Popisná statistika 2

Zdeněk Mikulášek, Ústav teoretické fyziky a astrofyziky, PřF Masarykovy Univerzity, Brno

Dříve než přikročíme k analýze časových řad, vyplatí se provést důkladný rozbor naměřených dat, abyste pak mohli zvolit optimální způsob dalšího zpracování. K tomu slouží nástroje popisné statistiky nastíněné v . I zde budeme předpokládat, že máme k dispozici soubor n naměřených hodnot vytvářející datový soubor $D = \{x_i\}$. Vhodné je též přisoudit každému měření nebo skupině měření přisoudit jistou váhu, která by měla souviset s nejistotou, s níž jednotlivá měření zjišťujeme. Je-li odhad nejistoty i-tého měření δx_i, pak bychom takovému měření měli přisoudit váhu w_i, kde $w_i = n\,(\delta x_i)^{-2}/\sum_{i=1}^{n}(\delta x_i)^{-2}$. Pokud se domníváme, že nejistoty všech měření jsou stejné, pak klademe váhu každého z měření rovnu 1 ($w_i = 1$), takže součet vah je roven n.

Jednotkovou váhu používáme i tehdy, je-li rozdíl v kvalitě jednotlivých měření malý, nebo je-li očekávaná vnitřní nejistota jednotlivých měření viditelně menší než jejich celkový rozptyl v rámci souboru. Naopak použití vah je deklarováno, zejména při transformaci měřených veličin nějakou nelineární funkcí ($\log x, 1/x$) nebo při robustní regresi.

V dalším zavádíme následující konvenci:

$$\overline{x^k} = \sum_{i=1}^{n} x_i^k\, w_i/n. \tag{1}$$

1 Nástroje pro normální rozdělení. Průměr, standardní odchylka

Pro prvotní popis pozorovaných dat je dobré uvést dvě charakteristiky - nějakou hodnotu, kolem níž se pozorovaná data kupí - nějaký střed datového souboru, a pak veličinu, která popisuje charakteristickou vzdálenost pozorovaných dat od tohoto středu.

Z hlediska nejčastějšího z nástrojů zpracování datových souborů - *metody nejmenších čtverců*, je přirozenou mírou popisující střed studovaného datového souboru veličina a nazývaná též *aritmetický průměr*, respektive *průměr*, obecně *váhovaný průměr*. Pro tuto veličinu platí, že suma váhovaných čtverců odchylek jednotlivých měření od centra a, $S(a)$, je minimální:

$$S(a) = \sum_{i=1}^{n}(x_i - a)^2\, w_i = n\,\overline{(x-a)^2} = n\left[\overline{x^2} - 2\,a\,\overline{x} + a^2\right]; \quad \frac{\partial S(a)}{\partial a} = 0 \Rightarrow a = \overline{x}. \tag{2}$$

Ze vztahu (2) plyne, že tímto aritmetickým průměrem a je váhovaná střední hodnota $\overline{x}$ definovaná ve smyslu konvence (1) pro $k = 1$. Pro popis míry rozptýlení od aritmetického průměru a zavedeme *váhovanou standardní odchylku* s, jejíž kvadrát odhadujeme vztahem:

$$s^2 = \frac{S(\bar{x})}{n-1} = \frac{n\,\overline{(x-\bar{x})^2}}{n-1} = \frac{n\,(\overline{x^2} - 2\,\bar{x}\,\bar{x} + \bar{x}^2)}{n-1} = \frac{n\,(\overline{x^2} - \bar{x}^2)}{n-1}. \tag{3}$$

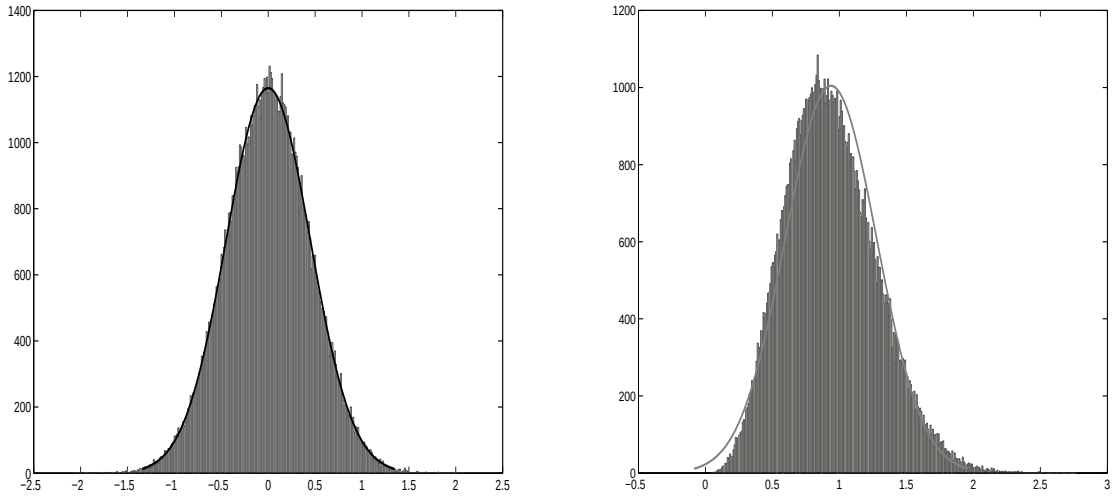

Fig. 1. Simulace výsledků odhadů centra v aritmetickém průměru (vlevo) a standardní odchylky (vpravo) pro normální rozdělení s centrem v 0 a standardní odchylkou 1 při výběru pouhých 5 bodů. Tento výběr byl ovšem opakován 100 000krát. Zatímco odhad průměru se chová tak, jak bychom čekali, v případě standardní odchylky vidíme, že rozdělovací křivka se dosti liší od očekávaného normálního rozdělení. Aritmetický střed odhadu je systematicky menší. Pokud byste ze standardní odchylky přešli na její čtverec, pak sice obdržíte lepší shodu průměru rozdělení s běžně používaným vztahem, ale rozdělovací funkce bude ještě daleko odchylnější od normálního rozložení. Závěr: Běžný odhad velikosti standardní odchylky je nutno pro malý počet měření brát s velkou rezervou.

Při pokusech se simulovanými daty se ukázalo, že pokud chceme pro odhad standardní odchylky s užít prostou odmocninu střední odhadu podle rovnice (3) pro soubory s "malým počtem měření" n, dostáváme systematicky menší menší hodnoty, než bychom měli. Zavádíme proto tzv. modifikovanou standardní odchylku s_{mod} s upraveným jmenovatelem

$$s_{\mathrm{mod}} = \sqrt{\frac{n\,\overline{(x-\bar{x})^2}}{n-1.46}} = \sqrt{\frac{n\,(\overline{x^2}-\bar{x}^2)}{n-1.46}}, \tag{4}$$

který i pro $4 < n < 15$ poskytuje patřičné výsledky. Příčinu této diskrepance, na kterou učebnice neupozorňují, je skutečnost, že rozdělovací funkce pro standardní odchylku se pro malá malá n citelně odchylují od normálního - názorně je to patrno na Fig. 1.

Pokud je rozptyl pozorování určen zejména náhodnými ději (statistika fotonů, atmosférická scintilace atp.), je dáno rozdělení odchylek kolem centra symetrickou *normální rozdělovací funkcí* (Gaussovou). Funkce hustoty pravděpodobnosti $f(x)$, normovaná na 1 a je popsána dvojicí parametrů - středem rozdělení μ a disperzí σ:

$$f(x) = \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right]. \tag{5}$$

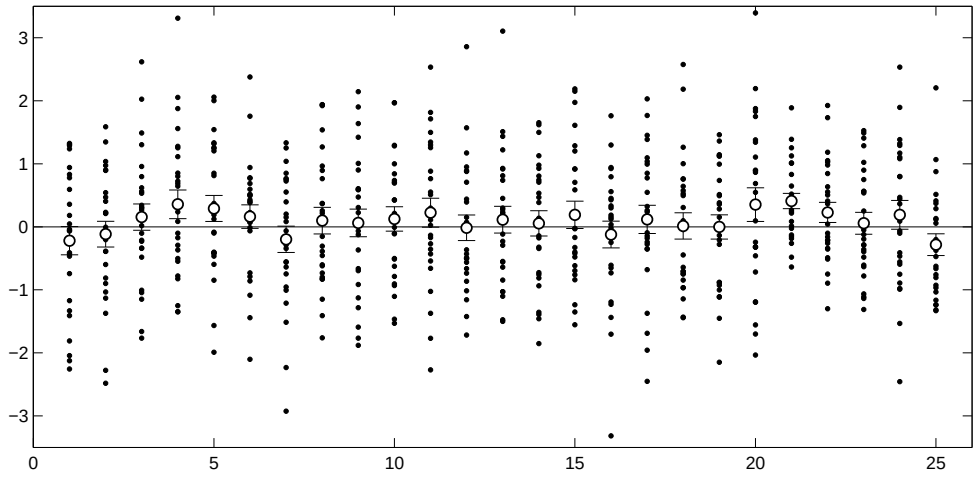

Fig. 2. Simulace výsledků 26 měření pro normální rozdělení s centrem v 0 a standardní odchylkou 1. Jednotlivá měření jednotlivých sad jsou znázorněna nad sebou plnými kotoučky, průměr s jeho nejistotou je naznačen větším prázdným kroužkem a chybovou úsečkou. Povšimněte si jak odlišné může být rozložení těchto bodů v jednotlivých sadách, rovněž tak, že body s odchylkou 3 σ jsou zcela běžné - v tomto případě tedy nejde o odlehlé body. Prostudujte i obr. 3, který je dalším zpracování této simulace.

Abychom tyto parametry určili, museli bychom mít k dispozici nekonečné množství pozorování. Pokud máme k dispozici jen n pozorování, může učinit jen odhad zmíněných veličin a odhadnout jejich neurčitost. $\delta\mu$ a $\delta\sigma$.

$$\mu \cong \bar{x}; \quad \delta\mu = \frac{\sigma}{\sqrt{n}}, \quad \sigma \cong s; \quad \delta\sigma = \frac{\sigma}{\sqrt{2\,n}}. \tag{6}$$

To, proč zde hovoříme jen o odhadech příslušných veličin, dostatečně ilustrují obrázky 2 a 3 pořízené na základě počítačových simulací. Znovu ovšem uvádíme, že výše uvedené vztahy fungují zcela správně jen tehdy, je-li reálná rozdělovací funkce blízká normální. Metody, jak si to ověřit, jsou uvedeny v následující kapitole.
Poznámka: Relativní přesnost určení rozptylu a chyby průměru $\rho = 1/\sqrt{2\,n}$ primárně závisí na počtu měření n, a to tak, že 10% činí pro 50 měření, 3% pro 560 a pro 1% již 5000 měření. Těmto skutečnostem byste měli podřídit počet míst a způsob zaokrouhlování (v těchto případech se přimlouvám zaokrouhlovat vždy spíše nahoru).

2 Odchylky od normálního rozdělení

V astronomické praxi se ovšem často setkáváme s tím, že rozdělovací funkce naměřených veličin se více či méně odlišuje od ideálního normálního rozdělení. Projevuje se to tím, že rozdělovací funkce jeví asymetrii (tedy nenulovou šikmost) nebo odlišné proporce než Gaussova křivka (odlišná špičatost). Příčin odchylek od normálního rozdělení je celá řada,

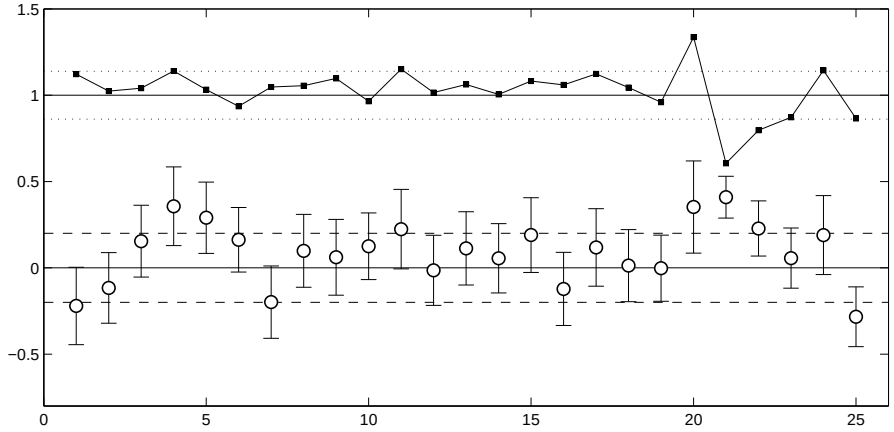

Fig. 3. Simulace výsledků měření pro normální rozdělení s centrem v 0 a standardní odchylkou 1. Každý z 25 výsledků byl zkonstruován z 26 individuálních měření. Je patrné jak kolísání polohy aritmetického průměru kolem nuly, tak i kolísání hodnoty naměřené směrodatné odchylky.

velmi často je to výskyt *odlehlých bodů* (angl. outliers) zvyšujících šikmost nebo reálná proměnnost zkoumaného zdroje, případně přístrojové efekty nejrůznějšího charakteru.

2.1. Šikmost a špičatost

Asymetrii rozdělovací funkce vyjadřuje jednoduchá bezrozměrná veličina nazývaná *šikmost* (angl. skewness). Označuje se nejčastěji jako γ_1. Nulová šikmost značí, že hodnoty náhodné veličiny jsou rovnoměrně rozděleny vlevo a vpravo od střední hodnoty. Kladná šikmost značí, že vpravo od průměru se vyskytují odlehlejší hodnoty nežli vlevo (rozdělení má tzv. pravý ocas) a většina hodnot se nachází blízko vlevo od průměru. U záporné šikmosti je tomu naopak.

Symetrická rozdělení, včetně normálního rozdělení, mají šikmost nula. Pro rozdělení s kladnou šikmostí obvykle platí, že jeho modus (nejčastěji se vyskytující hodnota) je menší nežli medián a ten je menší nežli střední hodnota. Pro zápornou šikmost je tomu opět naopak.

$$\gamma_1 = \frac{\overline{(x-\bar{x})^3}}{s^3}; \qquad \delta\gamma_1 = \frac{2.4}{\sqrt{n}}, \tag{7}$$

kde n je počet hodnot a $\delta\gamma_1$ je odhad neurčitosti pro případ normálního rozdělení. Tento odhad je důležitý pro to, abyste mohli kvalifikovaně rozhodnout, zda příslušná rozdělovací funkce je či není asymetrická. U mnohých funkcí může být zjevná asymetrie jen dílem náhody. Odhad neurčitosti šikmosti je například u stovky bodů zatížen střední neurčitostí 0,24, takže jisti byste si mohli být jen pro šikmosti, větší než 0,7 a menší -0,7. Tento příklad jen ukazuje na to, že šikmost je dobrým instrumentem jen v případě, že máte co do činění se soubory mnoha tisíci měření.

Koeficient špičatosti (excesu) (angl. excess kurtosis) je rovněž bezrozměrné číslo, které charakterizuje rozdělení náhodné veličiny tím, že jej porovnává s normálním rozdělením pravděpodobnosti. Koeficient špičatosti se obvykle označuje γ_2. Normální rozdělení má

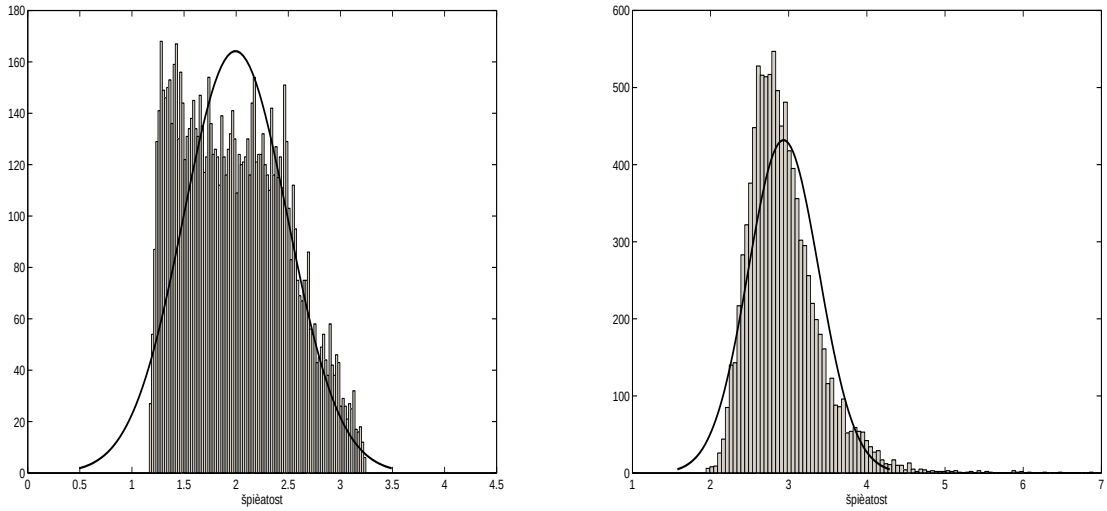

Fig. 4. Simulace výsledků odhadů špičatosti (vlevo) (vpravo) pro normální rozdělení s centrem v 0 a standardní odchylkou 1 při výběru pouhých 5 bodů (vlevo) 100 bodů (vpravo). Tento výběr byl ovšem opakován 100000krát. Je zřejmé, že ani při stovce bodů není rozdělovací křivka normální, což je dokazuje, že bychom měli být při odhadvání špičatosti obezřetní, nebo lépe, neměli bychom tuto funkci používat vůbec.

špičatost nula. Kladná špičatost značí, že většina hodnot náhodné veličiny leží blízko její střední hodnoty a hlavní vliv na rozptyl mají málo pravděpodobné odlehlé hodnoty. Křivka hustoty je špičatější, nežli u normálního rozdělení. Záporná špičatost značí, že rozdělení je rovnoměrnější a jeho křivka hustoty je plošší než-li u normálního rozdělení.

$$\gamma_2 = \frac{n+1}{n-1}\frac{\overline{(x-\bar{x})^4}}{s^4} - 3; \qquad \delta\gamma_2 = \frac{4.65}{\sqrt{n}}, \tag{8}$$

kde $\delta\gamma_2$ je odhad neurčitosti pro případ normálního rozdělení. Tento odhad je důležitý pro to, abyste mohli kvalifikovaně rozhodnout, zda se příslušná rozdělovací funkce odchyluje či neodchyluje od normálního rozdělení. U mnohých funkcí může být zjevná odchylnost jen výsledkem souhry náhod. Odhad neurčitosti koeficientu špičatosti je například u stovky bodů zatížen střední neurčitostí 0,47, takže jisti byste si mohli být jen pro rozdělovací funkce s excesem špičatosti, větší než 1.2 a menší -1.2. Tento příklad dokazuje, že pro soubory s menším počtem měření bychom tento, jinak velmi užitečný instrument měli používat s jistou obezřetností.

Budiž poznamenáno, že parametr excesu špičatosti γ_2 v tvaru (8) nejspíš nikde nenajdete, protože jde o tvar modifikovaný, který platí v rozsahu od $n > 4$. Běžně uváděný vztah pro exces špičatosti dává systematicky nižší hodnoty než by měl a platit začne až pro soubory od několika set prvků výše! Modifikace spočívá ve vložení multiplikativního členu s n.

2.2. Odlehlé body a jejich eliminace

Astrofyzikální data se dosti často musí potýkat s fenoménem odlehlých bodů. Vznikají nejčastěji v důsledku nedostatků a nestabilit v měřící aparatuře, hrubých chyb a dalších příčin nesouvisejících se zkoumaným objektem. Odlehlé body mají jiný, obecně větší, rozptyl než reálná nepoškozená měření a mění koeficient excesu špičatosti na záporný. Jejich vliv na statistiku vedenou běžnými nástroji, jako je průměr a standardní odchylka, bývá zničující.

Astrofyzikální data ovšem trpívají i opačným neduhem, kdy jsou v rámci neodborné závěrečné "kosmetizace"datových souborů nenávratně odstraněna i některá měření, která mají tu smůlu, že se zdají být příliš odchylena od centra. Řada i zkušených astronomů se domnívá, že všechny body vzdálené od centra více než $3\,\sigma$ jsou automaticky odlehlé body a lze je tedy beztrestně vymazat. Bohužel, tak jednoduché to není, obrázek 2 sestrojený výlučně z dat s normálním rozdělením, obsahuje takových nepravých "odlehlých"bodů několik. V průměru sice představují o něco méně než 0.3% (náhodně jich ale může být i několikrát víc), jejich odmazání ale má za následek neoprávněné snížení standardní odchylky v průměru o 1.5%.

K eliminaci vlivu odlehlých bodů bylo vypracováno několik postupů. Nejtriviálnější je ten, že si např. sestrojíte histogram naměřených hodnot a identifikujete body viditelně se odchylující od normálního rozdělení. Tyto body buď opravíte, je-li zřejmé, jak ona odlehlost povstala (nesprávný řád, chybějící číslice, nesprávný formát atp.) nebo je jednoduše vymažeme ze souboru. Tento postup byste ale neměli používat pro body vzdálené od centra jen o $3.5\,\sigma$, protože ty bývají v souborech dat docela běžné (viz. Obr. 2). Jinou možností je tzv. *robustní regrese*, což je nejčastěji iterativní metoda přisuzující bodům vzdáleným od centra menší váhu než bodům v blízkosti centra - více v podkapitole 2.4. Další možnost představuje zpracování nástroji, které nejsou tak citlivé na přítomnost odlehlých bodů jako jsou medián a veličina mad.

2.3. Medián, střední velikost odchylky

Medián (označován med(x) nebo $\tilde{x}$) je hodnota, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny. Ve statistice patří mezi míry centrální tendence. Platí, že nejméně 50% hodnot je menších nebo rovných a nejméně 50% hodnot je větších nebo rovných mediánu. Pro nalezení mediánu daného souboru stačí hodnoty seřadit podle velikosti a vzít hodnotu, která se nalézá uprostřed seznamu. Pokud má soubor sudý počet prvků, obvykle se za medián označuje aritmetický průměr hodnot na místech $n/2$ a $n/2+1$.

To ovšem platí pro případ, kdy jsou si váhy jednotlivých měření rovny, v opačném případě je nalezení váhovaného mediánu složitější. Postup má tyto kroky:

- Seřadíme všechny hodnoty i s jejich váhami x podle velikosti, takže $x_1 < x_2 ... < x_k < ... x_n$.
- Každému z bodů x_k přiřadíme funkční hodnotu $W_k = \frac{1}{n}\left(\sum_{i=1}^{k-1} w_i + \frac{1}{2}w_k\right)$.
- Nyní hledám po sobě následující dvojici, pro niž by platilo $W_j < 0.5 < W_{j+1}$.
- Hodnota medianw$(x, w) = \tilde{x} = [(W_{j+1} - 0.5) * x_j + (0.5 - W_j) * x_{j+1}]/(W_{j+1} - W_j)$ je pak oním hledaným váhovaným mediánem.

Předpis občas nevybere hodnotu mediánu jednoznačně, ale to většinou nevadí.

Robustní třídou měr rozptýlení je tzv. *váhovaná střední (absolutní) odchylka* (weighted mean (absolute) deviation - wmd) počítaná obecně vůči zvolenému centru a:

$$\mathrm{md} = \overline{|x-a|}; \qquad \mathrm{wmd} = \frac{1}{n}\sum_{i=1}^{n} |x_i - a|\, w_i = \overline{|X-a|}; \tag{9}$$

Počítání střední odchylky se obvykle vztahuje k váženému aritmetickému průměru, tedy $a = \bar{x}$. Lze se ovšem setkat i s jinou (dle mého soudu odůvodněnější) variantou, kdy centrem je vážený medián $a = \tilde{x}$. V tomto případě bude mít vážená suma absolutních hodnot odchylek svou minimální hodnotu.

Ještě robustnější vlastnosti má *váhovaný medián absolutní odchylky* (weighted median absolute deviation - wmad) centrovaný tentokrát vždy k mediánu:

$$\mathrm{mad} = \mathrm{median}(|x-\tilde{x}|); \qquad \mathrm{wmad} = \mathrm{medianw}(|x-\tilde{x}|); \tag{10}$$

V případě, že aplikujeme mead nebo mad na soubor dat s normálním rozdělením, pak pro disperzi σ dostanete:

$$\sigma(x) \cong 1.482\sqrt{\frac{n}{n-1.61}}\,\mathrm{mad}(x); \qquad \sigma(x) \cong 1.253\sqrt{\frac{n}{n-1}}\,\mathrm{md}(x); \tag{11}$$

V případě, že máte data s množstvím odlehlých bodů a potřebujete dobrý odhad disperze, pak můžete tyto vztahy s výhodou použít. Je třeba ještě odhadnout očekávanou relativní nepřesnost určení veličin mad$x$ a md(x), za předpokladu, že je aplikujete na normální rozdělení:

$$\delta\mathrm{mad}(x) = 0.6745\,\mathrm{std}(x); \qquad \mathrm{md}(x) = \sqrt{2/\pi}\,\mathrm{std}(x); \tag{12}$$

$$\frac{\delta\mathrm{std}_{\mathrm{mod}}(x)}{\sigma(x)} = \frac{0.71}{\sqrt{n}}; \qquad \frac{\delta\mathrm{mad}(x)}{\mathrm{mad}(x)} = \frac{1.17}{\sqrt{n}}; \qquad \frac{\delta\mathrm{md}(X)}{\mathrm{md}(x)} = \frac{0.76}{\sqrt{n}}; \tag{13}$$

Zde je jasně vidět, nejvíce informací přenáší nejméně robustní metoda určení rozptylu - std(x), v závěsu je nástroj md(x) a nejméně mad(x), který tak platí za svou robustnost.
Příklad: Dejme tomu, že máte 24 bodů s normálním rozdělením s centrem v nule, $\sigma = 1$ a jedním odlehlým bodem, $x_{25} = 10$.
Aritmetický průměr: $x = 0.40$, střední medián: $\tilde{x} = 0.052$, standardní odchylka: $s = 2.23$. Střední odchylka s centrem v aritmetickém středu $\mathrm{md}(x,\bar{x}) = 1.196 \Rightarrow s_{\mathrm{pred}} = 1.498$, střední odchylka s centrem v mediánu $\mathrm{md}(x,\tilde{x}) = 1.140 \Rightarrow s_{\mathrm{pred}} = 1.498 \Rightarrow s_{\mathrm{pred}} = 1.428$. Konečně nejrobustnější odhad s mediánem střední odchylky $\mathrm{mad}(x,\tilde{x}) = 0.687 \Rightarrow s_{\mathrm{pred}} = 1.019$!
Je tedy zjevné že v tom třetím případě odhad reálných charakteristik nebyl přítomností odlehlého bodu takřka vůbec ovlivněn. Jako počáteční odhad pro robustní regresi je optimální a ušetří spoustu zdlouhavých iterací, pokud se ovšem nespokojíme se samotným odhadem.

2.4. Robustní regrese

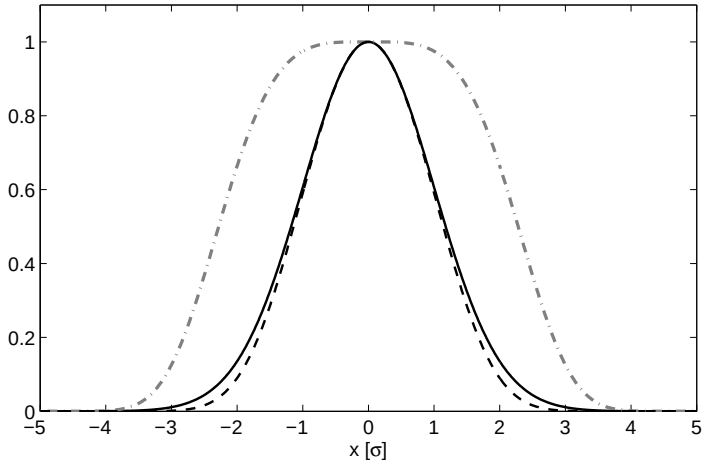

Fig. 5. Robustní regrese. Plnou čarou je naznačena normální rozdělovací funkce, čerchovanou čarou je naznačen filtr robustní regrese, čárkovaně je znázorněn součin filtru a normální rozdělovací funkce.

Metoda nejmenších čtverců právem patří mezi ty nejpoužívanější metody zpracování astrofyzikálních dat. Pomocí ní lze velice dobře určit parametry různých modelů, kterými se snažíme popsat pozorovanou realitu, přičemž je lhostejné, zda jde o modely fyzické nebo jenom fenomenologické. Bohužel častá přítomnost odlehlých bodů tuto metodu silně poškozuje a způsobuje, že mnohé výsledky bývají sporné, zejména pak není možné se spolehnout na odhad neurčitostí parametrů modelu a na neurčitost předpovědí.

Aby bylo možné zachovat komfort poskytovaný MNČ a současně přitom eliminovat vliv odlehlých bodů i chybějících neprávem vymazaných bodů, se používá řady metod, které se řadí do kategorie robustních regresí, které jsou jen minimálně citlivé na hrubé chyby a přítomnost odlehlých bodů.

Popíšu zde svou vlastní metodu, která je jednoduchá, rychle iteruje a mám ji dobře odzkoušenou na řadě reálných i simulovaných úloh.

Nechť $\{x_i, w_i\}$ je soubor sestávající z proměnné veličiny (nejčastěji jde o rozdíl mezi pozorovanou veličinou a modelovanou funkcí), která by měla mít povahu náhodné veličiny s rozdělení blízkým normálnímu, a vahou jednotlivých bodů určenou na základě jejich očekávaných individuálních neurčitostí. Nechť $\bar{x}_0$ a s_0 jsou počáteční odhady váženého průměru a standardní odchylky získané nejlépe pomocí nástrojů zmiňovaných v předcházející podkapitole. V dalším kole iterace se váhy upraví multiplikativním faktorem $\xi_{i,j}$ (i je číslo měření v souboru, j je pořadové číslo iterace) podle vztahu:

$$\xi_{i,j} = 1.06 \, \exp\left[- \left(\frac{x_i - \bar{x}_{j-1}}{2.5 \, s_{j-1}} \right)^4 \right]; \quad (14)$$

$$n_j = \sum_{i=1}^{n} \xi_{i,j} - 0.12; \quad \bar{x}_j = \frac{\sum_{i=1}^{n} x_i \, \xi_{i,j} \, w_i}{\sum_{i=1}^{n} \xi_{i,j} \, w_i}; \quad s_j = \sqrt{\frac{1.23 \, n_j}{(n_j - 1)} \, \frac{\sum_{i=1}^{n} (x_i - \bar{x}_j)^2 \, \xi_{i,j} \, w_i}{\sum_{i=1}^{n} \xi_{i,j} \, w_i}}. \quad (15)$$

Dle zkušenosti stačí šest iterací, pak už se výsledky nemění. Během iterací se postupně zpřesňují hodnoty parametrů $\bar{x}_j$, s_j a n_j, což odhad reálného počtu měření (bez odlehlých

a s body neoprávněně vyřazenými), tak, aby se rozdělovací funkce co nejvíce podobala normální.

Vyzkoušíme robustní regresi na náš příklad z předchozí podkapitoly, střední hodnota vychází jako přesná nula, disperze je 0.995.

Ze statistických simulací vyplývá, že střední nejistota určení parametru n_j, $\delta n_j = 0.10\sqrt{n}$. Při stovce bodů je tedy přesnost 1%, při $n = 10000$ pak desetkrát méně! Je-li $n_j > n$ lze očekávat, že v datovém souboru nějaká měření chybějí, v opačném případě, že soubor obsahuje odlehlé body. Pokud jde o relativní přesnost určení standardní odchylky, tak platí $\delta s_j/s_j = 0.81/\sqrt{n}$, což je jen o malounko horší, než při standardním postupu (viz vztahy ve (13)). Důvod je zřejmý - příspěvek bodů vzdálených centra, které především určují rozptyl, je v této verzi robustní regrese potlačen.

2.5. Testy normality pozorovaných rozdělovacích funkcí

Pokud je jasnost hvězdy konstantní, pak by měla být rozdělení normální. Normálnost (normalitu) rozdělovací funkce lze nejsnáze testovat pomocí šikmosti γ_1 a excesem špičatosti γ_2, jen je třeba rozvážit, zda jsou tyto diagnostické nástroje dostatečně jemné a účinné. Zmíněné nástroje v sobě obsahují třetí nebo dokonce i čtvrtou mocninu vzdálenosti od centra, takže "nadržují"nejvíce vzdáleným bodům. Těch je relativně málo, takže jejich statistika bývá vrtkavá.

2.6. Indexy normality

Pro exces špičatosti lze zkonstruovat tzv. *index normality* rozdělovací funkce $\Lambda_{\rm kurt}$, vycházející z poměru γ_2 a jeho neurčitosti $\delta\gamma_2$ (viz vztah 8), tedy

$$\Lambda_{\rm kurt} = \gamma_2/\delta\gamma_2 = 0.22\,\sqrt{n}\,\left(\frac{n+1}{n-1}\,\frac{\overline{(x-\bar{x})^4}}{s^4} - 3\right). \qquad (16)$$

Leží-li index normality Λ v intervalu od -2.5 do 2.5, pak je rozdělení nejspíš normální, je-li větší, pak jsou v datech zastoupeny odlehlé body, v opačném případě rozdělovací funkce naznačuje, že buď v datovém souboru někdo odmazal nějaká měření nebo že zdroj je proměnný.

Index normality spojený s excesem špičatosti lze úspěšně nahradit robustnějším indexem $\Lambda_{\rm mad}$, vycházejícím z poměru robustní míry rozptýlení mad(x) a standardní odchylka std(x). Je-li rozdělení normální, pak je bezrozměrný střední poměr $\bar{r} = \overline{\rm mad(x)/std(x)} = 0.6744$. Standardní odchylka tohoto poměru daná náhodným rozložením bodů závisí na počtu bodů: $\mathrm{std}(r) = 0.622/\sqrt{n}$.

$$\Lambda_{\rm mad} = \frac{\bar{r} - r}{\mathrm{std}(x)} = \sqrt{n}\left(1.084 - \frac{1.608\,\mathrm{mad}(x)}{\mathrm{std}(x)}\right). \qquad (17)$$

Obdobně lze definovat index normality $\Lambda_{\rm RR}$ pomocí parametru n_j vycházejícího z robustní regrese popsané v podkapitole 2.4

$$\Lambda_{\rm RR} = \frac{n - n_j}{\mathrm{std}(n_j)} = \frac{10.0\,(n_j - n)}{\sqrt{n}}. \qquad (18)$$

Na řadě konkrétních příkladů lze ukázat, že naposled zmiňovaný index je ze všech tří zmiňovaných indexů normality nejcitlivější.

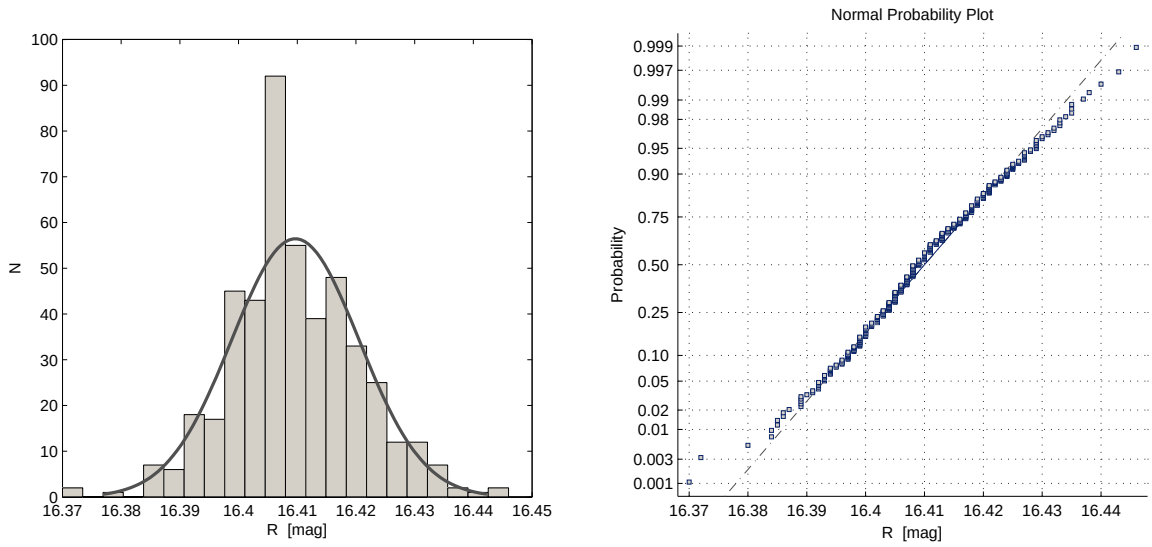

Fig. 6. Vlevo: Histogram naměřených světelných změn jedné hvězdy ve Velkém Magellanově mračnu podezřelé z příslušnosti k CP hvězdám. Naznačená křivka je pokusem o proložení normální rozdělovací funkce. Rozdělovací funkce je však zjevně špičatější s množstvím odlehlých bodů, což svědčí o tom, že tato hvězda bude nejspíš neproměnná. **Vpravo:** Do velkých detailů lze odchylky od normální funkce vysledovat na grafu s normální pravděpodobností. Všimněte si například výskytu odlehlých bodů na horní a dolní části grafu: vzhledem k tomu, že pravděpodobnost jednotlivých bodů je pevně dána počtem měření, graf ukazuje, že při normálním rozdělení by měly být body na okrajích více přimknuty ke středu.

2.7. Další metody testování normality

Názorným prostředkem umožňujícím posoudit případné odchylky rozdělovací funkce od normální funkce je utilita v Matlabu označovaná 'normplot.m'. Je to v podstatě kumulativní distribuční funkce s upravenou osou y. Na x-ovou osu vynášíme naměřenou veličinu (třeba hvězdnou velikost) a na osu y pak pravděpodobnost, že daná veličina naměří. Zobrazení na ose y je takové, že pokud je rozděleni normální, pak jsou jednotlivé body rozloženy podél přímky. Parametry normálního rozdělení se přitom určují z mediánu a mezikvartilního rozpětí, což je robustní míra rozptýlení založena na vzdálenosti mezi 1. a 3. kvartilem.

3 Diagnostika proměnnosti

Na proměnnost zkoumaných objektů lze usuzovat z řady okolností. Především, pokud zjistíte, že rozptyl hodnot sledované veličiny je zjevně větší, než by vyplývalo z nejistoty jejího měření, lze to nejpřirozeněji vysvětlit tak, že kromě náhodných chyb jsou tu ve hře i reálné časové změny zkoumané veličiny. Je však třeba se mít na pozoru, pokud na nejistotu měření usuzujete pouze na základě udávané vnitřní chyby měření. Tyto údaje občas bývají podceněné, což občas mívá tu příčinu, že udávaná nepřesnost sice dobře vystihuje relativní kvalitu jednotlivých pozorování, ale v absolutní velikosti neodpovídají. Zde je pomoc snadná - zjistíme-li, že rozptyl veličin korelace s udanou neurčitostí jednotlivých měření,

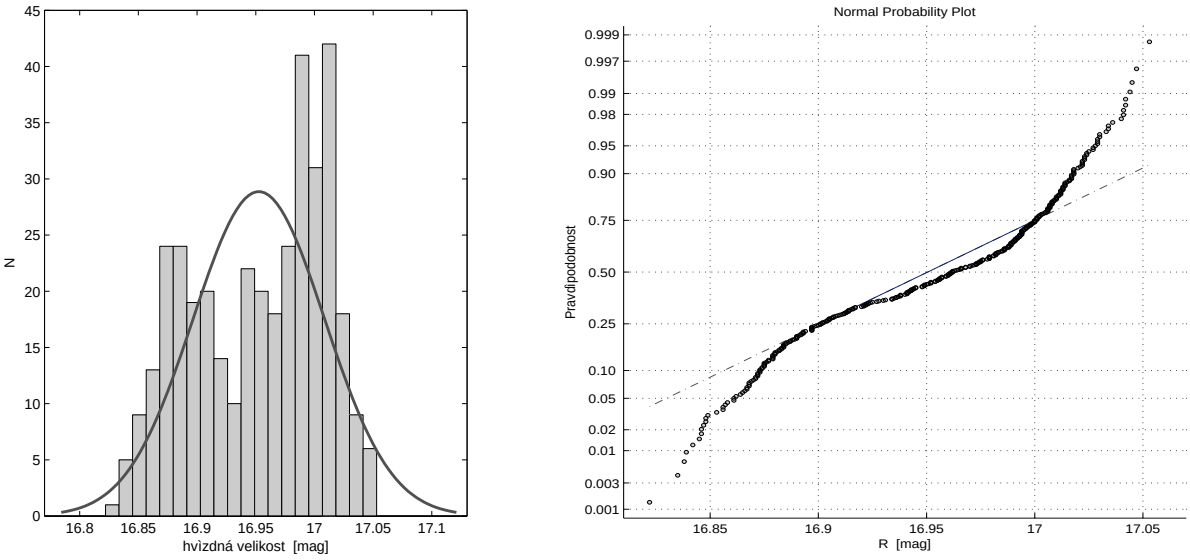

Fig. 7. Vlevo: Histogram naměřených světelných změn jedné trpasličí cefeidy ve Velkém Magellanově mračnu. Naznačená křivka je pokusem o proložení normální rozdělovací funkce. Rozdělovací funkce je však zjevně bimodální, což jasně svědčí o tom, že nejspíš jde o periodicky proměnnou hvězdu. **Vpravo:** Do velkých detailů lze odchylky od normální funkce vysledovat na grafu s normální pravděpodobností. Zde si povšimněte velké odchylnosti výsledné křivky od ideálu normálního rozdělení jak v okolí centra, tak i v křídlech. Typický vzhled pro periodickou světelnou křivku.

je třeba udané neurčitosti modifikovat vynásobením vhodnou konstantou.

Jiné to ovšem je, pokud zjistíme, že tu korelace není nebo je slabá, pak to může být způsobeno jak tím, že dotyčný zdroj se skutečně mění nebo se při stanovení vnitřní chyby nevzaly v úvahu veškeré příčiny pozorovaného rozptylu. Zde zase může hodně napovědět to, jak se chovají srovnávací objekty - tedy nejčastěji jasnosti srovnávací a kontrolní hvězdy. Pokud zjistíte, že diskrepance mezi pozorovaným a udávaným rozptylem existuje jen u objektu podezřelého z proměnnosti, pak je skoro jisté, že dotyčný objekt je z proměnnosti podezřelý právem.

Mnohé ovšem napoví i povaha závislosti dané veličiny na čase. Především se zjistí v jaké časové škále se pozorované změny odehrávají, tedy zda jde o dlouhodobé změny, nazývané jako trendy, nebo jde o periodické změny nejrůznějších příčin nebo změny epizodické, kataklizmické. Hodně informací v sobě ale obsahuje i sama distribuční funkce pozorované veličiny, jejíž analýza by měla předcházet další pokusy o výklad povahy proměnnosti. V dalším se proto budeme zamýšlet nad tím, jak rozdělovací funkce a její charakteristiky, zejména pak špičatost a šikmost souvisejí s povahou časových změn zkoumané veličiny.

Ve valné většině případů bývají proměnné hvězdy odhaleny na základě jejich světelných změn, budeme proto v dalším hovořit o světelných změnách, i když ony závěry je možné aplikovat i na proměnnost jiných veličin, jako je třeba magnetická indukce, radiální rychlost nebo intenzita spektrálních čar.

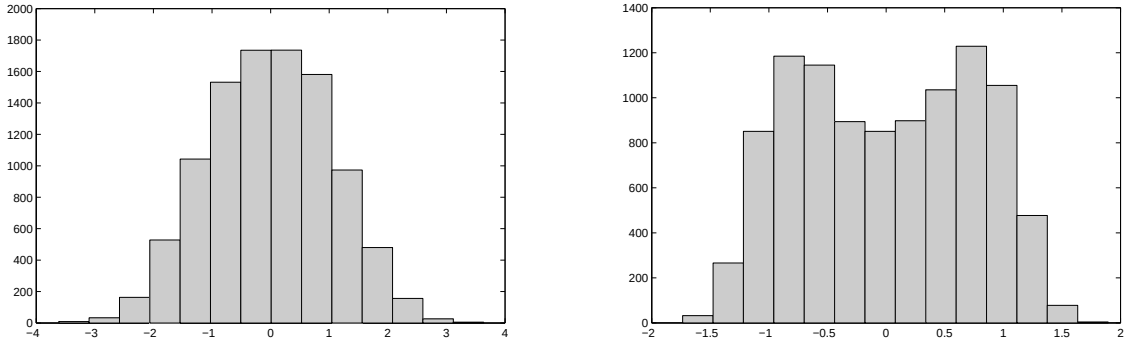

Fig. 8. Vlevo: Histogram sinových světelných změn se semifinalistkou 1 a rozptylem s $\sigma = 3/4$. **Vpravo:** U-histogram sinových světelných změn se semiamplitudou 1 a rozptylem s $\sigma = 1/4$.

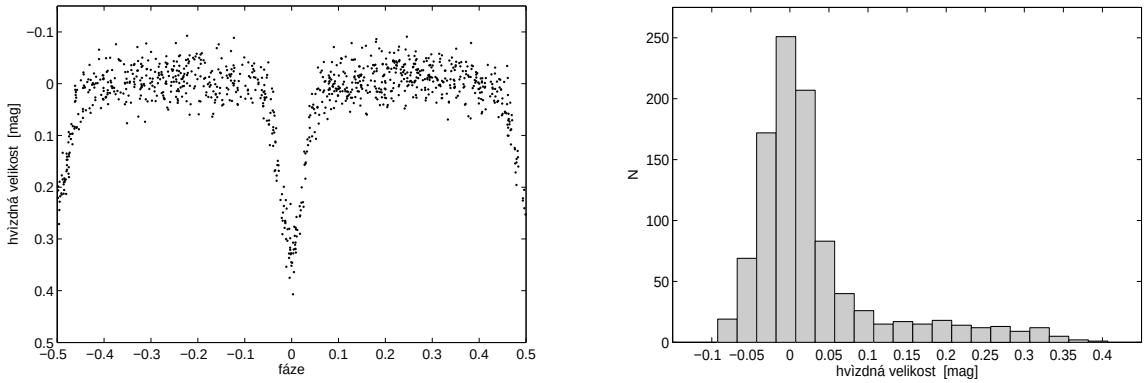

Fig. 9. Vlevo: Simulace světelné křivky zákrytové dvojhvězdy mírně zašumění. **Vpravo:** Asymetrický histogram charakteristický pro zákrytovou dvojhvězdu.

Reference