejspíš první, kdo navrhl, aby byla paralaxa[42] Slunce určena pozorováním oběžné dráhy Venuše, byl astronom Edmond Halley. Venuše měla být sledována při svém zdánlivém přechodu přes líc Slunce. Při průchodu Venuše přes sluneční disk při pozorování z různých míst na Zemi je možné zjistit, jak se Venuše promítá na různá místa slunečního kotouče. Pomocí změření těchto rozdílů lze dopočítat vzdálenost Venuše od Slunce a následně i vzdálenost Země od Slunce. Nedostatkem zmíněného postupu je fakt, že přechod Venuše přes sluneční disk je poměrně vzácný jev. První zaznamenaný průchod Venuše se odehrál v roce 1639 a byl pozorován pouze v Anglii. Další průchody byly očekávány až v roce 1761 a 1769. Aby bylo možné provést smysluplné měření, bylo potřeba provést pozorování na dostatečně vzdálených místech na Zemi. Do roku 1761 se podařilo vytvořit pozorovatelny na Mysu dobré naděje, v Římě, Kalkatě, Stockholmu a ve většině evropských hvězdáren. Data, kterými se zabýval James Short, pochází právě z roku 1761.

3.2 Zamítnutí odlehlých pozorování

Jedním ze statistických problémů, který souvisí s robustními odhady, je zamítání odlehlých hodnot. O tomto problému byla zmínka již v předchozí kapitole o metodě nejmenších čtverců. Jak ale zjistit, která pozorování jsou opravdu odlehlá a která máme tudíž vyřadit z dalších výpočtů? První návrh pro stanovení odlehlých hodnot publikoval roku 1852

[41] **James Short** (1710–1768) – astronom a známý výrobce dalekohledů.

[42] **Paralaxa** je v astronomii úhel, o který se změní poloha nebeského tělesa na obloze, pokud je pozorováno z krajních bodů vhodně zvolené základny. Výpočet paralaxy se používá pro měření vzdáleností ve vesmíru. Zvláštní postavení má tzv. sluneční paralaxa. Sluneční paralaxa je úhel, pod kterým by byl pozorován rovníkový poloměr Země ze středu Slunce (8,794148″).

[70] Benjamin Peirce[43]. Nicméně Peirce, stejně jako mnoho dalších, se příliš nezajímal o vlastnosti odhadu, který je udělán po zamítnutí odlehlých hodnot. Spíše implicitně předpokládal, že poté, co zamítl odlehlá pozorování, je možné vytvořit odhad bez ohledu na to, jaká informace mohla být ztracena. Tento nedostatek samozřejmě nezůstal dlouho bez povšimnutí. Britský astronom George Biddell Airy[44] v roce 1856 ve svém listu kritizuje použití Peirceova kritéria. Tím se rozpoutává mezinárodní debata, která nebyla uspokojivě rozřešena. Nicméně na americké půdě se Peirce aspoň částečně stává vítězem. V letech 1852–1874 působil v U. S. Coast Survey. Tato společnost měla za úkol zaměřit a zmapovat novou zemi. A právě v letech, kdy zde Peirce působil, byl jeho test na zamítnutí odlehlých hodnot běžně používán členy této organizace.

Obrázek 3.1. *Benjamin Peirce a Simon Newcomb*

3.3 Simon Newcomb a směsi normálních rozložení

Simon Newcomb (1835–1909) je znám jako nejvýznamnější americký astronom devatenáctého století. Určil mnoho astronomických konstant, které jsou dodnes uznávány. Byl také nadaným matematikem a spoluzakladatelem (a mnoho let editorem) *American Journal of Mathematics*. Díky své knize *Principles of Political Economy* se stal také předním americkým

[43] **Benjamin Peirce** (1809–1880) – profesor astronomie a matematiky na Harvardu od r. 1842 až do své smrti. Promoval na stejné univerzitě. Jeho nejvýznamnějším učitelem byl překladatel Laplace Nathaniel Bowditch (1773–1838). Z dnešního hlediska je Peirce ceněn z několika důvodů. Tím prvním je jeho učitelská a vědecká práce. V roce 1870 vydává spis *Linear Associative Algebra*. Dalším důvodem je jeho test na zamítnutí odlehlých hodnot, který vzbudil rozsáhlou debatu o vhodnosti takového počínání. Jeho synem byl Charles Sanders Peirce (1839–1914). Dnes je C. S. Peirce znám především jako filozof a logik. Pracoval, stejně jako jeho otec, v U. S. Coast Survey.

[44] **George Biddell Airy** (1801–1892) – anglický matematik a astronom. Zabýval se planetárními oběžnými dráhami, určením střední hustoty Země. Studoval v Cambridge. V letech 1835–1881 působil jako britský královský astronom.

teoretickým ekonomem. Newcomb byl, jak se zdá, první, kdo představil směs normálních hustot jako model pro rozdělení s těžkými konci a použil tento model k získání odhadu polohy, který byl robustnější jak výběrový průměr.

Simon Newcomb se narodil v Novém Skotsku, kde byl jeho otec učitelem. V šestnácti letech šel do učení k lékaři, které mělo trvat pět let. Nicméně z jeho kariéry lékaře po dvou letech sešlo za poměrně dramatických okolností. V osmnácti se živil jako učitel. V roce 1858 ukončil Harvard`s Lawrence Scientific School.

Newcomb běžně používal vážený průměr ve svých odhadech astronomických konstant. Přitom váhy určoval subjektivně na základě svého vlastního posouzení relativní správnosti a přesnosti v průběhu pozorování. Newcomb při svých výpočtech také zamítal odlehlé hodnoty. Ale pouze v případě, že byly opravdu silně vychýlené.

Při pozorování oběžné dráhy Merkuru (hodnoty byly naměřené 6. 5. 1878) zjistil, že množina 684 reziduí, založených na těchto pozorování, má mnohem těžší konce, než příslušné normální rozdělení. Množina pozorování pocházela z měření s různými stupni přesnosti. Newcomb zjišťuje, že v tomto případě nebude aritmetický průměr vhodnou charakteristikou. Přišel s myšlenkou popsat naměřená data pomocí směsi normálních rozdělení s různými parametry.

O několik let později (v roce 1886) Newcomb publikuje v *American Journal of Mathematics* článek *A Generalized Theory of the Combination of Observations so as to Obtain the Best Result* [67]. V tomto článku nejprve kritizuje přílišné užívání kritéria pro zamítání odlehlých hodnot a poté představuje svůj model směsi rozdělení. Dále navrhuje odhad založený na tom, že menší váhy jsou dávány příliš odlišným pozorováním.

Ve svých pozdějších pracích dokonce Newcomb vytváří jednoduchou verzi Tukeyho citlivostní křivky[45].

3.4 Lineární funkce pořádkových statistik

Lineární funkcí pořádkových statistik označujeme váženou lineární kombinaci pozorování, kde jsou váhy přiřazeny pouze na základě pořadí daného pozorování. Tyto lineární funkce pořádkových statistik můžeme použít k odhadování středních hodnot. Do této třídy patří i velmi dobře známý medián a variační rozpětí. Jak medián, tak variační rozpětí, mají poměrně dlouhou historii.

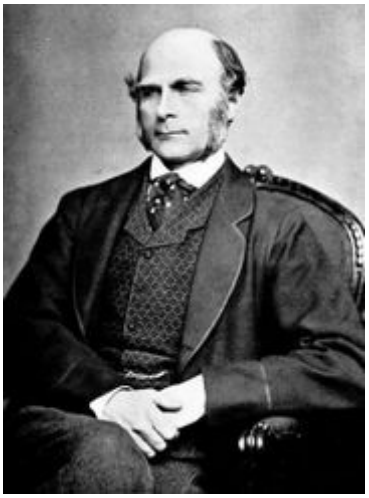
[45] John Wilder Tukey začal používat v roce 1970 citlivostní křivku (sensitivity curve) ke zkoumání vlastností odhadů pro konečné vzorky. Více o Tukeyho citlivostní funkci lze nalézt v knize [1].

Mediánem se zabývá ve své práci *Théorie Analytique des Probabilités* [54] z roku 1812 Laplace. Laplace zde uvažuje problém, který bychom dnes nazvali lineární regresí:

$$a_i = \nu_i y + \varsigma_i ,$$

kde a_i, p_i jsou známé, odhadujeme y, x_i jsou chyby, které pocházejí z libovolného symetrického rozdělení. Laplace hledal odhad, který by minimalizoval součet absolutních hodnot reziduí. Dospěl k tomu, že takovým odhadem je medián a_i v případě, že $p_i \equiv$. Laplace odvodil hustotu takového odhadu a ukázal, že s rostoucí velikostí výběru se tato hustota blíží hustotě normálního rozdělení.

Mnoho prací o rozlišných použitích mediánu publikoval v roce 1875 Francis Galton[46]. Medián je také dnes často využívaným odhadem i z toho důvodu, že je robustní k odchylkám od normálního rozdělení. Galtonovou motivací však byla, spíše než nedůvěra k normálnímu rozdělení, jednoduchost výpočtu mediánu a snadnost jeho interpretace. Galton navrhuje použití mediánu v případech, kdy lze očekávat těžší konce než u normálního rozdělení. Mnoho podobných znaků lze nalézt v nezávislé práci Gustava Theodora Fechnera[47].



Obrázek 3.2. *Francis Galton*

V roce 1889 Galton navrhuje komplikovanější lineární odhad průměru a směrodatné odchylky ve tvaru:

[46] **Francis Galton** (1822–1911) – anglický psycholog, statistik a antropolog. Je zakladatelem eugeniky (sociálně-filozofický směr zaměřený na studium metod, které povedou k dosažení co nejlepšího genetického fondu člověka). Francis Galton byl vnukem Erasma Darwina. Po Galtonovi je pojmenován přístroj modelující normální rozdělení (Galtonův přístroj, quincunx) a také vysokofrekvenční píšťala užívaná pro výcvik psů, tzv. Galtonova píšťala.

[47] **Gustav Theodor Fechner** (1801–1887) – německý fyziolog, fyzik, psycholog, zakladatel psychofyziky, jeden ze zakladatelů experimentální psychologie a estetiky, profesor univerzity v Lipsku.

$$\mu = \frac{\xi_{} X^{(nq)} - \; _{} X^{(np)}}{\xi_{} - \; _{}},$$

$$\hat{\sigma} = \frac{Y^{(np)} - \; \zeta^{(nq)}}{\xi_{} - \; _{}},$$

kde p, q jsou libovolné, ale pevné $0 < p < q < \; _{}$, $\xi_p, \xi_{} \;$ jsou p a q percentily standardizovaného normálního rozdělení, $X^{(np)}$ a $X^{(nq)}$ jsou výběrové $100p$ a $100q$ percentily. Důkaz sdružené asymptotické normality Galtonových odhadů v případě, že populace je normální, dal v roce 1899 William Fleetwood Sheppard[48] [82].

3.5 Percy John Daniell

V roce 1920 Percy John Daniell publikuje v *American Journal of Mathematics* článek *Observations Weighted According to Order*. Nicméně tento článek byl nejspíš naprosto přehlédnut. Trvalo dalších třicet let, než byly Daniellovy výsledky znovuobjeveny. Pokusme se podívat na důvody tohoto opomenutí.

Percy John Daniell (1889–1946) byl anglický matematik. Vystudoval na univerzitě v Cambridge. Jeho pobyt zde se sice překrýval s R. A. Fisherem, ale protože byli na různých fakultách, nemuseli se vůbec potkat. Po ukončení studií krátce pobýval v Göttingenu a Liverpoolu. V roce 1912 odchází do Rice Institute v Houstonu v Texasu. V roce 1920 se zde stává řádným profesorem. Právě v Houstonu vytváří svoji nejvýznamnější práci z integrálního počtu, kde objevuje to, co dnes známe pod názvem Daniellův integrál. V roce 1924 se vrací do Anglie na univerzitu v Sheffieldu, kde zůstává až do své smrti. Článek z roku 1920, který se zabývá lineárními pořadovými statistikami, napsal Daniell během svého působení v Rice. A to byl nejspíš jeden z důvodů, proč tato jeho práce upadla v zapomnění. Jak Rice, tak Sheffield byly izolované od aktivního statistického výzkumu. Navíc Daniell byl znám jako matematik a práce z roku 1920 se mohla zdát, že pouze vzdáleně souvisí se statistikou. U Daniella se tato nešťastná situace opakovala hned o rok později. I jeho článek z 1921 *Integral Products and Probability*, uveřejněný opět v *American Journal of Mathematics*, upadl v zapomnění. Tentokrát zde Daniell jako první prezentuje matematické zpracování spojitých Markovových procesů, včetně Chapman-Kolmogorovy rovnosti.

[48] **William Fleetwood Sheppard** (1863–1936) – právník, matematik a statistik. Narodil se v Austrálii. Své vzdělání už ale dokončil v Anglii v Cambridge. Po dokončení studií pracoval jako právník. V roce 1896 se ale rozhodl věnovat vzdělávání. V roce 1926 se stal členem Edinburgh Mathematical Society. Byl také členem Matematical Association, které dokonce v letech 1928–1929 předsedal.

Vraťme se ale k Daniellově práci z roku 1920. Jeho práce je inspirována Poincarého knihou *Calcul des Probabilités* z roku 1912, protože hned v úvodu se na tuto práci odkazuje v souvislosti se zamítáním odlehlých pozorování v případě počítání aritmetického průměru. Hned ale pokračuje (přeloženo podle [13]):

Kromě takových ořezaných[49] *průměrů, můžeme vymyslet jiné, ve kterých by byly váhy přiřazeny měřením podle jejich pořadí. Ve skutečnosti běžný průměr, medián, ořezaný průměr, odchylka (od mediánu) a kvartilová odchylka mohou být také posuzovány, že byly získány postupem, ve kterém jsou měření násobena činiteli, které jsou funkcemi pořadí. Hlavní cíl této práce je získat vzorec pro střední kvadratickou odchylku jakéhokoliv takového výrazu. Tento vzorec by pak mohl být použit pro výpočet relativních přesností všech takovýchto výrazů.*[50]

Ve druhé části své práce Daniell zavádí statistiku

$$\bar{t} = \sum_{r=1}^{n} f_r t_r \ ,$$

kde $t_1, t_2, \ldots, t_n$ jsou měření uspořádaná podle velikosti tak, že $t_1 \leq t_2 \leq \ldots \leq t_n$. Tato měření pocházejí z rozdělení s hustotou $p(x)$. Váhy f_r jsou dány výrazem: $f_r = \frac{1}{n} f(x_r)$, kde veličina x_r je zavedena jako $x_r = \frac{r}{n+}$, takže vztah se dá přepsat:

$$f_r = \frac{1}{n} f\left(\frac{r}{n+}\right)$$

Daniell zavádí pravděpodobnostní integrální transformaci. Postupným odvozováním se dostává k výrazu pro asymptotický rozptyl $\bar{t}$:

[49] **Ořezaný průměr** – Daniell zde použitá anglický výraz „discard-average“, což by se dalo do češtiny přeložit jako „odložený“, „odřezaný“ nebo „ořezaný“ průměr. Dnes mluvíme o „trimmed mean“, neboli useknutém průměru. Ten ale vznikne symetrickým ořezáním souboru měření. Daniell měl nejspíš na mysli obecnější ořezání, protože hovoří o tom, že je možné zamítnout jednu nebo i více extrémních hodnot. Proto jsem v jeho citaci použila trochu nestandardní pojem „ořezaný“ průměr, aby bylo zřejmé, že se nemusí obecně jednat o tentýž průměr.

[50] *Besides such a discard-average we might invent others in which weights might be assigned to the measures according to their order. In fact the ordinary average or mean, the median, the discard-average, the numerical deviation (from the median, which makes it minimum), and the quartile deviation can all be regarded as calculated by a process in which the measures are multiplied by factors which are functions of order. It is the general purpose of this paper to obtain a formula for the mean square deviation of any such expression. This formula may then be used to measure the relative accuracies of all such expressions.* [13, str. 222]

$$S^2 = \int_{-\infty}^{\infty} \varphi(p)\,dt,$$

kde $\varphi(t) = \int_{c}^{t} f(x)\,dx$ a $x(p) = \int_{-\infty}^{t} p(u)\,du$.

Ve třetí části Daniell definuje přesnost $\bar{t}$ jako poměr σ / S^2, kde σ je teoretický rozptyl a S^2 asymptotický rozptyl $\bar{t}$, odvozený v předchozí části. Ve čtvrté části svého spisu se Daniell zabývá nejlepší váhovou funkcí, která bude minimalizovat S^2. V případě, že se bude jednat o normální rozdělení, pak doporučuje použít pro výpočet průměru všechny váhy stejné. Dále dodává, že u normálního rozdělení přesnost nejvhodnějšího odhadu σ bude stejná jako výběrové směrodatné odchylky. V případě, že se bude jednat o supernormální rozdělení (je zde více extrémních hodnot než u normálního rozdělení), navrhuje při výpočtu $\bar{t}$ snížit váhy odlehlým hodnotám tak, že pro příliš velké hodnoty *t* by mohly být tyto váhy dokonce záporné. Další možností pak může být to, že vezme všechny váhy stejné pro hodnoty v určitých mezích a zamítne všechna ostatní měření mimo tyto meze. Jak říká, toto lze nazývat ořezaným průměrem a v praxi počítat tak, že ořízneme vnější část měření. Pro výpočet odchylky pak navrhuje zamítnout ne vnější část měření, ale vnitřní.

Následující dvě kapitoly jsou zaměřeny na příklady použití. Daniell věnuje zvláštní pozornost ořezanému průměru. Symbolem *k* označujete u tohoto typu průměru centrální část, která zbyde po ořezání. Uvádí zde výraz pro asymptotický rozptyl ořezaného průměru, pro který $k = 1/2$ (z každé strany odřežeme čtvrtinu hodnot, nazývaný jako „quartile-discard average“). Dále se zabývá srovnáním přesnosti takto ořezaného průměru ($k = 1/2$) a prostého průměru a dává podmínky, za kterých je ořezaný průměr lepší. Další typ průměru, který v páté kapitole Daniell uvažuje, je medián-kvartilový průměr (median-quartile average), který se počítá jako průměr mediánu a dolního a horního kvartilu, neboli:

$$\bar{t} = \frac{1}{3}(Q_1 + M + Q_3).$$

Pokud je počet měření velký, Daniell ho považuje za přesnější než ořezaný průměr s $k = 1/2$. Ovšem v případě, že je počet měření malý, pak je těžké určit kvartily přesně, a tudíž (z [13]):

Po uvážení všech aspektů můžeme říci, že medián-kvartilový průměr a ořezaný průměr (s k = 1/2) jsou téměř stejně přesné.[51]

[51] *...taking everything into consideration, we may say the median-quartile average and the quartile-discard average are about equally accurate.*[13, str. 234]

3.6 Hendrik van de Hulst

Tím, co bychom dnes nazvali robustní statistikou, se zabýval i nizozemský astronom van de Hulst. Svoje poznámky ale nikdy nepublikoval. Část jeho otištěných poznámek je tak možné najít u van Zweta[52] [88]. Van de Hulst[53] byl tehdy studentem astronomie v Utrechtu a později profesorem teoretické astronomie v Leidenu. Van de Hulst svoje poznámky dával ke kontrole van Dantzigovi[54]. V době, kdy se van Dantzig začal zabývat matematickou statistikou, byla matematická statistika mezi nizozemskými matematiky téměř neznámým pojmem. Lepší pozici na tomto poli měli v Nizozemí astronomové a fyzikové. Když van de Hulst poslal své poznámky van Dantzigovi, van Dantzig pisatele pokáral za nedostatek matematické přesnosti. Nakonec ho však přece jenom povzbudil, aby v započaté práci pokračoval. Nicméně van de Hulst již v roce 1944 začal slavit úspěchy na poli teoretické astronomie, což byl nejspíš důvod k tomu, že opustil svoje bádání na poli robustních statistických metod.

Podívejme se teď na něco z jeho poznámek. Proč se vůbec začal van de Hulst problémem zabývat? K robustním odhadům se dostal díky problému, který řešil Ejnar Hertzsprung[55], ředitel observatoře v Leidenu. Tento problém se týkal určení pohybů hvězd v souhvězdí Plejád. V roce 1942 Hertzsprung popisuje ve věstníku astronomických institutů v Nizozemí svůj experiment, který provedl za účelem určení rozptylu useknutého průměru. Zabývá se otázkou, jak moc je váha výsledku v případě normálního rozdělení chyb oslabena tímto symetrickým zamítnutím odlehlých pozorování. Protože matematické hledání odpovědi považuje za příliš složité, výsledek hledá za pomocí experimentu. Na 12 534 lístečků byla

[52] **Willem Rutger van Zwet** (*1934) – nizozemský mateatik a statistik. Titul Ph.D. získal na univerzitě v Amsterodamu v roce 1964. Působil mj. na Mathematish Centrum v Amsterodamu, University of Leiden, jako hostující profesor pak např. na University of Oregon nebo University of California v Berkeley. V letech 1992–1999 zastával pozici ředitele Thomas Stieltjes Research Institute. Získal řadu ocenění, mezi jinými Van Dantzig Award (Netherlands Statistical Society, 1970), Bernoulli Medal (Taškent, 1986), Médaille de la Ville de Paris (1989), Adolphe Quételet Medal (International Statistical Institute, 1993), AKZO-Nobel Award (Nizozemí, 1996). Ocenění se mu dostalo i u nás. Roku 1997 získal čestný doktorát na Univerzitě Karlově v Praze.

[53] **Hendrik Christoffel van de Hulst** (1918–2000) – nizozemský fyzik a astronom. Narodil se v Utrechtu jako šesté dítě známého spisovatele knih pro děti. Jeho studia na univerzitě v Utrechtu byla přerušena v roce 1939 všeobecnou moblizací. Jako astronom se zabýval rozptylem světla. Je autorem knih *Light Scattering by Small Particles* (1957) a *Multiple Light Scattering* (1980).

[54] **David van Dantzig** (1900–1959) – věnoval se algebře, diferenciální geometrii, elektromagnetismu, termodynamice, pravděpodobnosti a statistice. V roce 1927 se stal asistentem na Delftské Technické univerzitě. V roce 1932 získal doktorát za svoji práci *Studiën over topologische Algebra*. V roce 1940 se stal řádným profesorem. Nicméně ještě téhož roku byl při německé okupaci Holandska donucen i s rodinou opustit Haag a odstěhovat se do Amsterodamu. Ke statistice se dostal právě během druhé světové války. Po válce byl jmenován profesorem na univerzitě v Amsterodamu. Zde se věnoval především statistice a vzdělávání nové generace matematických statistiků. Byl jedním ze zakladatelů Mathematish Centrum v Amsterodamu.

[55] **Ejnar Hertzsprung** (1873–1967) – astronom a chemik. V letech 1919–1946 pracoval ve hvězdárně v Leidenu v Nizozemí, od roku 1937 na pozici ředitele. Jeho jméno známe především díky tzv. Hertzsprung-Russellovu diagramu, který v letech 1911–1913 odvodil společně s H. N. Russellem.

zapsána čísla symetrická kolem nuly. Tato čísla byla zapisována na dvě desetinná místa jako odchylky od nulové střední hodnoty tak, aby vyjadřovaly normální rozdělení. Takže 50 lístečků bylo popsáno číslem 0,00, dalších 50 lístečků neslo 0,01, dalších 50 mělo na sobě 0,01 atd. Poté byl tisíckrát vybrán vzorek o velikosti 24. Každá z těchto 24-prvkových množin byla seřazena podle velikosti odchylky a byl počítán průměrný čtverec sumy těchto odchylek tak, že byl spočítán nejprve pro všech 24 hodnot. Poté byly dvě extrémní hodnoty (nejmenší a největší) zamítnuty a počítalo se pouze s 22 zbývajícími hodnoty. Tento postup se opakoval, až se počítalo pouze se dvěma posledními hodnotami.


Obrázek 3.3. *Hendrik Christoffel van de Hulst a Ejnar Hertzsprung*

Abychom mohli uvést Hertzsprungovy výsledky, musíme zavést jistou symboliku. Nechť tedy $X_1,\ldots,X_n$ jsou nezávislá pozorování, stejně rozdělená se střední hodnotou nula, konečným rozptylem a s hustotou f, která je symetrická kolem nuly. Dále nechť $X_{n:1} < X_{n:2} < \ldots < X_{n:n}$ jsou uspořádaná pozorování podle velikosti. Useknutý průměr pak označíme jako[56]

$$\overline{X}_{n,k} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X_{n:i}\,.$$

Příslušný rozptyl je potom

$$\sigma_{n,k} = E\overline{X}_{n,k}^2 - \left(E\overline{X}_{n,k}\right)^2 = E\overline{X}_{n,k}^2\,.$$

Hertzsprung se zabývá případem, kdy hustota f je hustota standardizovaného normálního rozdělení, $n = 24$ a $k = 0, 1, 2, \ldots, 11$. Hertzsprung si pak definuje veličiny

[56] **Useknutý průměr** – zde používám trochu jinou notaci než později pro useknutý průměr v další části. Zde se totiž bude jednat o „ořezávání“ konkrétního počtu pozorování (dvou, čtyř, …), zatímco později se budeme bavit o „usekávání“ nějaké procentní části pozorování. V tomto případě mi tak tato symbolika přišla vhodnější.

$$n\sigma_{n,k} = \frac{\sigma_{n,k}}{\sigma_{n,0}},$$

které odhaduje pomocí příslušných poměrů výběrových rozptylů tisíce vzorků. Dostal se k následujícím výsledkům:

k	$s^2_{n,k}/s^2_{n,0}$	k	$s^2_{n,k}/s^2_{n,0}$	k	$s^2_{n,k}/s^2_{n,0}$
0	1,000	4	1,095	8	1,283
1	1,013	5	1,139	9	1,345
2	1,037	6	1,184	10	1,407
3	1,069	7	1,232	11	1,489

Tabulka 3.1. *Hertzsprungovy odhady* $n\sigma_{n,k}$ *pro* $n = 4$ *a* $k = 0, 1, 2, \ldots, 11$

Nakonec ještě Hertzsprung ukazuje formuli, která poměrně dobře vystihuje data v tabulce 3.1.:

$$n\sigma_{n,k} = +0{,}53 \cdot (k/n)^{3/2}.$$

Van de Hulst se rozhodl nalézt $n\sigma_{n,k}$ matematicky. Ve svých dopisech Hertzsprungovi z dubna a června 1942 počítá hodnoty $\sigma_{n,11}$ a $\sigma_{n,2}$. I poté se zabývá tímto problémem a své výsledky si zapisuje do svého zápisníku. Zajímavé ale je, že van de Hulst začíná své zápisy s asymptotickým rozptylem *M*-odhadu (z [88]):

Věta: Pokud n pozorování X_i *(n velmi velké) má rozdělení se symetrickou hustotou f a určíme číslo M podle*

$$\sum_{i=1}^{n} \psi(X_i - M) = 0,$$

kde ψ *je nějaká lichá funkce, pak*

$$EM^2 = \frac{\int \psi^2(x) f(x)\,dx}{n\left[\int f(x)\,d\psi(x)\right]^2}.$$ [57]

[57] *Theorem.*

Van de Hulst píše, že tento výsledek je dobře známý a odkazuje se na Zernikeho[58] kapitolu *Pravděpodobnost a matematická statistika* ve třetím svazku *Handbuch der Physik* z roku 1928. Je překvapující, že již v roce 1928 byl fyzikům známý koncept *M*-odhadu. Nicméně není zde uveden důkaz tohoto tvrzení. Van de Hulst tento důkaz poskytuje. Ve svém dokazování však nepostupuje striktně matematicky, což je věc, za kterou ho kritizoval ve svém dopise z února 1943 van Dantzig.

Vraťme se ale zpátky k useknutým průměrům. Nechť tedy $X_1,\ldots,X_n$ jsou nezávislá pozorování, stejně rozdělená se střední hodnotou nula, konečným rozptylem a s hustotou *f*, která je symetrická kolem nuly, *F* je distribuční funkce příslušná k *f*. Dále opět uvažujme useknutý průměr a jemu příslušný rozptyl ve tvaru:

$$\overline{X}_{n,k} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X_{n:i},$$

$$\sigma_{\cdot,k} = \mathrm{E}\overline{X}^2_{n,k} - \left(\mathrm{E}\overline{X}_{n,k}\right)^2 = \mathrm{E}\overline{X}^2_{n,k}.$$

Pokud nyní označíme odseknutou část na každé straně jako α a a bude horní α-bod distribuční funkce *F*, tj.

$$\alpha = k/n,\ F(a) = 1-\ \cdot.$$

Nechť funkce ψ je definována jako

$$\psi_{\cdot} = \begin{cases} -a & \text{pro } x < -a, \\ x & \text{pro } -a \le x \le a, \\ a & \text{pro } x > a. \end{cases}$$

Van de Hulst ukazuje, že pokud je α dostatečně daleko od 1/2 a $n \to$, pak

$$n\sigma_{\cdot,k} \overset{as.}{\approx} (1-2\alpha)^{-2}\int \psi_{\cdot}(x)^2 f(x)\,dx.$$

If n obeservations X_i (n very large) are distributed according to the symmetric probability (density) f, and one determines the number M by $\sum_{i=1}^{n}\psi(X_i - M) = 0$, where ψ is some odd function, then $EM^2 = \dfrac{\int \psi(x)^2 f(x)\,dx}{n\left[\int f(x)\,d\psi(x)\right]^2}$. [88, str. 84]

[58] **Frits Zernike** (1888–1966) – holandský fyzik. Doktorát získal v roce 1915 na univerzitě v Amsterdamu. Roku 1953 získal Nobelovu cenu za fyziku za zkonstruování fázově-kontrastního mikroskopu. Na princip fázově-kontrastního mikroskopu přišel při zkoumání vad optických přístrojů.

Toto však dokázal už Percy John Daniell, o kterém byla zmínka v předchozím textu. Nicméně, jak bylo také uvedeno, jeho práce zůstala velice dlouho nepovšimnuta. Van de Hulstův důkaz lze nalézt v [88].

Pokud funkce *f* bude rovna hustotě standardizovaného normálního rozdělní ϕ, pak můžeme vztah pro $n\sigma_{\ldots,k}$ upravit do tvaru:

$$n\sigma_{\ldots,k} \approx \left(-\ldots\right)\left(-\ldots - \ldots a\phi\left(\ldots\right) + \ldots\right)^{2}.$$

Pokud dosadíme $n = \ldots$, dostaneme hodnoty uvedené v tabulce 3.2.

k	$n\sigma_{\ldots,k}$	k	$n\sigma_{\ldots,k}$	k	$n\sigma_{\ldots,k}$
0	1,000	4	1,114	8	1,294
1	1,021	5	1,153	9	1,352
2	1,048	6	1,195	10	1,417
3	1,079	7	1,242	11	1,483

Tabulka 3.2. Asymptotické hodnoty $n\sigma_{\ldots,k}$ pro $f = \ldots$, $n = \ldots$ a $k = \ldots, 1, 2, \ldots, 11$.

Tyto hodnoty docela dobře souhlasí s Hertzsprungovými empirickými daty v tabulce 3.1. Když ale van de Hulstovi výsledky komentoval van Dantzig, vyjádřil se následovně [88]:

Vcelku si myslím, že nemá smysl pokoušet se vysvětlit malé odchylky mezi vašimi výpočty a Hertzsprungovými výsledky; podle mého mínění shoda je příliš dobrá na to, aby to bylo správně.[59]

Van de Hulst se také zabýval asymptotickým rozptylem mediánu. Po několika neúspěšných pokusech se mu podařilo najít pro *n* (počet hodnot, ze kterých je počítán medián) liché a $f = \ldots$ rozvoj v řadu

$$n\sigma\left(\ldots_{n}\right) = \frac{\pi}{2}\left(1 - \frac{4-\pi}{2n} - \frac{3\pi-4}{n^{2}} + \frac{13\pi^{2}}{24n^{2}} + \ldots\right).$$

[59] *On the whole, I don't think it makes sense to try to explain the small deviations between your calculations and Hertzsprung's results; in my opinion, the agreement is too good to be true.* [88, str. 90]

Tento výsledek ovšem nelze použít pro $n = 24$. Van de Hulst se také pokoušel zlepšit aproximaci pro $\sigma_{n,1}$. Použil k tomu variační rozpětí, definované jako:

$$M_n = \frac{1}{2}(X_{1:n} + X_{n:n})$$

Odvodil vztah mezi rozptyly useknutého průměru $\bar{X}_{n,1}$, průměru $\bar{X}_{n,0}$ a variačního rozpětí M_n v případě, že $f = 6$:

$$\frac{\sigma_{n,1}}{\sigma_{n,0}} - = \frac{4}{(6 - 2)}\left(\frac{EM_n^2}{\sigma_{n,0}} - \right)$$

Numerickými metodami pro $n = 24$ a $f = 6$ nalezl $n\sigma_{n,1} = 1,018$.

V prosinci 1943 píše van de Hulst Hertzsprungovi, o tom, že nalezl matematicky prakticky stejné výsledky jako Hertzsprung, co se týče rozptylů useknutých průměrů 24 pozorování. Dále uvádí, že je možné aplikovat jeho vzorec $n\sigma^2_{n,k} \overset{as.}{\approx} (1 - 2\alpha)^{-2} \int \psi^2(x) f(x)\,dx$ nejen na normální rozdělení, ale na skutečné rozdělení chyb měření včetně těch, obsahujících odlehlé hodnoty. Navrhuje také určit rozdělení chyb v Hertzsprungově pokusu a nalézt optimální procentní část, která by měla být useknuta. Pravděpodobně sám van de Hulst tento problém dále neřešil. Zabýval se ale rozdělením chyb s rozsahem od -4,24 do 4,24. V tomto případě navrhuje 10% oříznutí na obou stranách.

Kapitola 4:

Vývoj robustních odhadů a robustních metod v lineárním regresním modelu

4.1 Robustnost

Slovo „robustní“ (popř. robustnost) bylo v osmnáctém století používáno k vyjadřování o někom, kdo je silný, avšak surový a vulgární. Statistický význam dal slovu až roku 1953 George Edward Pelham Box[60]. Ve svém článku *Non-Normality and Tests of Variances* píše nejprve o tom, které testy jsou v jeho době používané při ověřování shodnosti průměrů a rozptylů v několika skupinách. A zmiňuje, že tyto testy jsou odvozeny za podmínky, že jsou splněny některé předpoklady. Jedním z těchto předpokladů je normální rozdělení pozorování. Poté používá slovo robustnost v následujícím kontextu:

Zdá se, že tato vlastnost „robustnosti“ vzhledem k nenormalitě, kterou tyto testy pro porovnávání středních hodnot mají, a bez níž by byly mnohem méně vhodné pro potřeby experimentátora, není samozřejmá u jiných statistických testů a zvláště ne u testů rovnosti rozptylů, zmíněných výše.[61]

Před rokem 1885 se vědci zabývali tím, co bychom dnes mohli nazvat „robustnost“ ve smyslu necitlivosti procedur na odchylky z předpokladů, především z předpokladu normality. O tomto tématu bylo pojednáno již v předchozí části. V současné době existují různé definice, více či méně matematicky přesné, ale obecně robustní znamená necitlivý na malé odchylky z idealizovaných předpokladů, pro které je odhad optimalizovaný.

Proč se ale vůbec robustní metody začaly rozvíjet? Když se podíváme na klasické statistické postupy, zjistíme, že se většinou jedná o parametrické postupy. Tj. model je určen až na hodnoty několika parametrů, které nabývají reálných nebo vektorových hodnot. V mnoha případech se jedná o parametry rozdělení pravděpodobností náhodných chyb

[60] **George Edward Pelham Box** (*1919) – vystudoval matematickou statistiku na University College v Londýně. Zabýval se robustností, časovými řadami, nelineárními odhady a aplikací statistiky. Box získal mnoho ocenění za svůj přínos statistice, mj. Wilks Memorial Medal od Americké statistické společnosti a Shewhart Medal od American Society for Quality Control.

[61] *It would appear, however, that this remarkable property of „robustness“ to non-normality which these tests for comparing means possess, and without which they would be much less appropriate to the needs of the experimenter, is not necessarily shared by other statistical test, and in particular is not shared by the tests for equality of variances, mentioned above.* [9, str. 318]

měření. Opakem parametrických metod jsou metody neparametrické. Ty jsou nezávislé, případně jen málo závislé, na tvaru rozdělení pravděpodobností. Klasických příkladem jsou pořadové testy statistických hypotéz. Tyto neparametrické metody mají dobré vlastnosti pro celou třídu distribučních funkcí. Nicméně za tuto „univerzálnost“ zaplatíme jistou daň. A tou je v tomto případě ztráta vydatnosti.

Jak bylo uvedeno již v předcházející kapitole, i malé odchylky od normálního rozdělení mohou značně ovlivnit odhady, získané metodou nejmenších čtverců, ale i například klasickým *F*-testem a dalšími klasickými postupy. A právě robustní metody mají tu vlastnost, že si zachovávají dobré vlastnosti v okolí nějakého základního rozdělení pravděpodobností. A na rozdíl od neparametrických testů jsou vydatnější.

4.2 Teoretické základy robustních statistických metod

Protože se dostáváme k moderním robustním odhadům, ráda bych úvodem zavedla několik pojmů, které se budou nadále poměrně často vyskytovat. Spíše však pro „ujasnění pojmů“. Tyto definice vychází z knihy *Robustní statistické metody* [49], kde je možné najít podrobnější informace.

Jedním z požadavků na statistický odhad je fisherovská konzistence, kterou zavedl v roce 1921 Ronald Aylmer Fisher[62]. Z hlediska robustnosti je tato vlastnost odhadu důležitější než jeho nestrannost.

Definice 4.1. Řekneme, že odhad $\hat{\theta}$ založený na pozorováních $X_1, X_2, \ldots, X_n$ s rozdělením pravděpodobnosti P je fisherovsky konzistentním odhadem parametru θ, jestliže, pokud ho zapisujeme jako funkcionál $\theta = T(P_n)$ empirického rozdělení pravděpodobností vektoru $X_1, X_2, \ldots, X_n$, $n = 1, \ldots$, pak platí $T(P) = \theta$.

[62] **Ronald Aylmer Fisher** (1890–1962) – studoval matematiku a astronomii v Cambridge, zajímal se také o biologii. V letech 1915–1919 pracoval jako učitel matematiky a fyziky. V roce 1919 začal pracovat v zemědělské experimentální stanici v Rothamstedu jako biolog. Dosáhl zde významných výsledků ve statistice i genetice. Zabýval se organizací experimentů, analýzou rozptylu, testováním hypotéz. V roce 1921 Fisher zavedl pojem věrohodnosti. Fisher publikoval několik významných spisů, např. *Statistical Methods for Research Workers* (1925), *The Genetical Theory of Natural Selection* (1930), *The design of experiments* (1935). V roce 1929 byl Fisher přijat do Královské společnosti a v roce 1948 získal Darwinovu cenu Královské společnosti. Následovala Copleyho cena Královské společnosti v roce 1955.

Důležitým pojmem při výkladu robustních odhadů je tzv. influenční funkce. Abychom mohli zavést influenční funkci, je nutné se nejprve vypořádat s pojmy Gâteauxova derivace a kontaminace rozdělení.

Definice 4.2. Funkcionál T je diferencovatelný v Gâteauxově smyslu podle P ve směru Q, existuje-li limita

$$T'_Q(P) = \lim_{t \to 0} \frac{T(P + tQ - tP) - T(P)}{t}.$$

$T'_Q(P)$ pak nazveme Gâteauxovou derivací T podle P ve směru Q.

Definice 4.3. Kontaminací P rozdělení Q v poměru t nazveme rozdělení pravděpodobností

$$P_t(Q) = (1-t)P + tQ,$$

kde $P, Q \in \mathcal{P}$ a $t \in \langle 0;1 \rangle$.

Definice 4.4. Influenční funkcí funkcionálu T v rozdělení pravděpodobností P nazveme Gâteauxovu derivaci T podle P ve směru δ_x, $x \in \mathcal{X}$, tedy

$$IF(x; T, P) = T'_{\delta_x}(P) = \lim_{t \to 0} \frac{T(P_t(\delta_x)) - T(P)}{t},$$

kde $P_t(\delta_x) = (1-t)P + t\delta_x$.

Empirický funkcionál odpovídající vektoru pozorování $(X_1, X_2, ..., X_n)$ nyní označíme jako $T_n = T_n(P_n) = T_n(X_1, ..., X_n)$. Pokud k pozorováním $X_1, X_2, \ldots, X_n$ přidáme další pozorování Y, pak nás bude zajímat vliv Y na T_n. Tento vliv budeme charakterizovat jako

$$T_{n+1}(X_1, ..., X_n, Y) - T_n(X_1, ..., X_n) = I(T_n, Y).$$

To nás vede k pojmu citlivost funkcionálu k přidání dalšího pozorování.

Definice 4.5. Citlivostí funkcionálu $T_n(X_1, ..., X_n)$ k přidání dalšího pozorování při daných $X_1, X_2, \ldots, X_n$ nazveme číslo

$$S(T_n) = \sup_Y |I(T_n(X_1, \ldots, X_n), Y)|.$$

Pokud chceme měřit robustnost odhadů, nastává problém, jak to udělat. Existují totiž různé charakteristiky robustnosti. Již bylo řečeno, že influenční funkce je jednou z nejdůležitějších charakteristik statistického funkcionálu. Hodnota influenční funkce $IF(x;T,P)$ měří vliv kontaminace funkcionálu T hodnotou x. Pokud má být tedy funkcionál robustní, měl by mít ohraničenou influenční funkci. Na influenční funkci jsou založeny dvě často používané charakteristiky funkcionálu T – lokální a globální citlivost.

Definice 4.6. Lokální citlivostí funkcionálu T pro rozdělení pravděpodobnosti P nazveme hodnotu

$$\lambda = \sup_{x,y;x\neq y} \left| \frac{IF(y;T,P) - IF(x;T,P)}{y - x} \right|.$$

Tato hodnota měří vliv nahrazení hodnoty x hodnotou y na funkcionál T.

Definice 4.7. Globální citlivostí funkcionálu T pro rozdělení pravděpodobnosti P nazveme hodnotu

$$\gamma^* = \sup_{x\in} |IF(x;T,P)|.$$

Často se také setkáme s další charakteristikou robustnosti odhadu, a tou je tzv. bod selhání.

Definice 4.8. Nechť $\mathbf{x}^{(0)} = (x_1,\ldots,x_n)$ je náhodný výběr, kterému přísluší odhad funkcionálu T s hodnotou $T_n(\mathbf{x}^{(0)})$. Nyní nahraďme libovolných m hodnot v původním výběru $\mathbf{x}^{(0)}$ jakkoliv zvolenými hodnotami. Označme $\mathbf{x}^{(m)}$ nový výběr, který vznikne při co nejnepříznivějších nahrazení co nejnepříznivějšími hodnotami. Příslušnou hodnotu odhadu pak označme $T_n(\mathbf{x}^{(m)})$. Bodem selhání odhadu T_n ve výběru $\mathbf{x}^{(0)}$ pak nazveme číslo

$$\varepsilon^*(T_n, \mathbf{x}^{(0)}) = \frac{m^*(\mathbf{x}^{(0)})}{n},$$

kde $m^*(\mathbf{x}^{(0)})$ je nejmenší celé číslo m, pro které platí

$$\sup_{\mathbf{x}^{(m)}} \left\| T_n(\mathbf{x}^{(m)}) - T_n(\mathbf{x}^{(0)}) \right\| = \infty.$$

Jedná se tedy o nejmenší podíl pozorování, které po nahrazení libovolnými hodnotami mohou vést k hodnotám T_n nekonečno.

4.2.1 Robustní odhady reálného parametru

Nechť $X_1,\ldots,X_n$ je náhodný výběr z populace s rozdělením pravděpodobností P. Naším úkolem je najít odhad parametru θ, který lze vyjádřit jako funkcionál rozdělení P, tedy $T(P)$. Existují tři nejpoužívanější třídy odhadů reálného parametru: M-odhady, L-odhady a R-odhady. Tyto třídy odhadů lze rozšířit např. na lineární regresní model.

1. *M*-odhady

M-odhad T_n je definován jako řešení minimalizace

$$\sum_{i=}^{n} \rho(X_i,\theta) = \min \text{ vzhledem k } \theta \in),$$

kde ρ je vhodně zvolená funkce. Pokud je funkce ρ diferencovatelná vzhledem k θ se spojitou derivací ψ, pak T_n je řešením rovnice

$$\sum_{i=}^{n} \psi(X_i,\theta) =) \qquad \theta \in)$$

Influenční funkci M-odhadu pak můžeme zapsat ve tvaru

$$IF(x;T,P) = \frac{\psi(x,T(P))}{-\int_X \dot{\psi}(y,T(P))\,dP(y)},$$

kde $\dot{\psi}(y,T(P)) = \left[\frac{\partial}{\partial\theta}\psi(y,\theta)\right]_{\theta=T(P)}$.

M-odhad T_n je ekvivariantní vzhledem k posunutí, tj. platí

$$T_n(X_1+c,\ldots,X_n+c) = T_n(X_1,\ldots,X_n)+c, \quad c\in .$$

Nicméně není ekvivariatní vzhledem k měřítku, tj. obecně neplatí

$$T_n(X_1,\ldots,cX_n) = cT_n(X_1,\ldots,X_n) \quad c\in R,\ c>).$$

Zatím jsme blíže nespecifikovali funkci ρ, popř. její derivaci ψ. Mezi takové funkce patří např. Huberova funkce, Andrewsova sinusová funkce nebo Tukeyho funkce biweight, o kterých bude řeč později v souvislosti s jejich objevem. Můžeme se setkat i s odhadem skipped mean, který je generován funkcí

$$\psi(x) = \begin{cases} -1 & -k \le x < 0 \\ 0 & |x| > k \\ 1 & 0 \le x \le k. \end{cases}$$

Jak již bylo řečeno, *M*-odhad není obecně ekvivariantní vzhledem k měřítku. Tuto vadu je možné odstranit tzv. studentizací *M*-odhadu škálovou statistikou $S_n = S_n(X_1,\ldots,X_n)$, která splňuje:

(i) $S_n(\cdot) > 0$ s.v. pro $\mathbf{x} \in \mathbb{R}^n$

(ii) $S_n(x_1 + c,\ldots,x_n + c) = S_n(x_1,\ldots,x_n) + c$, $c \in \mathbb{R}$, $\mathbf{x} = (x_1,\ldots,x_n) \in \mathbb{R}^n$

(iii) $S_n(cx_1,\ldots,cx_n) = cS_n(x_1,\ldots,x_n)$, $c > 0$, $\mathbf{x} \in \mathbb{R}^n$.

Navíc předpokládáme, že

$$\sqrt{n}(S_n - S(F)) = o_p(1) \text{ pro } n \to \infty,$$

kde $S(F)$ je statistický funkcionál S_n.

Tím dostáváme studentizovaný *M*-odhad, který je ekvivariantní k posunutí i k měřítku, jako řešení rovnice

$$\sum_{i=1}^{n} \rho\left(\frac{X_i - \theta}{S_n}\right) = \min \text{ vzhledem k } \theta \in \mathbb{R}.$$

Pokud je funkce ρ diferencovatelná vzhledem k θ se spojitou derivací ψ, pak T_n je řešením rovnice

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta}{S_n}\right) = 0 \qquad \theta \in \mathbb{R}$$

Jako škálová statistika S_n se používá např. výběrová směrodatná ochylka, mezikvartilové rozpětí[63] nebo mediálnová absolutní odchylka[64].

2. *L*-odhady

Tento typ odhadů je založen na uspořádaných pozorováních neboli pořádkových statistikách. Mějme náhodný výběr $X_1, X_2, \ldots, X_n$. Jeho pořádkové statistiky budeme označovat (pozorování uspořádaná podle velikosti) $X_{n:1}, X_{n:2}, \ldots, X_{n:n}$. Tedy platí: $X_{n:1} \le X_{n:2} \le \ldots \le X_{n:n}$. *L*-odhad pak je řešením rovnice

$$T_n = \sum_{i=1}^{n} c_{ni} h(X_{n:i}) + \sum_{j=1}^{k} a_j h^*(X_{n:\lfloor np_j \rfloor + 1}),$$

[63] **Mezikvartilové rozpětí** – rozdíl horního a dolního kvartilu.

[64] **Mediánová absolutní odchylka** – medián z absolutních hodnot rozdílu hodnot X_i a mediánu, tedy $S_n = \underset{1 \le i \le n}{\text{med}} \left| X_i - \tilde{Y} \right|$.

kde $0 < p_1 < \ldots < p_k <$, h, h^* jsou dané funkce, $c_{n1}, \ldots, c_{nn}$ a $a_1, \ldots, a_k$ jsou dané koeficienty. Navíc koeficienty $c_{ni}, 1 \leq \leq \imath$ jsou ohraničeny váhovou funkcí J : ,1 $\rightarrow$ buď jako

$$c_{ni} = \int_{\frac{i-}{n}}^{\frac{i}{n}} J(s)\,ds, i = ,\ldots,n,$$

nebo přibližným způsobem jako

$$c_{ni} = \frac{1}{n} J\left(\frac{i}{n+}\right) \quad i = ,\ldots,n.$$

Jak vidíme, *L*-odhad se tak skládá ze dvou složek. Velká část *L*-odhadů je tvořenou pouze jednou z těchto částí. Z toho také vyplývá označení těchto odhadů jako *L*-odhad typu I a typu II.

Známými příklady *L*-odhadů jsou medián nebo variační rozpětí[65], α -useknutý průměr, definovaný jako

$$\overline{X}_{n\alpha} = \frac{1}{n - 2[n\alpha} \sum_{i=[n\alpha+1}^{n-[n\alpha} X_{n:i},$$

kde $0 < \alpha$: ,5.

Méně známým *L*-odhadem je Giniho průměrná diference (Gini mean difference)

$$G_n = \frac{1}{n \cdot (n-1)} \sum_{i,j=1}^{n} \left| X_i - X_j \right|.$$

Mezi *L*-odhady patří také tzv. α-winsorizovaný průměr

$$\overline{W}_{n\alpha} = \frac{1}{n}\left\{ [n\alpha] X_{n:[n\alpha]+1} + \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n:i} + [n\alpha] X_{n:n-[n\alpha]} \right\}.$$

Tento průměr tedy dostaneme tak, že extrémní hodnoty (nejmenší a největší hodnoty v souboru, stejnou část na obou koncích) nahradíme posledními „neuseknutými“ kvantily.

3. *R*-odhady

Nechť $X_1, X_2, ..., X_n$ je náhodný výběr populace se spojitou distribuční funkcí. Nechť R_i je pořadí X_i mezi $X_1, ..., X_n$, i = 1,…,*n*. Formálně lze pořadí vyjádřit jako $R_i = \sum_{j=}^{n} I\left[X_j \leq X_i\right], i = ,..., n.$ Pak $R_i = \imath F_n(X_i), i = ,..., n$, kde F_n je empirická distribuční funkce $X_1, ..., X_n$. Pořadí jsou invariantní ke třídě ryze monotónních transformací. Pořadové

[65] **Variační rozpětí** – rozdíl maximální a minimální hodnoty ve výběru.

testy mají tu vlastnost, že rozdělení testového kritéria a platnosti hypotézy nezávisí na distribuční funkci pozorování. *R*-odhady jsou inverzí pořadových testů.

Omezme se na případ, kdy $X_1,...,X_n$ mají spojitou distribuční funkci $F(x-\theta)$ se středem symetrie θ. Hypotézu $H_0: \theta = \theta_0$ o středu symetrie testujeme znaménkovým pořadovým testem založeným na statistice

$$S_n(\theta) = \sum_{i=1}^{n} sign(X_i - \theta_0)\, a_n(R_{ni}^+(\theta)),$$

kde $R_{ni}^+(\theta)$ je pořadí $|X_1 - \theta_0|,...,|X_n - \theta_0|$ a $a_n(1) \le ... \le a_n(n)$ jsou dané skóry, generované neklesající skórovou funkcí $\varphi: (0,1) \to R^+, \varphi(u) \ge 0$ jako $a_n(i) = \varphi\left(\frac{i}{n+1}\right), i = 1,...,n$. Jestliže platí $\theta = \theta_0, F(x) + F(-x) = 1, x \in R$, jsou $sign(X_i - \theta_0)$ a $R_{ni}^+(\theta)$ stochasticky nezávislé a $S_n(t)$ je nerostoucí a schodovitá funkce *t*. Z toho vyplývá, že $E_\theta S_n(\theta) = 0$ a $S_n(\theta)$ je za platnosti H_0 symetrické kolem 0. Odhadem θ je hodnota *t*, která je řešením rovnice $S_n(t) = 0$. Kvůli nespojitosti $S_n(t)$ nemusí mít tato rovnice řešení. Proto definujeme *R*-odhad ve tvaru

$$T_n = \frac{1}{2}(T_n^- + T_n^+),$$

$$T_n^- = \sup\{t; S_n(t) > 0\},$$

$$T_n^+ = \inf\{t; S_n(t) < 0\}.$$

R-odhady jsou ekvivariantní vzhledem k posunutí v poloze i ke změně měřítka. Mezi *R*-odhady patří i medián nebo tzv. Hodges-Lehmannův odhad

$$T_{nH} = \mathrm{med}\left\{\frac{X_i + X_j}{2}, 1 \le i \le j \le n\right\}.$$

4.2.2 Robustní odhady v lineárním modelu

Jak už bylo uvedeno dříve, zmíněné robustní odhady lze rozšířit na robustní odhady v lineárním modelu:

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,$$

kde $Y_1,\ldots,Y_n$ jsou jednotlivá pozorování, $\mathbf{x}_i \in \mathbf{R}^p, i = 1,\ldots,n$ jsou dané vektory, $\varepsilon_1,\ldots,\varepsilon_n$ jsou vzájemně nezávislé náhodné chyby se stejnou distribuční funkcí a $\boldsymbol{\beta} \in R^p$ je neznámý vektor, který chceme odhadnout. Regresní maticí pak nazýváme matici

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}.$$

Běžným odhadem tohoto parametru je odhad získaný metodou nejmenších čtverců

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Tento odhad je sice nejlepším nestranným lineárním odhadem parametru $\boldsymbol{\beta}$, nicméně má i jisté nedostatky. Je totiž citlivý k odchylkám od normálního rozdělení chyb ε (pro rozdělení s těžkými konci je odhad metodou nejmenších čtverců naprosto nevhodný) i k odlehlým pozorováním Y_i. Navíc odhad $\hat{\boldsymbol{\beta}}$ je citlivý k odlehlým hodnotám prvků regresní matice $\mathbf{X}$. Proto se hledaly alternativní odhady k metodě nejmenších čtverců, které by byly robustní.

1. *M*-odhady

M-odhad M_n parametru $\boldsymbol{\beta}$ v lineárním regresním modelu je definován jako řešení

$$\sum \rho(Y_i - \mathbf{x}_i'\mathbf{t}) = \min \text{ vzhledem k } \mathbf{t} \in \mathbb{R}^p,$$

kde ρ je vhodně zvolená funkce $\rho\ \mathbf{R} \to$, absolutně spojitá. Takto konstruovaný *M*-odhad je ekvivariantní vzhledem k regresi, tj. platí

$$M_n(\mathbf{Y} + \mathbf{Xb}) = M_n(\mathbf{Y}) + \mathbf{b}, \forall \in {}^p.$$

Není však ekvivariantní vzhledem k měřítku. Pokud bychom chtěli odhad *M*-odhad ekvivariantní jak vzhledem k regresi, tak vzhledem k měřítku, použijeme tzv. studentizaci. O studentizaci byla řeč již v části výkladu robustních metod reálného parametru. Podrobnosti o studentizaci v tomto modelu zde najít v některé z učebnic robustních odhadů, např. v [48, 49].

2. *L*-odhady

Rozšíření *L*-odhadů na lineární regresní model se povedlo až roku 1978, a to Koenkerovi a Bassettovi, kteří definovali tzv. regresní α-kvantil $\hat{\boldsymbol{\beta}}(\alpha)$, $0 < \alpha$:, pro lineární regresní model. Tento regresní α-kvantil je definován za podmínky, že β označuje absolutní člen a regresní matice splňuje podmínku

$$x_{i1} = , i = ,\ldots,n.$$

Regresní α-kvantil $\hat{\boldsymbol{\beta}}(\alpha)$ je pak definován jako řešení

$$\sum_{i=1}^{n} \rho_{\alpha}(y_i - \mathbf{x}_i'\mathbf{t}) = \min, \ \mathbf{t} \in \mathbf{R}^p ,$$

kde $0 < \alpha < 1$ a $\rho_{\alpha}(x) = x(\alpha - I(x < 0))$, $x \in R$. Tato minimalizace se dá řešit simplexovou metodou. Pomocí regresních kvantilů je možné definovat *L*-odhady v lineárním regresním modelu. Známý je např. regresní medián. Jedná se o regresní α-kvantil s α = 0,5.

4.3 Počátky moderních robustních odhadů

V roce 1931 Egon S. Pearson[66] vydává svůj článek *The Analysis of Variance in Cases of Non-Normal Variation* [69]. Zde si všímá citlivosti klasických metod analýzy rozptylu k odchylkám od normálního rozdělení. Své závěry ukazuje na příkladu dvou set testů měření izolačního odporu jistého materiálu. Tyto testy byly provedeny ve společnosti Bell Telephone Laboratories v New Yorku. Pearson zde také ukazuje, že testy pro porovnání dvou rozptylů jsou velmi citlivé na odchylky od normality. Tato citlivost je ještě větší, když se jedná o porovnání více než dvou rozptylů.

Charles Frederick Mosteller[67] v roce 1946 vydává svůj článek *On some useful „inefficient" statistics* [66]. V něm představuje třídu odhadů, které jsou jednoduchou lineární kombinací malého počtu pořádkových statistik.

V roce 1953 E. P. Box vydává svůj, již zmiňovaný, článek *Non-Normality and Tests on Variances* [9]. Ve své práci mj. shrnuje poznatky, které byly v této oblasti učiněny. Závěrem píše:

Věřím, že robustnost je ve skutečnosti ještě důležitější než fakt, že by testy měly mít maximální sílu a že by použitá statistika měla být zcela účelná… Na druhé straně si nemyslím, že nutně potřebujeme chodit do extrémů používáním neparametrických testů, když mohou být nalezeny silnější robustní parametrické testy.[68]

---

[66] **Egon Sharpe Pearson** (1895–1980) – syn Karla Pearsona. Spolupracoval s Neymanem (Neyman-Pearsonova teorie testování statistických hypotéz). V roce 1955 byl oceněn Zlatou medailí Královské statistické společnosti. V letech 1955–56 byl zvolen prezidentem této společnosti.

[67] **Charles Frederick Mosteller** (1916–2006) – americký statistik. Byl jedním ze zakladatelů Harvardské fakulty statistiky. Přestal vyučovat v roce 1987, ale pokračoval v práci a publikování na Harvardu až do roku 2003. Mosteller napsal přes 50 knih a 350 spisů s více jak 200 spoluautory.

[68] *The property of robustness I belive to be even more important in practice than that the test should have maximum power and that the statistics employed should be fully efficient… On the other hand, I do not think that we need necessarily go to the extreme of using nonparametric tests when it may well be that more powerful robust parametric tests can be found.* [9, str. 333–334]

O dva roky později Box společně s Andersonem [10] prezentují názor, že dobrý statistický test (obecně postup) má být necitlivý na změny parametrů, které se ho netýkají, nebo jsou přímo rušivé. Naopak má být citlivý ke změnám parametrů, které jsou středem jeho zájmu. Kladou tak požadavek vydatnosti.

V roce 1963 vychází článek Hodgese[69] a Lehmanna[70] *Estimates of Location Based on Rank Tests* [36]. Hned v úvodu se zabývají otázkou, že *t*-testy a *F*-testy jsou velice citlivé na velké chyby. Mnohem lépe jsou na tom, podle jejich názoru, Wilcoxonův test a Kruskal-Wallisův *H*-test. Ve třech kapitolách tak pojednávají o odhadech polohy. V dalších kapitolách jsou dokazovány některé vlastnosti těchto odhadů. Z našeho hlediska je podstatné, že se tu objevují tzv. *R*-odhady, které jsou inverzí pořadových testů. *R*-odhady mají tu výhodu, že jsou ekvivariantní nejen vzhledem k posunutí v poloze[71], ale také ke změně měřítka[72].

Odhad, který dnes známe pod názvem Hodges-Lehmannův odhad, můžeme psát jako:

$$T_{nH} = med\left\{\frac{X_i + X_j}{2} : 1 \le i \le j \le n\right\},$$

kde $X_1, X_2, ..., X_n$ je náhodný výběr z populace se spojitou distribuční funkcí. Jak vidíme, jedná se o medián z průměrů všech možných dvojic, tj. i takových, kde $i = j$. Speciálním *R*-odhadem je tak i medián. Nicméně pouze tyto dva *R*-odhady (medián a Hodges-Lehmannův odhad) se dají vyjádřit explicitně. Ostatní musí být počítány iteračně.

4.4 Peter Jost Huber

V roce 1964 Peter J. Huber[73] vydává článek *Robust Estimation of a Location Parameter*. V něm se objevuje hned několik důležitých novinek. Jak uvádí Huber hned v úvodu [41]:

---

[69] **Joseph Lawson Hodges** (1922–2000) – po získání titulu PhD. v roce 1949 zůstal v Berkeley po zbytek života, kromě krátkého působení na University of Chicago (1951–1952) a ve Švédsku (1956–1957). Jeho dlouholetým přítelem a častým vědeckým spolupracovníkem byl Erich Lehmann. V letech 1961–1964 působil jako editor *The Annals of Mathematical Statistics*.

[70] **Erich Leo Lehmann** (*1917) – německý statistik, většinu svého života ale prožil v USA. Zabýval se neparametrickým testováním hypotéz. Byl editorem *The Annals of Mathematical Statististics*, ředitelem Institute of Mathematical Statistics, členem National Academy of Science.

[71] **Ekvivariantní vzhledem k posunutí v poloze** - $T_n(X_1 + c, ..., X_n + c) = T_n(X_1, ..., X_n) + c, \quad c \in$.

[72] **Ekvivariantní ke změně měřítka** - $T_n(X_1, ..., cX_n) = cT_n(X_1, .., X_n), \quad c \in R, c >$).

[73] **Peter Jost Huber** (*1934) – narozen v Wohlenu ve Švýcarsku. Doktorskou práci z matematiky (*Homotopy Theory in General Categories*) napsal na švýcarské Eidgenössiche Technische Hochschule v Curychu. Poté ale přešel ke statistice. V letech 1961–1963 působil na Berkeley, kde také napsal svojí první a nejznámější práci z robustní statistiky *Robust Estimation of a Location Parameter*. Po působení na Cornellově univerzitě odešel na Eidgenössiche Technische Hochschule. V letech 1978–1988 pracoval na Harvardské univerzitě, v letech 1988–1992 na MIT a poté až do svého odchodu do důchodu na University of Bayreuth.

Tato práce obsahuje nový přístup k teorii robustních odhadů; podrobně pojednává o asymptotické teorii odhadu parametru polohy pro kontaminovaná normální rozdělení a ukazuje odhady – prostředníky mezi prostým průměrem a mediánem – které jsou asymptoticky nejrobustnější (ve smyslu, který bude specifikován) mezi všemi odhady invariantními vzhledem k posunutí.[74]

Huber předpokládá, že pozorování $X_1,\ldots,X_n$ jsou nezávislá a pochází z rozdělení s distribuční funkcí $F(x-\theta)$. Úkolem je odhadnout parametr polohy θ. Huber zde zavádí kontaminované rozdělení $F=(1-\varepsilon)\Phi+\varepsilon H$, kde $0\le\varepsilon<1$ je známé číslo, $\Phi(x)=(2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}s^2\right)ds$ je distribuční funkce standardizovaného normálního rozdělení a H je neznámé kontaminující rozdělení. Takové rozdělení můžeme využít v případě, kdy předpokládáme, že pozorování pocházejí z normálního rozdělení s rozptylem 1, ale část ε z těchto pozorování je ovlivněna velkými chybami.

Dále se Huber zabývá metodou nejmenších čtverců, jak byla zavedena a jaké mohou mít malé odchylky od normality následky. Pokládá si otázku, jestli by nemohl být získán robustnější odhad minimalizováním jiné funkce chyb než součtu jejich druhých mocnin. Tím se dostává k zavedení odhadu pro parametr polohy:

$T=T_n(X_1,\ldots,X_n)$, který minimalizuje $\sum_{i=1}^{n}\rho(X_i-T)$, kde ρ je nekonstantní funkce.

Jedná se vlastně o zavedení *M*-odhadů, jak je i Huber pojmenovává[75]. Pokud položíme $\rho(t)=t^2$, dostáváme prostý průměr. Při $\rho(t)=|t|$ dostaneme jako odhad medián.

Huber zároveň navrhuje funkci ρ:

$$\rho(t)=\frac{1}{2}t^2 \quad \text{pro } |t|<k,$$

$$\rho(t)=k|t|-\frac{1}{2}k^2 \quad \text{pro } |t|\le k.$$

Zároveň uvádí, že takový odhad je nejrobustnější mezi všemi odhahy ekvivariantními vzhledem k posunutí. Dále, že prostý průměr ($k=\infty$ a medián ($k=0$) jsou hraniční případy

[74] *This paper contains a new approach toward a theory of robust estimation; it treats in detail the asymptotic theory of estimating a location parameter for contaminated normal distributions, and exhibits estimators – intermediaries between sample mean and sample median – that are asymptotically most robust (in a sense to be specified) among all translation invariant estimators.* [41, str. 73]

[75] Huber *M*-odhady vlastně zde pojmenovává (*M*)-odhady. Nicméně brzy se začalo používat označení bez závorek.

odpovídající $\varepsilon = 1$ a $\varepsilon =$ a tento odhad je úzce spojen a asymptoticky roven winsorizovanému průměru[76]. Odhad generovaný touto funkcí ρ je necitlivý k odlehlým pozorováním.

Huber také studuje asymptotickou normalitu *M*-odhadů pro konvexní i nekonvexní funkci ρ. V příkladech počítá asymptotický rozptyl a dochází k následujícím závěrům (volně přeloženo podle [41]):

(i) Prostý průměr je velmi citlivý ke chvostům *F*.

(ii) Medián je velmi citlivý k průběhu F v mediánu a nevšímá si chování jinde.

(iii) „Winsorizování“ netrpí těmito nedostatky. Zřejmě je to spojeno s faktem, že příslušná funkce ψ je monotónní, ohraničená a spojitá.

(iv) „Ořezávání“ je spíše citlivé k průběhu *F* v bodech oříznutí $\pm$. Vysoká hustota v těchto bodech může zničit odhad. Tyto nedostatky jsou běžné i u jiných procedur, kdy se zamítají pozorování; zde by bylo možné se tomuto problému vyhnout pomocí vyhlazení ρ v $\pm$.

V páté části své práce se Huber zabývá minimaximálně robustními odhady. Symbolem *C* je označena množina všech distribučních funkcí tvaru $F = (1-\varepsilon)G + \varepsilon H$, kde $0 \le \varepsilon <$ je pevně dané číslo, G je pevná a H proměnná distribuční funkce. Předpokládejme, že G je distribuční funkce s dvakrát diferencovatelnou hustotou g takovou, že $-\log g(t)$ je konvexní v t. Nechť T_n je M-odhad s příslušnou funkcí ρ, $\psi = \rho'$, $c = \psi(F)$ tak, že $\int \psi(t - c(F))\,dF(t) = 0$. Pak asymptotický rozptyl je:

$$V(\psi, F) = \frac{E_F \psi^2(t - T(F))}{\left(E_F \psi'(t - T(F))\right)^2}.$$

Nyní se Huber pokouší minimalizovat supremum $\sup_F V(\psi, F)$. Dokazuje tak následující větu (přeloženo z [41]):

Věta: Asymptotický rozptyl $V(\psi, F)$ má sedlový bod: existuje $F_0 = (1-\varepsilon)G + \varepsilon H_0$ a ψ_0 tak, že

[76] **Winsorizovaný průměr** – je definován jako:

$$\overline{W}_{n\alpha} = \frac{1}{n}\left(\lfloor n\alpha \rfloor X_{n:\lfloor n\alpha \rfloor + 1} + \sum_{i=\lfloor n\alpha \rfloor + 1}^{n - \lfloor n\alpha \rfloor - 1} X_{n:i} + \lfloor n\alpha \rfloor X_{n:n-\lfloor n\alpha \rfloor} \right).$$

Tento průměr vlastně vypočítáme tak, že pozorování uspořádáme podle velikosti a jistou část nejmenších a největších hodnot nahradíme kvantilem $X_{n:\lfloor n\alpha \rfloor}$ nebo $X_{n:n-\lfloor n\alpha \rfloor}$.

$$\sup_F V(\psi, F) = V(\psi, F_0) = \inf_\psi V(\psi, F_0),$$

kde F probíhá distribuční funkce z C, pro které platí $E_F\psi = 0$. Nechť $t_0 < t_1$ jsou koncové body intervalu, kde $|g'/g| \le k$ (jeden nebo oba tyto body mohou být nekonečno) a k je s ε ve vztahu

$$\frac{1}{1-\varepsilon} = \int_{t_0}^{t_1} g(t)\,dt + \frac{g(t_0) + g(t_1)}{k}.$$

Potom hustota f_0 příslušející F_0 je:

$$\begin{aligned} f_0(t) &= (1-\varepsilon)\,g(t_0)\,e^{k(t-t_0)} && \text{pro } t \le t_0, \\ &= (1-\varepsilon)\,g(t) && \text{pro } t_0 < t < t_1, \\ &= (1-\varepsilon)\,g(t_1)\,e^{-k(t-t_1)} && \text{pro } t \ge t_1. \end{aligned}$$

$\psi = -f_0'/f_0$ je monotónní a ohraničená a odpovídá maximálně věrohodnému odhadu parametru polohy, pokud F_0 je výchozí rozdělení.

Tato věta je ještě doplněna poznámkou, že je možné omezit G na symetrickou distribuční fukci a nechat H probíhat přes všechny symetrické distribuční funkce. Tím se vyhneme tomu, že třída, přes kterou H probíhá, závisí na ψ.

Důležitý je i speciální případ minimaximálně robustního odhadu v modelu kontaminovaného normálního rozdělení. Budeme postupovat podle předchozí věty, kdy za G dosadíme distribuční funkci standardizovaného normálního rozdělení ϕ. Stejně postupoval i Huber. Použil odhad $T_n = T_n(X_1, \ldots, X_n)$, který minimalizuje

$$\sum_{i=1}^{n} \rho(X_i - T),$$

kde

$$\rho(t) = \frac{1}{2}t^2 \text{ pro } |t| < k,$$

$$\rho(t) = k|t| - \frac{1}{2}k^2 \text{ pro } |t| \ge k,$$

$\psi = \rho'$, ε a k jsou ve vztahu:

$$\frac{1}{1-\varepsilon} = \int_{-k}^{k} \rho(t)\,dt + \varphi(k)\,k.$$

Pak jako nejméně příznivé rozdělení Huberovi vychází to s hustotou:

$$f_0(t) = (1-\varepsilon)(2\pi)^{-\frac{1}{2}} e^{-\rho(t)}$$

V dalších kapitolách své práce se Huber ještě zabývá např. případem, kdy kontaminující funkce H nebude symetrická, minimaximální teorií a odhadem parametru měřítka. Svůj minimaximální přístup Huber ukazuje na příkladu hry mezi přírodou a statistikem, kdy příroda si vybírá rozdělení dat v okolí modelu a statistik si vybírá odhad v dané třídě. Výplatou statistikovi je asymptotický rozptyl odhadu v daném rozdělení.

4.5 Princetonská studie

Významnou událostí z hlediska rozvoje moderních robustních odhadů je seminář v Princetonu, kde působil jako profesor statistiky John Wilder Tukey[77], v akademickém roce 1970–1971. Původně se předpokládalo, že tento rok dlouhý seminář vyústí v konferenci. K tomuto výzkumu přispěli nejen zaměstnanci a studenti Princetonu, ale i pozvaní hosté z jiných univerzit i z Bell Telephone Laboratories. Základem této skupiny byli: David F. Andrews[78], Peter J. Bickel[79], Frank R. Hampel[80], Peter J. Huber, W. H. Rogers a John W. Tukey. Princetonská skupina upozornila na nedostatky klasických odhadů a ukázala vlastnosti několika možných alternativních (robustních) odhadů. Na konci května 1971 byly příspěvky vybrány a následně posouzeny, doplněny a uspořádány. V roce 1972 tak vychází v Princetonu kniha *Robust Estimates of Location: Survey and Advances*. Jak už napovídá sám název, kniha je zaměřena výhradně na odhady polohy. Nicméně nepublikované semináře a neformální diskuze se zaměřovaly na širší okruh témat. Autoři této knihy si stanovili dva úkoly:

(i) poznat další vlastnosti robustních odhadů,

(ii) odvodit a studovat nové odhady.

Z tohoto důvodu studovali celkem 68 odhadů. Některé z nich byly již známé, jiné byly odvozeny až v průběhu tohoto bádání. Část z těchto odhadů je možné počítat ručně, další

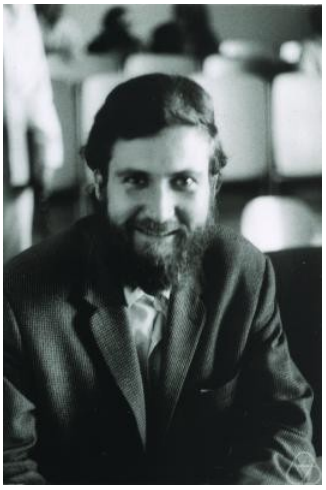
[77] **John Wilder Tukey** (1915–2000) – americký statistik. Vystudoval matematiku a chemii na Brown University. Doktorát z matematiky získal v roce 1939 na univerzitě v Princetonu. Během druhé světové války Tukey působil v úřadu Fire Control Research, kde se zabýval matematickou statistikou. Po válce zastával v Princetonu místo statistika. Tukey krom toho pracoval ještě v AT&T Bell Laboratories. V roce 1965 společně s J. W. Cooleym publikuje důležitý rychlý algoritmus Fourierovy transformace. Tukey poukazoval nejen na užitečnost, ale také na omezení matematické statistiky. Tukey se zabýval také analýzou rozptylu.

[78] **David Francis Andrews** (*1943) – až do konce roku 2001 působil jako profesor na University of Toronto. Zajímal se o statistickou genetiku.

[79] **Peter J. Bickel** (*1941) – profesor statistiky na University of California v Berkeley. Jeho výzkum zahrnuje několik oblastí – semiparametrické modely (je spoluautorem knihy *Efficient and Adaptive Estimation for Semiparametric Models*), Markovovy procesy, rozpoznávání řeči, molekulární biologie.

[80] **Frank R. Hampel** (*1941) – narodil se v Heidelbergu. Studoval matematiku a fyziku na univerzitě v Göttingenu a Mnichově a statistiku na University of California v Berkeley, kde získal v roce 1968 titul PhD. Od roku 1979 až do svého odchodu do důchodu v roce 2006 působil na ETH v Curychu. Zabýval se robustní statistikou, small sample asymptotics, statistickou analýzou dat.

vyžadují počítačové zpracování. Počítačové programy, které byly použity, jsou uvedeny v dodatku knihy. Ve zkoumání byly použity jak malé, tak velké soubory.



Obrázek 4.1. *F. R. Hampel a P. J. Huber*

Druhá kapitola obsahuje výčet použitých odhadů. Všechny tyto odhady mají tu vlastnost, že odhadují střed symetrického rozdělení. Na odhady byl mj. kladen požadavek, že všechny odhady musí být popsány algoritmy, které se dají počítačově zpracovat. Jednotlivé odhady jsou označeny určitou značkou, kterou je pak využívána v dalším textu. Někdy je použito slovo, označující celou skupinu odhadů. Např. 5% useknutý průměr je označován prostým symbolem 5%. Všechny useknuté průměry jsou označovány jako „trims“ a mají analogicky označení od 5% do 50%. Slůvko „hampels“ je používano pro celou skupinu Hampelových odhadů se značkami 12A až 25A. Huberovy návrhy H07 až H20 jsou pak označovány jako „hubers“. Do studie jsou tedy zahrnuty jak lineární kombinace pořádkových statistik (useknuté průměry, mediány, lineární kombinace pouze vybraného počtu pořádkových statistik), tak *M*-odhady, skipped odhady a další. Vlastnosti odhadů byly zkoumány při 40 různých typech vzorků. Použity byly vzorky, pocházející z normálního rozdělení, Cauchyho rozdělení, *t*-rozdělení a různých kontaminovaných rozdělení. Skupina kontaminovaných normálních rozdělení byla vytvořena tak, že pevný počet proměnných pocházel z kontaminovaného normálního rozdělení s velkým rozptylem a zbytek ze standardizovaného normálního rozdělení. Navíc byly brány vzorky různých velikostí.

A jaké byly výsledky celé studie? Tukey, Adrews, Hampel, Huber a Bickel vyjádřili svůj názor přímo v sedmé kapitole této knihy. Obecně by se dalo říci, že se nepodařilo jednoznačně určit, který z robustních odhadů je nejlepší. Spíše se podařilo identifikovat skupiny odhadů, které jsou vhodné, mají dobré vlastnosti. Samozřejmě výběr odhadu také

závisí na konkrétní aplikaci a na preferencích konkrétního statistika. Jak uvádí Tukey [1]: *Naučili jsme se, že je lepší uvažovat o třídách souvisejících odhadů než uvažovat o jednotlivých odhadech.*[81] Pokud se naopak podíváme na nejhorší odhad, pak Hampel uvádí aritmetický průměr. Ve vzorcích o velikostech alespoň 10 se badatelé se vcelku shodli, že vhodné jsou *M*-odhady. Konkrétně velice dobře v této studii vycházel např. odhad označovaný 17A, kde ψ je Hampelova po částech lineární funkce:

$$\psi(x) = \operatorname{sgn} x \cdot \begin{cases} |x| & 0 \le |x| < a, \\ a & a \le |x| < b, \\ \dfrac{c-|x|}{c-b}\, a & b \le |x| < c, \\ 0 & |x| \ge c. \end{cases}$$

Takových odhadů je zde konstruováno několik dosazením různých konstant za parametry a, b, c. V případě odhadu 17A $a = 1{,}7; b = 3{,}4; c = 8{,}5$. Vysoce byl také hodnocen *M*-odhad označovaný AMT, kde ψ je Andrewsova sinusová funkce:

$$\psi(x) = \begin{cases} \sin(x/2{,}1) & |x| < 2{,}1\pi \\ 0 & \text{jinak.} \end{cases}$$

Lineární kombinace pořádkových statistik, jako useknutý průměr, v Princetonské studii poněkud zaostávaly. Důvodem byla nemožnost kombinovat dobré lokální vlastnosti s těmi globálními (např. požadavek vysokého bodu selhání). Jako možná alternativa k *M*-odhadům byly vybrány odhady, založené na pořadových testech (jako Hodges-Lehmannův odhad).

Ve velmi malých vzorcích badatelé tak jednotní nebyli. Jak uvádí Huber [40]: *Zdá se, že s nejmenšími vzorky (velikosti 1 až 4) toho nemůžeme příliš dělat: prostý medián je pravděpodobně stejně dobrý jako jakékoliv jiné robustní odhady, zvláště pokud neznáme měřítko.*[82] Nicméně ani pro vzorky o rozsahu 5 až 9 konsenzu dosaženo nebylo. Výsledky shrnuje ve svém článku [6] také Peter Bickel.

Jedním z podstatných výsledků studie byl však sám fakt, že badatelé stanovovali kvality a nedostatky odhadů pomocí Hampelova bodu selhání a influenční funkce.

[81] *We have learned that much more is gained by thinking about families of related estimates than by thiking about individual estimates.* [1, str. 223]
[82] *For the very smallest sizes (range 1 to 4) it seems that not much can be done: the sample median is probably as good there as any other robust estimator, especially if scale is unknown.* [40, str. 3]

4.6 Frank Hampel

Jak již bylo uvedeno v předchozí části, influenční funkci a bod selhání studoval Frank R. Hampel. Psal o nich ve své dizertační práci *Contributions to the Theory of Robust Estimation* z roku 1968 a později v roce 1971 také v článku *A General Qualitative Definition of Robustness* [30] a následně v *The Influence Curve and Its Role in Robust Estimation* [35], který vyšel v roce 1974.

Nicméně koncept influenční funkce v různých podobách se objevil ve statistické literatuře již dříve, např. ve článku *On the Asymptotic Distributions of Differentiable Statistical Functions* [65] z roku 1947, jehož autorem byl Richard von Mises[83].

Ve své dizertační práci si Hampel stanovil určité aspekty robustnosti. Jedním z nich byl požadavek, aby malé změny (ve smyslu porušení) měly pouze malé účinky. Dalším byl požadavek, který vedl k zavedení bodu selhání; jak velké může být porušení, než všechno selže.

Článek z roku 1971 obsahuje dvě úzce související definice robustnosti posloupnosti odhadů, které berou v úvahu typy odchylek z parametrických modelů, které se obvykle vyskytují. Tyto definice využívají vlastnosti Prochorovovy vzdálenosti pravděpodobnostních měr[84]. Hampel předpokládá pozorování, která mohou být popsána nějakým parametrickým modelem (např. model nezávislých pozorování, pocházejících z totožného normálního rozdělení). Úkolem je odhadnout parametry tohoto modelu (nebo nějakou jejich funkci). Nicméně Hampel předpokládá, že tento parametrický model není naprosto korektní. Model se mírně liší. Hampel rozlišuje tři hlavní důvody těchto odchylek od parametrického modelu: zaokrouhlování pozorování, výskyt velkých chyb a fakt, že model může být pouze aproximací

---

[83] **Richard von Mises** (1883–1953) – jeden z nejvýznamnějších aplikovaných matematiků 20. století. Zabýval se mechanikou, hydrodynamikou, aerodynamikou, konstrukční geometrií, diferenciálními a integrálními rovnicemi, teorií pravděpodobnosti, matematickou statistikou a filozofií. Narodil se ve Lvově. Vystudoval gymnázium ve Vídni, poté studoval vídeňskou techniku. Ještě za dob svých studií působil jako asistent na brněnské německé technice. V roce 1908 získal na vídeňské technice doktorát za svoji práci Die Ermittlung der Schwingmassen im Schubkurbelgetriebe. Když mu bylo 26 let, byl jmenován profesorem aplikované matematiky na univerzitě ve Štrasburku. Po první světové válce působil krátce jako profesor hydrodynamiky a aerodynamiky v Drážďanech. V roce 1920 se stává ředitelem na institutu aplikované matematiky na univerzitě v Berlíně. Hned v následujícím roce zakládá časopis *Zeitshrift für angewandte Mathematik und Mechanik*. Nicméně kvůli nástupu Adolfa Hitlera musí odejít z berlínské univerzity a odchází neprve do Istanbulu a v roce 1939 do Spojených států. V roce 1944 v Harvardu získává místo profesora aerodynamiky a aplikované matematiky.

[84] **Prochorovova vzdálenost pravděpodobnostních měr** - definice použitá Hampelem [30]: Nechť $(\Omega, \mathcal{A})$ je měřitelný prostor takový, že Ω je metrický prostor, úplný a separabilní a $\mathcal{A}$ je σ-algebra generovaná topologií. Pro $A \subset \Omega, A \in \mathcal{A}$ nechť A^{ε} označuje množinu všech bodů, jejichž vzdálenost od A (tj. od alespoň jednoho bodu v A) je menší než ε. Nechť P a Q jsou dvě pravděpodobnostní míry (nebo obecněji dvě konečné míry) na $(\Omega, \mathcal{A})$. Pak jejich Prochorovova vzdálenost $\pi(P,Q)$ je definovaná jako

$$\pi(P,Q) = \inf\{\varepsilon : P(A) \le Q(A^{\varepsilon}) + \varepsilon \text{ a } Q(A) \le P(A^{\varepsilon}) + \varepsilon \text{ pro všechny } A \in \mathcal{A}\}.$$

výchozího náhodného mechanismu. Hampel ukazuje, že je možné popsat odchylky mezi rozděleními odhadů pomocí Prochorovovy vzdálenosti. Následuje ještě uvedení konceptu bodu selhání. Zde opět s použitím Prochorovovy vzdálenosti [30].

Definice: Nechť $\{T_n\}$ je posloupnost odhadů. Bod selhání δ^* posloupnosti $\{T_n\}$ v nějaké pravděpodobnostní míře F je definován následovně:

$$\delta^* = \delta^*(\{T_n\}, F) = \sup\{\delta \leq 1 : \exists \text{ kompaktní množina } K = K(\delta), \text{ která je vhodnou podmnožinou parametrického prostoru takovou, že } \pi(F, G) < \delta \Rightarrow G\{T_n\} \in K \} \to 1 \text{ pro } n \to \infty .$$

Práce z roku 1974 se skládá z osmi částí. Článek je pojednáním o influenční funkci (Hampel ji nazývá influenční křivkou), o její interpretaci, vlastnostech a možnostech použití v teorii robustních odhadů.

Influeční funkci zavádí následovně [35]:

Definice: Nechť R je reálná přímka, T reálný funkcionál definovaný na nějaké podmnožině množiny všech pravděpodobnostních měr na R a F označuje pravděpodobnostní míru na R pro kterou je T definován. Označme δ_x pravděpodobnostní míru určenou hromadným bodem 1 v jakémkoliv bodě $x \in R$. Směs F a nějaké δ_x zapíšeme jako $(1-\varepsilon)F + \varepsilon\delta_x$ pro $0 < \varepsilon < 1$. Pak influenční křivka $\mathrm{IC}_{T,F}(x)$ nebo („odhad“) T v („základní pravděpodobnostní distribuci“) F je definována bodově

$$\mathrm{IC}_{T,F}(x) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon\delta_x] - T(F)}{\varepsilon},$$

pokud tato limita je definována v každém bodě $x \in R$.

Následně je určována influenční funkce pro aritmetický průměr, medián, useknutý průměr apod. Z těchto konkrétních příkladů Hampel odvozuje obecný vlastnosti influeční funkce. Ve svém shrnutí mj. píše [35]:

(Influenční funkce) nám dává informace o detailním chování odhadu a o tom, jak jednotlivá pozorování přispívají k odhadované hodnotě.[85]

Ve svém článku [43] z roku 1972 Huber popisuje Hampelovu influenční křivku a poté dodává:

Podle mého názoru Hampelova influenční funkce je nejdůležitější heuristický nástroj pro konstruování robustních odhadů se stanovenými vlastnostmi. Někdo může usilovat o influenční funkce, které jsou ohraničené (aby omezil vliv jednotlivého „chybného" pozorování, které jsou spojité v x (k dosažení necitlivosti proti zaokrouhlování a efektům seskupování) a které jsou spojité jako funkce F (ke stabilizování asymptotického rozptylu odhadu za malých změn F).[86]

Hampel přichází s novým pojmem, a tím je bod selhání. Stručně ho interpretuje [35]: *Je to nejmenší procento kontaminace, které může přivést hodnotu odhadu přes všechny hranice.*[87] Kontaminací se zde míní odlehlé hodnoty, velké chyby i chybné hodnoty. Bod selhání tedy měří globální hledisko robustnosti. Bod selhání se stal široce používaným. Stalo se trendem za vhodné považovat ty odhady, které mají bod selhání roven jedné polovině. Otázkou, zda je nutné používat toto striktní omezení, se zabýval mj. i Stigler. O jeho studii z roku 1977 je pojednáno dále.

Dalším kritériem robustnosti, které Hampel zmiňuje, je globální citlivost (gross-error-sensitivity). Tuto hodnotu ve své dizertační práci označuje jako σ , ve článku z roku 1974 již jako γ^*. Jedná se supremum absolutních hodnot influenční funkce, tedy

$$\gamma^* = \sup_x |IC(x)|.$$

Globální citlivost je dobrou kvalitativní charakteristikou robustnosti, nicméně, jak Hampel uvádí, jsou situace, kdy by měl být člověk při jejich používání opatrný. Jako příklad uvádí odhady, odvozené z pořadových testů.

---

[85] *(Influence function) tell us a lot about detailed behavior of the estimator and about how the separate observations contribute to the estimated value.* [35, str. 386]

[86] *In my opinion, Hampel`s influence function is the most important single heuristic tool for constructing robust estimates with specified properties. One will strive for influence functions which are bounded (to limit the influence of any single „bad" observation), which are reasonably continuous in x (to achieve insensitivity against roundoff and grouping effects) and which are reasonably continuous as a function of F (to stabilize the asymptotic variance of the estimate under small changes of F).* [43, str. 1052]

[87] *It is the smallest percentage of free contamination which can carry the value of the estimator over all bounds.* [35, str. 388]

4.7 Povaha reálných dat

Podmínka, aby bod selhání byl roven jedné polovině, se stala populární. Stalo se tak módou vytvářet odhady, které budou mít bod selhání jedna polovina, a všechny „horší“ odhady (tj. s bodem selhání menším než jedna polovina) považovat za nedostatečné. Ale je opravdu nutné, aby byl tento požadavek na bod zvratu splněn? Je příroda skutečně tak pesimistická? K této otázce se vyslovuje Stigler ve svém článku *Do robust estimators work with real data?* [79] v roce 1977.

Stigler v úvodu zmiňuje, že díky rozvoji počítačů je možné získat velké objemy pseudonáhodných čísel. Výhody simulace jsou známé již od roku 1930, kdy E. S. Pearson propagoval její použití v sérii článků, otištěných v časopisech Biometrika. Díky vhodné transformaci pseudonáhodných čísel může badatel napodobit vzorek z libovolného matematicky definovaného rozdělení pravděpodobností. Badatel má navíc tu výhodu, že přesně ví, jaký mechanismus produkuje jeho data. Nicméně Stigler upozorňuje na hlavní nedostatek takové simulace:

Hlavní nedostatek takové simulace je, že bez ohledu na to, jak dobře badatel zvolí specifikace výběrového rozdělení, neexistuje záruka, že pseudo výběry, které generuje, jsou skutečnou reprezentací reálných dat. Opravdu mnoho simulačních studií robustnosti statistických procedur se zaměřuje spíše na úzký okruh alternativ k normalitě: nezávislé, stejně rozdělené výběry ze symetrických spojitých rozdělení s dlouhými konci. Ale proč by se mělo předpokládat, že reálná data nejsou korelovaná, zešikmená, asymetrická, heterogenní a nemají žádné nespojitosti?[88]

Stigler navrhuje alternativu, a to zhodnotit statistické procedury (v tomto případě robustní odhady polohy) tak, že je aplikuje na skutečná data. Aby se jednalo o reálná data, vybírá datové soubory na základě tří požadavků:

1. Nesmí být porušena integrita dat. Hodnoty musí být v té formě, v jaké byly zaznamenány. Tj. bez odstranění hodnot, které mohl badatel považovat za chybné.

[88] *The principal disadvantage of such simulation is that no matter how clever the investigator is in his choice of specifications for sampling distributions, there is no guarantee that the pseudo-samples he generates are actually representative of real data. Indeed, most simulation studies of the robustness of statistical procedures have concentrated on a rather narrow range of alternatives to normality: independent, identically distributed samples from long-tailed symmetric continuous distributions. But why should real data not be expected to be correlated, biased, asymmetric, heterogeneous and exhibiting some discreteness (or granularity)?* [79, str. 1056]

2. Data musí pocházet z měření dobře definovaných fyzikálních veličin (fyzikálních konstant), které nemusely být známé v době, kdy byla měření dělána, ale dnes jsou již považovaná za známá s jistým stupněm přesnosti.
3. Získávání dat bylo spojeno s relativní neznalostí měřené veličiny, takže se nedá předpokládat zaujatost nebo předsudky badatele.

Na základě těchto požadavků pak vybral historická data, která jsou podle něj k danému účelu nejvhodnější. Získává tak celkem dvacet základních souborů. Osm z nich pochází od Jamese Shorta, který v roce 1761 analyzoval pohyby Venuše, aby tak získal odhad průměrné vzdálenosti Země od Slunce. Devět souborů má svůj původ v určování rychlosi světla. Tyto experimenty prováděl v letech 1870–1888 Albert Abraham Michelson[89] a Simon Newcomb[90]. Poslední tři soubory jsou založeny na zkoumání hustoty Země Henry Cavendishem[91] v roce 1798[92]. Jediným zásahem do dat bylo rozdělení čtyř velkých souborů do měnších tak, aby

[89] **Albert Abraham Michelson** (1852–1931) – narodil se v Polsku, ale již ve svých dvou letech se stěhuje s rodiči do Spojených států. V letech 1880–1882 studoval v Evropě (Berlín, Paříž). Věnoval se především optice. Zpočátku se zabýval analýzou světla a světelných jevů. V letech 1880–1881 provedl v Berlíně pokus týkající se existence tzv. éterového větru. Tento pokus zopakoval s větší přesností v Clevelandu v roce 1887. Pro své pokusy Michelson zkonstruoval přístroj interferometr. Michelson ukázal, že světlo se šíří všemi směry stejnou rychlostí nezávisle na pohybu jeho zdroje. V roce 1892 se stal profesorem fyziky na univerzitě v Chicagu. V letech 1892–1893 Michelson pomocí interferometru srovnával délku metru s délkou světelné vlny a zkoumal strukturu spektrálních čar. Později sestrojil hvězdný interferometr, který umožňoval měřit úhlové průměry hvěz. V letech 1923–1927 byl prezidentem Národní akademie věd USA. V roce 1907 získal Michelson Nobelovu cenu za fyziku.

[90] **Pokusy Michelsona a Newcomba** – Poznání, že světlo cestuje konečnou rychlostí a není přenášeno okamžitě, pochází ze sedmnáctého století. V letech 1849 a 1850 francouzští fyzici Fizeau a Foucault navrhli metody měření rychlosti světla. A právě Michelson s Newcombem využívají ke svým měřením Foucaulltovu metodu, kterou sami ještě později poněkud vylepšili. Foucaultova metoda je založena na průchodu světla od rychle se otáčejícího zrcadla ke vzdálenému pevnému zrcadlu a odtud zpět k rotujícímu zrcadlu. Výpočet rychlosti světla je pak založen na měření vzdáleností mezi zrcadly, rychlosti rotujícího zrcadla a úhlovém posunutí přijatého obrazu od svého zdroje. V roce 1879 Michelson pracoval se vzdáleností zrcadel 600 metrů, získal tak sto určení rychlosti světla. V letech 1880–1882 Newcomb tento experiment překonal tím, že použil vzdálenost 3721 metrů – ze západního břehu Potomac k pevnému zrcadlu, umístěnému u paty Washingtonského pomníku.

[91] **Henry Cavendish** (1731–1810) – chemik a fyzik. Od roku 1749 studoval v Cambridgi. Studium ukončil v roce 1753, až by dosáhl hodnosti. Vrátil se do Londýna, kde dělal pokusy s vedením tepla, magnetismem a elektřinou. Cavendishova samotářská povaha a nechuť publikovat vědecké výsledky vedly k nejasnostem při určování prvenství jeho objevů. Poté, co James Clerk Maxwell vydal v roce 1879 Cavendishovy práce, se zjistilo, že např. Cavendish objevil Coulombův zákon nebo Ohmův zákon ještě předtím, než je objevili vědci, po kterých jsou tyto zákony pojmenovány.

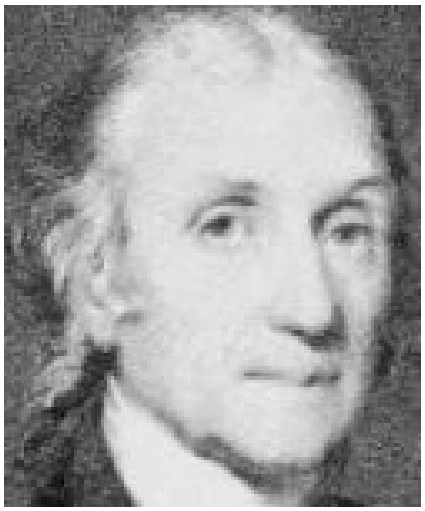
[92] Od konce osmnáctého století a v celém devatenáctém století se objevují pokusy určit střední hustotu Země. Tyto pokusy měly dále sloužit k určení přitažlivosti Země. Newtonův zákon gravitace nám říká, že síla přitažlivosti mezi dvěma hmotnými body je dána vzorcem

$$f = \frac{Gm_1m_2}{r^2},$$

kde m_1, m_2 jsou jejich hmotnosti, r je vzdálenost mezi jejich středy gravitace a G gravitační konstanta. A právě od konce osmnáctého století se vědci snažili určit tuto gravitační konstantu. Při předpokladu, že Země je koule s poloměrem R, se dá Newtonův gravitační zákon přepsat do tvaru:

$$G\Delta = \frac{3f}{4\pi\rho},$$

v každém z těchto nových souborů bylo asi dvacet měření. Tato úprava byla učiněna záměrně, aby bylo možné výsledky této studie srovnávat s již zmíněnou Princetonskou studií z roku 1971.



Obrázek 4.2. *Albert A. Michelson a Henry Cavendish*

Protože již v době, kdy studie vznikala, existovalo velké množství robustních odhadů, nebylo možné (především z hlediska hospodárnosti) zahrnout do porovnávání všechny tyto odhady. Pro porovnání – v Princetonské studii bylo uvažováno 65 základních odhadů. Stigler se rozhodl zařadit pouze 11 odhadů, a to:

1. Průměr $\overline{X}$.

2. Medián $\widetilde{X}$.

3. 10% useknutý průměr. Obecně 100α useknutý průměr je definován:

$$\overline{X}_{\alpha} = \frac{\left(p\,X_{\lfloor \alpha n+ \rfloor} + X_{n-\lfloor \alpha n \rfloor} + \sum_{i=\lfloor \alpha n+ \rfloor}^{n-\lfloor \alpha n+ \rfloor} X_i \right)}{n(1-\alpha)}.$$

4. 15% useknutý průměr.

---

kde Δ je střední hustota Země. Protože G a R můžeme pokládat za známé, určení střední hustoty Země Δ dostaneme změřením f. Z prvních experimentů bývá právě ten od Cavendishe považován za nejlepší. Je to díky jeho úplným popisům pokusů a velice dobrým použitým metodám.

5. 25% useknutý průměr.

Následující tři odhady jsou tzv. *M*-odhady, definované v tomto případě jako řešení (*T*) rovnice

$$\sum_{i=1}^{n} \phi\left(\frac{X_i - T}{s}\right) = 0,$$

kde *s* je odhad rozpětí (zde násobek mediánu absolutních reziduí kolem mediánu nebo kolem dřívější hodnoty odhadu, pokud se tato rovnice řeší iteračně), ϕ je vhodně zvolená funkce.

6. Huberův P15[93] je jednokrokový *M*-odhad, kde

$$\phi(u) = \min(k, \max(-k, u)),$$

k = 1,5, *s* je medián absolutních reziduí kolem mediánu. Tento odhad měl obecně dobré výsledky v Princentonské studii.

7. Andrewsův odhad AMT, kde

$$\phi(u) = \sin\left(\frac{u}{2{,}1}\right) \qquad |u| \leq 2{,}1\pi$$
$$\phi(u) = 0 \qquad \text{jinak,}$$

s je medián absolutní odchylky od předchozí hodnoty *T* (začíná se s mediánem), tento odhad se upravuje každou třetí iteraci.

8. Tukeyho biweight funkce je *M*-odhad, kde

$$\phi(u) = uw(u),$$
$$w(u) = (1 - u^2)^2 \qquad |u| \leq 1,$$
$$w(u) = 0 \qquad |u| > 1.$$

Tento *M*-odhad byl počítán pomocí iterace

$$T_{i+1} = \frac{\sum_j w\left(\frac{X_j - T_i}{cS_i}\right) X_j}{\sum_j w\left(\frac{X_j - T_i}{cS_i}\right)},$$

[93] Stigler používá toto označení, které odpovídá značení odhadů v Princetonské studii, aby bylo možné srovnání se zmíněnou studií. Huber totiž navrhl celou třídu těchto *M*-odhadů (viz odst. 4.5).

která byla provedena šestkrát. Začínalo se s mediánem, $c =$) a S_i jako medián hodnot $\left|X_j - \ _i\right|$.

9. Edgeworthův odhad, který je váženým průměrem dolního kvartilu, mediánu a horního kvartilu. Váhy jsou v poměru 5:6:5. Tento odhad byl Edgeworthem[94] navržen v roce 1893.

10. Outmean $\overline{X}^c_{0,25}$, který je vlastně průměrem těch měření, která byla nezahrnuta do počítání uříznutého průměru $\overline{X}_{0,25}$. Vypočte se tedy

$$\overline{X}^c_{0.25} = \ \overline{X} - \ \overline{\ }_{0,25}.$$

11. Hoggův odhad T_1 je definován

$$\begin{aligned} T_1 &= \ \overline{\ }^c_{0,25} \qquad Q < \ ,0 \\ T_1 &= \ \overline{\ } \qquad 2,0 \le \ ) \le \ ,6 \\ T_1 &= \ \overline{\ }_{3/16} \qquad 2,6 < \ ) \le \ ,2 \\ T_1 &= \ \overline{\ }_{3/8} \qquad 3,2 < \ ). \end{aligned}$$

Q označuje míru „váhy konce“ výběru, a to

$$Q = \frac{U(0,05) - L(0,05)}{U(0,5) - L(0,5)}.$$

$L(\alpha)$ je průměr dolních a $U(\alpha)$ horních 100α) hodnot X_i. Tento odhad byl navržen Robertem V. Hoggem v [37] v roce 1974.

Těchto 11 odhadů bylo počítáno pro každý z 24 datových souborů. Pro srovnání odhadů mezi sebou použl Stigler dva ukazatele. Tím prvním byl index relativní chyby $RE($ pro i –tý uvažovaný odhad. Tento index byl vytvořen k měření absolutní velikosti chyby odhadu vzhledem k velikostem chyb, dosažených jinými odhady pro stejný datový soubor. Pro j-tý datový soubor byla nejprve vypočtena průměrná absolutní chyba:

$$s_j = \frac{1}{11}\sum_{i=} \theta_{\ } - \theta_{\ }\Big|,$$

[94] **Francis Ysidro Edgeworth** (1845–1926) – na Trinity College v Dublinu vystudoval jazyky. V roce 1881 publikuje knihu *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sicences*. V roce 1891 se Edgeworth stává prvním editorem *Economic Journal*, který vydávala Royal Economic Society. Editorem zůstává až do roku 1926. V roce 1892 se Edgeworth zabýval korelací a metodami odhadu korelačních koeficientů. První z článků na toto téma byl *Correlated Averages* (1892). Více o Edgeworthovi např. v [80].

kde θ_j označuje skutečnou hodnotu pro *j*-tý datový soubor a $\hat{\theta}_j$ jsou hodnoty jedenácti odhadů pro *j*-tý datový soubor. Dále byla určena relativní chyba *i*-tého odhadu pro *j*-tý datový soubor jako:

$$e_{ij} = \frac{|\theta_j - \quad|}{s_j}.$$

Pokud je hodnota e_{ij} menší než jedna, znamená to, že pro datový soubor *j* udělal odhad *i* menší chybu než je průměrná chyba pro jedenáct odhadů. Konečně tzv. index relativní chyby $RE(\quad)$ dostaneme pro každý odhad průměrováním přes datové soubory

$$RE(\quad) = \frac{1}{n}\sum_{j=}^{n} e_{ij}$$

Druhým ukazatelem, který Stigler použil pro srovnání odhadů, je index relativního pořadí. Pro každý datový soubor *j* byl nalezen odhad *i* s nejmenší chybou $|\hat{\theta}_j - \quad|$ a tomu byl přiřazena hodnota $r_{ij} =$. Naopak odhad, který měl v tomto datovém souboru největší chybu $|\hat{\theta}_j - \quad|$, dostal přiřazenu hodnotu $r_{ij} = 1$. Poté byl pro každý odhad i vypočítán index relativního pořadí jako průměr přes datové soubory

$$RR(\quad) = \frac{1}{n}\sum_{j=1}^{n} r_{ij}\ .$$

Výsledky těchto výpočtů jsou v tabulce 4. . Indexy byly počítány zvlášť pro velké a malé datové soubory. Čísla v kulatých závorech u indexu relativní chyby se vypočítají jako

$$SE(\quad) = \left(\frac{1}{n-} \sum_{j=}^{n} e_{ij} - RE(\quad)^2 \right)^{1}.$$

Velké hodnoty $SE(\quad)$ odrážejí velké kolísání ve výkonnosti odhadu pro různé datové soubory. Naopak malé hodnoty jsou znakem stabilního výkonu odhadu. Analogicky čísla v kulatých závorkách u indexu relativního pořadí jsou vypočtena jako

$$SR(\quad) = \left(\frac{1}{n-} \sum_{j=}^{n} e_{ij} - RR(\quad)^2 \right)^{1}.$$

	Relative Error (RE)		Relative Rank (RR)	
	Small Samples	Large Samples	Small Samples	Large Samples
Mean	.931 (.20)	.924 (.19)	4.9 (3.2)	6.0 (4.6)
Median	1.149 (.28)	1.152 (.18)	7.1 (4.1)	8.1 (3.8)
Edgeworth	1.018 (.08)	.945 (.07)	6.4 (3.2)	3.9 (1.5)
Outmean	1.038 (.58)	.774 (.50)	5.1 (4.8)	6.0 (5.8)
10% Trim	.916 (.20)	.944 (.06)	4.6 (2.2)	4.5 (2.4)
15% Trim	.983 (.10)	.991 (.04)	6.0 (1.7)	5.5 (0.6)
25% Trim	1.039 (.08)	1.073 (.12)	6.8 (3.0)	6.1 (3.5)
Huber P15	.922 (.20)	.985 (.05)	5.3 (2.8)	5.5 (1.7)
Andrews AMT	.966 (.14)	1.032 (.13)	6.2 (2.5)	6.0 (3.4)
Tukey Biweight	1.023 (.13)	1.097 (.17)	6.6 (3.1)	7.0 (3.9)
Hogg T1	1.014 (.07)	1.084 (.13)	6.8 (2.5)	7.4 (2.5)

Tabulka 4.1. *Indexy relativní chyby a relativního pořadí počítané zvlášť pro malé a velké datové soubory. Malé hodnoty RE a RR ukazují na dobrý výkon odhadu i. Zdroj: [79].*

	Small Samples		Large Samples	
	Relative Error	Relative Rank	Relative Error	Relative Rank
Best	10% Trim Huber P15 Mean	10% Trim Mean	Outmean Mean 10% Trim Edgeworth	Edgeworth 10% Trim
Good	Andrews AMT 15% Trim	Outmean Huber P15	Huber P15 15% Trim	15% Trim Huber P15
Average	Hogg T1 Edgeworth Tukey Biweight Outmean 25% Trim	15% Trim Andrews AMT Edgeworth Tukey Biweight 25% Trim Hogg T1	Andrews AMT 25% Trim Hogg T1 Tukey Biweight	Mean Outmean Andrews AMT 25% Trim
Poor	Median	Median	Median	Tukey Biweight Hogg T1 Median

Tabulka 4.2. *Seřazení jedenácti odhadů podle indexu relativní chyby a indexu relativního pořadí. Zdroj: [79].*

Výsledky Stigler dále shrnuje v tabulce 4.1. Jak ale sám říká, tyto údaje umožňují pouze předběžné. závěry Nicméně i tak poskytují některé překvapivé výsledky. Např. to, že 10% useknutý průměr (odhad velice jednoduchý na provedení) vychází jako jeden z nejlepších uvažovaných odhadů. Zajímavé je i to, že medián, který byl dlouho považován za vysoce robustní a často se používá jako počátek při náročnějších iterativních odhadech, měl jedny

z nejhorších výsledků. Z moderních odhadů, uvažovaných v Princentonské studii, vychází nejlépe Huber P15 a Andrews AMT. Stigler se zde zabývá poněkud provokativní otázkou, jestli vůbec některé z moderních odhadů stojí za čas, který je nutné věnovat jejich výpočtu.

V dalších částech studie je diskutována otázka možného zešikmení dat a systematické chyby v uvažovaných datech.

Nemůžeme zaručit, že situace, které zde zkoumáme, jsou reprezentanty obvyklých aplikací, ale to není dostatečný důvod, abychom založili naše odhady na nerealisticky idealizovaných předpokladech.[95]

V následující části se Stigler věnuje otázce normality. Uvažované datové soubory sice mají trochu těžší konce než normální rozdělení, nicméně, nevykazují žádné významné abnormality. Proto autor uvádí (str. 1070):

(Tabulka a obrázek) ukazují, že uvažované datové soubory mají sklon k poněkud těžším koncům, než je tomu u normálního rozdělení, ale pohled na svět přes Cauchyho brýle se mi zdá být příliš pesimistický.[96]

Naráží tak na fakt, že reálná data se liší od simulovaných dat. A právě simulovaná data se používají ve většině robustních studií. Někteří „pesimističtí" moderní statistikové používají pro svoje výpočty právě Cauchyho rozdělení, které má podstatně těžší konce než rozdělení normální. Závěrem autor shrnuje své poznatky a říká, že nejlepší způsob, jak se u svých zkoumaných souborů vyrovnat s mírným porušením normality, je malé ořezání dat, ne více než desetiprocentní.

Po vydání článku v roce 1977 se rozpoutala diskuze. Některé ohlasy byly vesměs souhlasné (např. reakce George A. Barnarda[97] z University of Waterloo, Davida R. Coxe[98]

[95] *There can be no gurantee that the situations studied here are representative of current applications, but that is not adequate reason for basing our assessments on unrealistically idealistic assumptions.* [79, str. 1068]
[96] *(Table and Figure) suggest, that the data sets considered tend to have slightly heavier tails than the normal, but that a view of the world through Cauchy colored glases may be overly-pessimistic.* [79, str. 1070]
[97] **George Alfred Barnard** (1915–2002) – britský statistik. Studoval na univerzitě v Princetonu, odkud během druhé světové války odešel. V roce 1948 odešel na Imperial College London, kde se roku 1954 stal profesorem matematické statistiky. Odtud odešel na nově vytvořenou University of Essex. Až do roku 1981 ale trávil velkou část roku na University of Waterloo v Canadě. Je znám především díky svým pracím ze statistiky a řízení jakosti.
[98] **David Roxbee Cox** (*1924) – britský statistik. Matematiku vystudoval na St. John`s College v Cambridge. PhD. vzdělání získal na University of Leeds. David Cox napsal, nebo byl spoluautorem, 300 článků a knih.

z Imperial College v Londýně nebo Roberta V. Hogga[99] z University of Iowa). Jiné byly více kritické – třeba ty od D. F. Andrewse (University of Toronto, jeden z autorů Princentonské studie) nebo Churchilla Eisenharta[100] (National Bureau of Standards). Nicméně otevřela otázku povahy skutečných dat.

4.8 Small sample asymptotics

U robustních odhadů by bylo logicky žádoucí zjistit jejich alespoň přibližné rozdělení pravděpodobností. Označme příslušný odhad T_n s hustotou f_n, založený na pozorováních $X_1,\ldots,X_n$, která jsou nezávislá, se stejnou hustotou *f*. Pokud T_n a *f* nemají nějaký speciální tvar, pak lze těžko analyticky vypočítat rozdělení T_n. Alternativou je linearizovat statistiku T_n a dokázat, že tato linearizovaná statistika je ekvivalentní T_n, pokud $n \to$. Výsledné asymptotické rozdělení může být použito jako aproximace rozdělení T_n. Ovšem v případě, že vzorek, se kterým pracujeme, je příliš malý, toto asymptotické rozdělení neposkytuje příliš dobré výsledky. Navíc tato aproximace je často nepřesná na koncích rozdělení.

Obrázek 4.3. *Henry Ellis Daniels*

V letech 1966–1991 byl editorem Biometriky. Cox přispěl svými příspěvky k mnoha oblastem statistiky a aplikované pravděpodobnosti. Zabýval se např. proporcionálními modely za rizika, stochastickými procesy, analýzou binárních dat.

[99] **Robert Vincent Hogg** (*1924) – titul PhD. ze statistiky získal na University of Iowa, kde také poté zůstal na fakultě matematiky. Stal se vedoucím oddělení statistiky, když v roce 1965 toto oddělení na univerzitě vznikalo. Ve své funkci zůstal dalších 19 let. Po 51 letech na univerzitě se stal v roce 2001 emeritním profesorem. Zabýval se robustními a adaptivními odhady a neparametrickou statistikou.

[100] **Churchill Eisenhart** (1913–1994) – americký matematik, byl předsedou Statistical Engineering Laboratory (SEL) a Divize aplikované matematiky v National Bureau of Standards (NBS). Byl synem matematika Luthera Eisenharta. Do NBS se dostal z University of Wisconsin-Madison.

V roce 1954 používá Henry Ellis Daniels[101] ve svém článku *Saddlepoint Approximations in Statistics* [16] metodu sedlového bodu k odvození velice přesné aproximace rozdělení aritmetického průměru. Hlavní myšlenka je následující – hustota f_n může být zapsána jako integrál v komplexní rovině pomocí Fourierovy transformace. Protože integrand je ve tvaru $e^{nw(\zeta)}$, hlavní podíl na tomto integrálu pro velká *n* bude pocházet z okolí sedlového bodu, tj. bodu, pro který je $w'(\zeta)$ rovno nule. Hodnota integrandu bude zanedbatelná mimo blízké okolí tohoto sedlového bodu. Pomocí metody „steepest descent" pak Daniels odvozuje kompletní rozvoj pro f_n řádu n^{-}. Výhodou tohoto rozvoje je fakt, že již několik prvních členů, někdy dokonce již sám vedoucí člen, poskytuje velice přesnou aproximaci pro konce rozdělení a funguje dobře i pro velmi malé soubory dat. Daniels o metodě sedlového bodu píše i ve svých dalších článcích [14, 15].

Obdobnou problematikou se dále zabýval také Frank Hampel. Vhodně použil výraz „small sample asymptotics", který vystihuje podstatu těchto metod. V roce 1973 na pražském sympoziu[102] přednesl příspěvek *Some Small Sample Asymptotics* [33]. I na druhém pražském sympoziu v roce 1978 bylo o tomto tématu pojednáno, tentokrát Fieldem[103]. Ve sborníku z této konference je pak možné najít článek *Summary of Small Sample Size Asymptotics for Location Estimates* [21]. V roce 1982 Hampel společně s Fieldem uveřejňují v Biometrice článek *Small-Sample Asymptotic Distributions of M-Estimators of Location* [20]. Tato metoda je vlastně vylepšenou verzí sedlového bodu, se kterým přišel v roce 1954 Daniels.

Hampel ve svém příspěvku z roku 1973 odhaduje rozdělení aritmetického průměru a vybraných *M*-odhadů. Studuje dokonce vzorky o velikosti 1 nebo 2. Metoda totiž dává velice dobré výsledky i pro velmi malé vzorky. Jeho práce je zaměřena spíše na aplikace, než na ryzí matematickou teorii. Již dříve se používaly Edgeworthovy rozvoje pro distribuční funkci F_n nebo pro hustotu f_n rozdělení odhadu. Nicméně to mělo své nevýhody. Edgeworthovy rozvoje mohly vést k záporným hustotám a nedávaly příliš dobré výsledky pro konce rozdělení. Navíc rozvoje F_n, případně f_n, vedly ke složitým výrazům. Hampel tak přichází s myšlenkou udělat rozvoj výrazu f_n'/f_n. Zatímco rozvoje F_n nebo f_n vedly často k exponenciálním výrazům, první výraz rozvoje f_n'/f_n je úměrný *n* a první dva výrazy jsou

[101] **Henry Ellis Daniels** (1912–2000) – britský statistik. V letech 1974–1975 byl prezidentem Royal Statistical Society.
[102] Jednalo se o Prague Symposium on Asymptotic Statistics. Poprvé se v Praze konalo v roce 1973, následně v pětiletých intervalech.
[103] **Christopher A. Field** – působí na Dalhousie University v Halifaxu v Novém Skotsku v Kanadě. Mezi hlavní oblasti jeho výzkumu patří robustní statistika, molekulární biologie aproximace v sedlovém bodě.

lineární v *n*. Hampel se v této práci zaměřuje především na aritmetický průměr. Označme p_n hustotu aritmetického průměr prvních *n* pozorování, $X_1, X_2, \ldots$ jsou nezávislá, stejně rozdělená pozorování s hustotou *f,* nemající delší konce rozdělení jak exponenciální rozdělení, $T_n = \bar{X}_n$. Hampel tedy uvažuje rozvoj

$$-K_n(\cdot) = p_n'(\cdot)/p_n(\cdot) = -n\alpha(\cdot) - \beta(\cdot) - \gamma(\cdot)/n - \ldots$$

Empiricky bylo zjištěno, že již první dva členy, které jsou pro každé *t* lineární v *n*, poskytují velice dobrou aproximaci.

V následujícím Fieldově příspěvku z roku 1978 se objevuje zobecnění Hampelovy metody pro aproximaci hustoty aritmetického průměru. Metoda je aplikována na *M*-odhady pro jednorozměrná data a na aritmetický průměr pro vícerozměrná data. Numerické výsledky ukazují na vysokou přesnost aproximace i v případě extrémních konců rozdělení.

Nyní tedy konečně přistoupíme trochu podrobněji k samotné podstatě metody, kterou používají Hampel a Field. Od aproximačních metod, které byly používány (např. již zmíněné Edgeworthovy rozvoje), se liší v několika ohledech. Jedním z nich je fakt, že místo rozvoje vysokého řádu v jednom bodě je používán rozvoj nižšího řádu v každém bodě. Rozvoje vyššího řádu mohou být nanejvýš lokálně přesné a pro velká *n* jsou rozvoje vyššího řádu zbytečné. Dalším rozdílem je již zmiňovaný fakt, že není vytvářen rozvoj distribuční funkce, ani hustoty, ale je aproximována derivace logaritmu hustoty $-\zeta_n(\cdot) = p_n'(\cdot)/p_n(\cdot)$. Tato pomocná funkce je známá jako skórová funkce maximálně věrohodných odhadů. Normální rozdělení hraje výjimečnou úlohu ve statistice. A právě jednou z výhod tohoto postupu je fakt, že linearita *K* v *t* ukazuje právě na normální rozdělení.

Držme se tedy společné práce Fielda a Hampela z roku 1982 [20], kde shrnují své dosavadní výsledky. Hlavní část tvoří odvození formule druhého řádu pro *M*-odhady polohy s monotónní funkcí ψ jak pro p_n'/p_n, tak pro p_n. Následně jsou teoretické formule použity ve dvou situacích – pro Huberovy odhady za 5% kontaminace normálního rozdělení a pro Cauchyho rozdělení. Tyto výsledky jsou srovnány s Danielsovou metodou sedlového bodu.

Předpokládejme tedy *n* nezávislých pozorování $X_1, \ldots, X_n$ s hustotou $f(x-\theta)$, kde *T* a θ jsou definovány jako řešení rovnice $\sum_{i=1}^{n} \psi(x_i - T) = 0$. Autoři uvažují rostoucí funkci ψ, *f*, ψ po částech diferencovatelné a hustotu $m_t(\cdot) = c_t \exp(\alpha\psi(x-\cdot))f(\cdot)$, kde $c_t^{-1} = \int_{-\infty}^{\infty} \exp(\alpha\psi(x-t))f(x)\,dx$. Hustota *T* je označena $p_n(\cdot)$. Nyní ve výrazu $p_n'(\cdot)/p_n(\cdot)$ je

samostatně upravován a rozvinut pomocí Edgeworthova rozvoje jmenovatel a čitatel. Vlastní odvození a přesné podmínky (především pro existenci jednotlivých integrálů během odvozování) je možné nalézt v [20] na str. 33–37. Až poté jsou tyto dva výrazy vyděleny a po úpravě je získán výsledný výraz:

$$\frac{p'_n(x)}{p_n(x)} = \left(n - \frac{1}{2}\right) A_{2,t} + \frac{\lambda_{n,t} A_{4,t}}{2\sigma} - \frac{A_{4,t} A_{3,t}}{\sigma A_{1,t}} - \frac{A_{6,t}}{2\sigma} + \frac{A_{5,t}}{A_{1,t}} + O(n^{-1})$$

kde

$$A_{1,t} = \int_{-\infty}^{\infty} \psi(x - t - c_t) \exp\left[\alpha(x - t)\right] f(x)\,dx,$$

$$A_{2,t} = \int_{-\infty}^{\infty} c_t \exp\left[\alpha(x - t)\right] f'(x)\,dx,$$

$$A_{3,t} = \int_{-\infty}^{\infty} \psi(x - t)\,\psi(x - t - c_t) \exp\left[\alpha(x - t)\right] f(x)\,dx,$$

$$A_{4,t} = \int_{-\infty}^{\infty} \psi(x - t - c_t) \exp\left[\alpha(x - t)\right] f'(x)\,dx,$$

$$A_{5,t} = \int_{-\infty}^{\infty} \psi(x - t - c_t) \exp\left[\alpha(x - t)\right] f'(x)\,dx,$$

$$A_{6,t} = \int_{-\infty}^{\infty} \psi^2(x - t - c_t) \exp\left[\alpha(x - t)\right] f'(x)\,dx$$

a c_t, α splňují rovnosti

$$\int_{-\infty}^{\infty} c_t \exp\left[\alpha(x - t)\right] f(x)\,dx = 1,$$

$$\int_{-\infty}^{\infty} \psi(x - t - c_t) \exp\left[\alpha(x - t)\right] f(x)\,dx = 0,$$

$$\sigma^2 = \int_{-\infty}^{\infty} \psi^2(x - t - c_t) \exp\left[\alpha(x - t)\right] f(x)\,dx,$$

$$\lambda_{n,t} = \int_{-\infty}^{\infty} \psi^n(x - t - c_t) \exp\left[\alpha(x - t)\right] f(x)\,dx / \sigma^n.$$

Následně byla také počítána relativní procentuální chyba konců rozdělení jako 100(*N*–*E*)/(1–*E*), kde *N* je aproximativní a *E* přesná distribuční funkce. Pro kontaminované normální rozdělení s $\varepsilon = 5$ % vycházely relativní chyby kolem nebo pod 1 % pro $t = 0{,}5$ až po $n = 1$, pro $t = 1$ až po $n = 3$ a pro $t = 1{,}5$ až po $n = 5$. Pro kontaminované normální rozdělení byly

počítány aproximace prvního, druhého i vyšších řádů (zahrnují pouze první, první dva, více výrazů rozvoje $p'_{..}(\cdot)p_n(\cdot)$ a bylo zjištěno, že pro konce rozdělení zahrnutím druhého výrazu do aproximace bude dosaženo podstatného zlepšení. Nicméně přidání třetího výrazu již další výrazné zlepšení nepřináší. Více o small sample asymptotics např. v [24].

4.9 Ricardo Antonio Maronna

V současnosti jsou trendem robustní metody v mnohorozměrných modelech. Problémem je samozřejmě zjednodušení struktury mnohorozměrných dat nebo nahrazení mnohorozměrných modelů jednorozměrnými.

Na pole robustních odhadů přišel s mnohorozměrnými modely již v roce 1976 Ricardo Antonio Maronna[104]. Institute of Mathematical Statistics otiskuje jeho článek *Robust M-Estimators of Multivariate Location and Scatter* [62]. V tomto listu pojednává o robustním odhadu vektoru polohy $\mathbf{t}$ a kovarianční matice $\mathbf{V}$ pomocí *M*-odhadů, definovaných jako řešení soustavy rovnic:

$$n^{-1}\sum_{i=1}^{n} u_1\left[(\mathbf{x_i}-\mathbf{t})'\mathbf{V}^{-1}(\mathbf{x_i}-\mathbf{t})\right]^{\frac{1}{2}}(\mathbf{x_i}-\mathbf{t}) = \mathbf{0},$$

$$n^{-1}\sum_{i=1}^{n} u_2\left[(\mathbf{x_i}-\mathbf{t})'\mathbf{V}^{-1}(\mathbf{x_i}-\mathbf{t})\right](\mathbf{x_i}-\mathbf{t})(\mathbf{x_i}-\mathbf{t})' = \mathbf{V},$$

kde u_1, u_2 jsou funkce, splňující množinu níže uvedených obecných předpokladů. Pro usnadnění zápisu a možnost srovnání s jednorozměrným případem je zde definováno pro $s \geq 0$: $\psi_i(s) = su_i(s)$, $i = 1,2$. Požadavky na funkce u_1, u_2 jsou následující:

(i) u_1, u_2 jsou nezáporné, nerostoucí a spojité pro $s \geq 0$.

(ii) ψ_1, ψ_2 jsou ohraničené. Nechť $K_i = \sup_{s\geq 0}\psi_i(s)$.

(iii) ψ_1 je neklesající, v intervalu, kde $\psi_1 < K_2$, je rostoucí.

(iv) Existuje s_0 tak, že $\psi_2(s_0^2) > p$ a $u_1(s) > 0$ pro $s \leq s_0$ (a proto $K_2 > p$).

Maronnova práce je velkým krokem vpřed vzhledem k odhadu kovarianční matice. Výhodou zde uvažovaného přístupu k *M*-odhadům je fakt, že dává afinně invariantní odhady, které jsou automaticky pozitivně definitní. V třetí a čtvrté části spisu je dokázána existence a jednoznačnost řešení navržené soustavy rovnic. Maronna pro názornost používá dva konkrétní případy. Prvním z nich je vícerozměrný Huberův *M*-odhad, kde funkce $\psi(s,k)$ je definována

[104] **Ricardo Antonio Maronna** – doktorské studium ukončil na univerzitě v Buenos Aires v roce 1975.

jako $\psi(t,k) = \max(-k, \min(k,t))$. Druhým příkladem je maximálně věrohodný odhad pro Studentovo rozdělení. V následujících částech je dokázána konzistence a asymptotická normalita. V sedmé části je pojednáno o influenční funkci a bodu selhání. A právě nevýhodou těchto afinně invariantních *M*-odhadů je fakt, že mají nízký bod selhání. Ten je menší nebo roven $1/m$, kde m je řád matice. V osmé části je vybráno devět konkrétních odhadů – jedná se o pět Huberových *M*-odhadů a čtyři maximálně věrohodné odhady pro Studentovo rozdělení s různými stupni volnosti. Pro tyto odhady jsou počítány hodnoty některých měr robustnosti společně s asymptotickými rozptyly těchto odhadů. Na závěr své práce Maronna ukazuje vylepšení procedury pro numerický výpočet odhadů a ukazuje výsledky chování odhadů pro vzorky velikosti deset a dvacet při pokusu, kdy některé parametry rozdělení odhadů byly počítány metodou Monte Carlo.

4.10 Současnost robustní statistiky a její přínosy

V současnosti se robustní statistika zabývá především regresními modely a jejich dalšími rozšířeními, jako jsou mnohorozměrné modely, heteroskedastické proměnné, nelineární regrese. Uvažují se také výpočetní aspekty. Co lze, se rozšiřuje na časové řady, např. na autogresní modely, včetně mnohorozměrných. Vedle regrese se v současnosti uvažuje také predikce, klasifikační modely, diskriminace a kalibrace.

V dnešní době je robustnost součástí výuky statistiky na mnoha univerzitách a robustní metody jsou používány i v praxi aplikovanými statistiky. Nicméně toto přijetí robustních metod nebylo automatické. Ačkoliv česká matematická statistika byla vždy na vysoké úrovní, ani zde na začátku sedmdesátých let dvacátého století nebyla robustnost přijata kladně. Avšak již v roce 1980 zorganizovala Matematicko-fyzikální fakulta v Praze první letní školu Robust o netradičních metodách matematické statistiky v Načetíně u Chomutova. Od té doby se Robust koná každé dva roky, střídavě v zimě a v létě. Poslední letní škola byla od 8. 9. 2008 v Roháčích na Slovensku.

Práci bych ráda zakončila několika poznatky z Ronchettiho článku *The Historical Development of Robust Statistics* z roku 2006, kde se zabýval otázkou, jak mohou hlavní myšlenky a nástroje robustní statistiky přispět k obecnému rozvoji moderní statistiky. A zde jsou některé jeho výsledky (volně přeloženo a zkráceno podle [74]):

(i) Model je pouze aproximací reality

Tato myšlenka je samozřejmě zcela běžná ve všech vědách. Nicméně robustní statistika ukazuje, jak se mohou procedury, které byly při určitých podmínkách optimální, změnit,

když se nepatrně od těchto podmínek odchýlíme. To otvírá dveře hledání lepších alternativ a pro používání více metod pro analýzu dat.

(ii) Mnohonásobné analýzy a řešení analýzy dat

Robustní statistika přispěla k rozvoji myšlenky, že mnohonásobné nástroje jsou nezbytné pro analýzu reálných dat a že reálné problémy mohou mít více řešení.

(iii) Statistické funkcionály; Gateauxova a Fréchetova diferencovatelnost

Hampelův přístup k robustnosti je v řeči funkcionální analýzy. Zejména influenční funkce (Gateauxova derivace funkcionálu) se stala nejdůležitějším heuristickým nástrojem analýzy stability statistických procedur a vývoje nových robustních procedur. Statistické funkcionály hrály důležitou roli později v rozvoji neparametrických metod.

(iv) *M*-odhady

Huberovy M-odhady představují velmi flexibilní a obecnou třídu odhadů, které hrály důležitou roli v rozvoji robustní statistiky a v konstruování robustních procedur. Nicméně tato myšlenka je mnohem obecnější a je důležitým stavebním blokem v mnohem různých oborech, např. v ekonometrii nebo biostatistice.

(v) Bod selhání

Bod selhání je mírou globální stability pro statistický funkcionál a jako takový je typickou mírou robustnosti. Požadavek na vysoký bod selhání odhadů pomohl i vývoji např. obecných výpočtových technik.

Literatura

[1] Andrews D. F., Bickel P. J., Hampel F. R., Huber P. J., Rogers W. H., Tukey, J. W.: *Robust Estimates of Location. Survey and Advances*. Princeton University Press, Princeton, 1972.

[2] Archibald R. C.: *History of Mathematics After the Sixteenth Century*. The American Mathematical Monthly 56(1949), 35–56.

[3] Bassett G., Koenker R.: *On Boscovich`s Estimator*. The Annals of Statistics 13(1985), 1625–1628.

[4] Bečvář J., Fuchs E.: *Matematika v 19. století*. Edice Dějiny matematiky, Prometheus, Praha, 1996.

[5] Belzová L.: *M-odhady a jejich vlastnosti*. Diplomová práce, Katedra pravděpodobnosti a matematické statistiky, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze, 2002.

[6] Bickel P. J.: *Another Look at Robustness: A Review of Reviews and Some New Developments*. Scandinavian Journal of Statistics 3(1976), 145–168.

[7] Boscovich R. J., Maire C.: *De Litteraria Expeditione per Pontificiam ditionem ad dimetiendas duas Meridiani gradus*. Palladis, Rome, 1755.

[8] Boscovich R. J., Maire C.: *Voyage astronomique et géographique dans l`etat de l`eglise*. Tilliard, Paris, 1770.

[9] Box G. E. P.: *Non-Normality and Tests on Variances*. Biometrika 40(1953), 318–335.

[10] Box G. E. P., Anderson S. L.: *Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption*. Journal of the Royal Statistics Society 17(1955), 1–34.

[11] Celmins A.: *The Method of Gauss in 1799*. Statistical Science 13(1998), 123–135.

[12] Cohen M. L., Gastwirth J. L.: *Small Sample Behavior of Some Robust Linear Estimators of Location*. Journal of the American Statistical Association 65(1970), 946–973.

[13] Daniell P. J.: *Observations Weighted According to Order*. American Journal of Mathematics 42(1920), 222–36.

[14] Daniels H. E.: *Exact Saddlepoint Approximations*. Biometrika 67(1980), 59–63.

[15] Daniels H. E.: *Saddlepoint Approximations for Estimating Equations*. Biometrika 70(1983), 89–96.

[16] Daniels H. E.: *Saddlepoint Approximations in Statistics*. The Annals of Mathematical Statistics. Institute of Mathematical Statistics 25(1954), 631–650.

[17] David H. A.: *Early Sample Measures of Variability*. Statistical Science 13(1998), 368–377.

[18] David F. N.: *Games, Gods and Gambling, A History of Probability and Statistical Ideas*. Dover Publications, New York, 1998.

[19] Euler L.: *Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter*. V *Leonhardi Euleri*, Opera Omnia 25(1960), 130–141.

[20] Field Ch. A., Hampel F. R.: *Small-Sample Asymptotic Distributions of M-Estimators of Location*. Biometrika 69(1982), 29–46.

[21] Field Ch. A.: *Summary of Small Sample Size Asymptotics for Location Estimates*. Proceedings of the Second Prague Symposium on Asymptotic Statistics 1978, ed. P. Mandl a M. Hušková, 173–179.

[22] Finkel B. F.: *Biography: Karl Frederich Gauss*. The American Mathematical Monthly 8(1901), 25–31.

[23] Fischer H.: *Dirichlet's Contributions to Mathematical Probability Theory*. Historia Mathematica 21(1994), 39–63.

[24] Flinger M. A.: *Small Sample Asymptotics*. Journal of Educational Statistics 13(1988), 53–61.

[25] Gowing R.: *Halley, Cotes, and the Nautical Meridian*. Historia Mathematica 22(1995), 19–32.

[26] Hacking I.: *The Taming of Chance*. Cambridge University Press, New York, 1990.

[27] Hájek J.: *Asymptotically Most Powerful Rank-Order Tests*. The Annals of Mathematical Statistics 33(1962), 1124–1147.

[28] Hald A.: *A History of Probability and Statistics and Their Apllications before 1750*. John Wiley & Sons, Inc., New York, 1990.

[29] Hald A.: *A History of Mathematical Statistics from 1750 to 1930*. John Wiley & Sons, Inc., New York, 1998.

[30] Hampel F. R.: *A General Qualitative Definition of Robustness*. The Annals of Mathematical Statistics 42(1971), 1887–1896.

[31] Hampel F. R.: *Optimally bounding the gross-error-sensitivity and the influence of position in factor space*. Proceedings of the American Statistical Association Statistical Computing Section, American Statistical Association, Washington, D. C., 1978, 59–64.

[32] Hampel F. R., Rousseuw P. J., Ronchetti E., Stahel W.: *Robust Statistical Procedures. The Approach Based on Influence Functions*, New York, Wiley, 1986.

[33] Hampel F. R.: *Some Small Sample Asymptotics*. Proceedings of the Prague Symposium on Asymptotic Statistics 1973, ed. J. Hájek, 109–126.

[34] Hampel F. R, Rousseeuw P. J., Ronchetti E..: *The Change-of-Variance Curve and Optimal Redescending M-Estimators*. Journal of the America Statistical Association 76(1981), 643–648.

[35] Hampel F. R.: *The Influence Curve and Its Role in Robust Estimation*. Journal of the American Statistical Association 69(1974), 383–393.

[36] Hodges J. L., Lehmann E. L.: *Estimatest of Location Based on Rank Tests*. The Annals of Mathematical Statistics 34(1963), 598–611.

[37] Hogg R. V.: *Adaptive robust procedures: a partial review and some suggestions for future applications and theory*. Journal Amer. Statist. Assoc. 69(1974), 909–922.

[38] Huber P. J.: *A Robust Version of the Probability Ratio Test*. The Annals of Mathematical Statistics 36(1965), 1753–1758.

[39] Huber P. J.: *Current Issues in Robust Statistics*. Lisboa, 1983

[40] Huber P. J.: *Recent trends in robustness*. IEEE International Symposium on Information Theory, France, 1982.

[41] Huber P. J.: *Robust Estimation of a Location Parameter*. Annals of Mathematical Statistics 35(1964), 73–101.

[42] Huber P. J.: *Robust Statistics*. New York, Wiley, 1981.

[43] Huber P.: *The 1972 Wald Lecture Robust statistics: A Review.* The Annals of Mathematical Statistics 43(1972), 1041–1067.

[44] Chernoff H., Gastwirth J. L., Johns M. V.: *Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation*. The Annals of Mathematical Statistics 38(1967), 52–72.

[45] Chernoff H., Savage I. R.: *Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics*. The Annals of Mathematical Statistics 29(1958), 972–994.

[46] Jaeckel L. A.: *Robust Estimates of Location: Symmetry and Asymmetric Contamination.* The Annals of Mathematical Statistics 42(1971), 1020–1034.

[47] Jurečková J.: *Asymptotic Linearity of a Rank Statistics in Regression Parameter*. The Annals of Mathematical Statistics 40(1969), 1889–1900.

[48] Jurečková J., Picek J.: *Robust Statistical Methods with R*. Chapman & Hall/CRC, Boca Raton, 2006.

[49] Jurečková J.: *Robustní statistické metody*. Karolinum, Praha, 2001.

[50] Kleczek J.: *Velká encyklopedie vesmíru*. Academia, Praha, 2002.

[51] Koenker R., Bassett G.: *On Boscovich's Estimator*. The Annals of Statistics 13(1985), 1625–1628.

[52] Koshar R.: *Foucault and Social History: Comments on "Combined Underdevelopment"*. The American Historical Review 98(1993), 354–363.

[53] Kreager P.: *Histories of Demography: A Review Article*. Population Studies 47(1993), 519–539.

[54] Laplace P. S.: *Théorie Analytique des Probabilités*. Paris, 1812.

[55] Laplace P. S.: *Théorie de Jupiter et de Saturne*. Mém. Acad. Sci. Paris année (1787), 33–160.

[56] Legendre A. M.*: Nouvelles méthodes pour la détermination des orbites des cométes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux cométes de 1805*. F. Didod, Paříž, 1805.

[57] Lehmann E. L.: *Asymptotically Nonparametric Inference: An Alternative Approach to Linear Models*. The Annals of Mathematical Statistics 34(1963), 1494–1506.

[58] Lehmann E. L.: *Nonparametric Confidence Intervals for a Shift Parameter.* The Annals of Mathematical Statistics 34(1963), 1507–1512.

[59] Lehmann E. L.: *Robust Estimation in Analysis of Variance*. The Annals of Mathematical Statistics 34(1963), 957–966.

[60] Lloyd E. H.: *Least-Squares Estimation of Location and Scale Parameters Using Order Statistics*. Biometrika 39(1952), 88–95.

[61] Mačák K.: *Počátky počtu pravděpodobnosti*. Edice Dějiny matematiky, 9. svazek, Prometheus, Praha, 1997.

[62] Maronna R. A.: *Robust M-Estimators of Multivariate Location and Scatter*. The Annals of Statistics 4(1976), 51–67.

[63] Mayer T.: *Abhandlung über die Umwälzung des Mondes um seine Axe*. Kosmographische Nachrichten u. Sammlungen (1750), 52–183.

[64] Merriman M.: *On the History of the Method of Least Squares*. The Analyst 4(1877), 33–36.

[65] Mises R. v.: *On the Asymptotic Distribution of Differentiable Statistical Functions*. The Annals of Mathematical Statistics 18(1947), 309–348.

[66] Mosteller F.: *On Some Useful „Inefficient" Statitics*. The Annals of Mathematical Statistics 17(1946), 377–408.

[67] Newcomb S.: *A Generalized Theory of the Combination of Observations so as to Obtain the Best Result*. American Journal of Mathematics 8(1886), 343–366.

[68] Pearson E. S.: *Studies in the History of Probability and Statistics XVII: Some Reflections on Continuity in the Development of Mathematical Statistics, 1885-1920*. Biometrika 54(1967), 341–355.

[69] Pearson E.G.: *The Analysis of Variances in Cases of Non-Normal Variation*. Biometrika 23(1931), 114–133.

[70] Peirce B.: *Criterion for the Rejection of Doubtful Observations*. The Astronomical Journal 2(1852), 161–163.

[71] Plackett R. L.: *Studies in the History of Probability and Statitics. XXIX: Discovery of the Method of Least Squares*. Biometrika 59(1972), 239–251.

[72] Poincaré H.: *Calcul des Probabilités*. Gauthier-Villars, Paříž, 1912.

[73] Portnoy S., He X.: *A Robust Journey in the New Millennium*. Journal of the American Statistical Association 95(2000), 1331–1335.

[74] Ronchetti E.: *The Historical Development of Robust Statistics*. ICOTS 7(2006).

[75] Ruppert D., Carroll R. J.: *Trimmed Least Squares Estimation in the Linear Model*. Journal of the American Statitical Association 75(1980), 828–838.

[76] Short J.: *Second Paper Concerning the Paralax of the Sun Determined from the Observations of the Late Transit of Venus; in Which This Subject Is Treated of More at Length, and the Quantity of the Paralax More Fully Ascertained*. Philosophical Transactions of the Royal Society of London 53(1763), 300–345.

[77] Siegel A. F.: *Robust Regression Using Repeated Medians*. Biometrika 69(1982), 242–244.

[78] Stigler S.: *Boscovich, Simpson, and a 1760 Manuscript Note on Fitting a Linear Relation*. Biometrika 71(1984), 615–620.

[79] Stigler S.: *Do Robust Estimators Work with Real Data?* Annals of Statistics 5(1977), 1055–1098.

[80] Stigler S.: *Francis Ysidro Edgeworth, Statistician*. Journal of the Royal Statistical Society 141(1978), 287–322.

[81] Stigler S.: *R. H. Smith, A. Victorian Interested in Robustness*. Biometrika 67(1980), 217–221.

[82] Stigler S.: *Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885–1920*. Journal of the American Statistical Association 68(1973), 872–879.

[83] Stigler S.: *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, Massachusetts, 1999.

[84] Stigler S.: *Studies in the History of Probability and Statistics XXXVIII, R. H. Smith, a Victorian interested in robustness.* Biometrika 67(1980), 217–221.

[85] Stigler S.: *The History of Statistics: The Measurement of Uncertainty before 190*0. The Belknap Press of the Harvard University Press, Cambridge, Massachusetts, 1986.

[86] Tukey J. W.: *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachussets, 1977.

[87] Tukey J. W.: *The Future of Data Analysis*. The Annals of Mathematical Statistics 33(1962), 1–67.

[88] Zwet W. R. v.: *Van de Hulst on robust statistics. A historical note.* Statistica Neerlandica 39(1985), 81–95.