

Popisná statistika I

Zdeněk Mikulášek, Ústav teoretické fyziky a astrofyziky

Výsledkem série astrofyzikálních měření vybrané veličiny y nějakého objektu (hvězdná velikost, intenzita, radiální rychlost) bývá zpravidla datový soubor $D = \{t_i, y_i, \delta y_i\}$. Ten ve formě uspořádaných trojic čísel obsahuje informaci o jednotlivých měření souboru, speciálně tedy o okamžiku jednotlivého měření t_i , naměřené hodnotě zkoumané veličiny y_i a odhadu nejistoty jejího určení δy_i . Počet těchto trojic (počet řádků) pak udává celkový počet měření n . V datovém souboru se mohou vyskytovat tytéž hodnoty i vícekrát, zejména tehdy, mají-li veličiny diskrétní (nespojitou) povahu (počet rohlíků).

Pokud chceme soubor dat blíže poznat a kvantitativně popsat, můžeme k tomu s výhodou použít některý ze standardních nástrojů tzv. *popisné statistiky* (descriptive statistics). Popisná statistika, jakkoli je primárně určená k zpracování výsledků měření a jejich nejistot $\{y_i, w_i\}$, je může sloužit i k popisu časových údajů¹.

1 Nejistota jednotlivého měření a váha

Pokud není kvalita jednotlivých pozorování stejná, je vysoce žádoucí tuto kvalitu vyjádřit nezáporným číslem tzv. vahou - w_i . Váha se vztahuje vždy k jednomu, konkrétnímu měření, proto ji nezaměňujte s četností příslušného výsledku. Vzhledem k tomu, že většina metod zpracování výsledků je tak či onak založena na metodě nejmenších čtverců, pak by měla tato váha souviset s vnitřní nejistotou určení hodnoty konkrétního měření odhadem δy_i :

$$w_i = K \delta y_i^{-2}, \quad (1)$$

kde K je vhodně zvolené nezáporné číslo².

Váhy bychom neměli použít v případě, kdy se ukáže, že očekávaná nejistota jednotlivých měření v souboru je výrazně menší, než jejich celkový rozptyl v rámci souboru. Formálně to pak lze chápat tak, že všechna měření mají tutéž váhu, rovnu např. 1. Naopak jsme je povinni použít pokud jsou deklarovány (a my se nepřesvědčíme, že jsou nějakým způsobem vadné a tudíž nepoužitelné), a pak zejména při transformaci měřených veličin nějakou nelineární funkcí ($\log y$, $1/y$) nebo při některých robustních metodách zpracování výsledků.

V dalším textu budeme používat následující konvence:

$$\overline{y^k} = \frac{\sum_{i=1}^n y_i^k w_i}{S_w}, \quad \text{kde} \quad S_w = \sum_{i=1}^n w_i, \quad \overline{w} = \frac{S_w}{n}. \quad (2)$$

¹K prozkoumání závislosti měřených veličin na čase, čili k analýze časových řad používáme tzv. *regresní analýzu*, ta však není předmětem této kapitoly.

²Třeba $K = 1$, $K = n / \sum_{i=1}^n \delta y_i^{-2}$.

2 Míra polohy

2.1. Průměry

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 y_m nazývaná též *aritmetický průměr*, respektive *průměr*, obecně *váhováný průměr*. Pro aritmetický průměr veličinu platí, že suma váhovaných čtverců odchylek jednotlivých měření od tohoto centra y_m , minimální:

$$y_m = \bar{y} = \frac{1}{S_w} \sum_{i=1}^n y_i w_i, \quad \Rightarrow \quad \sum_{i=1}^n (y_i - y_m) w_i = \overline{(y - y_m)} = \overline{(y - \bar{y})} = 0. \quad (3)$$

Mnohem řidčeji se používá *geometrický průměr* y_{mG} , *harmonický průměr* y_{mH} nebo *kvadratický průměr* y_{mQ} :

$$y_{mG} = \sqrt[n]{\prod_{i=1}^n y_i}, \quad y_{mH}^{-1} = \overline{y^{-1}}, \quad y_{mQ}^2 = \overline{y^2}. \quad (4)$$

2.2. Medián

Medián (označován $\text{med}(y)$ nebo \tilde{y}) 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 y s jejich váhami w podle velikosti, takže $y_1 < y_2 \dots < y_k < \dots < y_n$.
- 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 $\text{medianw}(y, w) = \tilde{y} = [(W_{j+1} - 0.5) * y_j + (0.5 - W_j) * y_{j+1}] / (W_{j+1} - W_j)$ je pak oním hledaným váhovaným mediánem.

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

2.3. Histogram, kvantily, kumulativní distribuční funkce

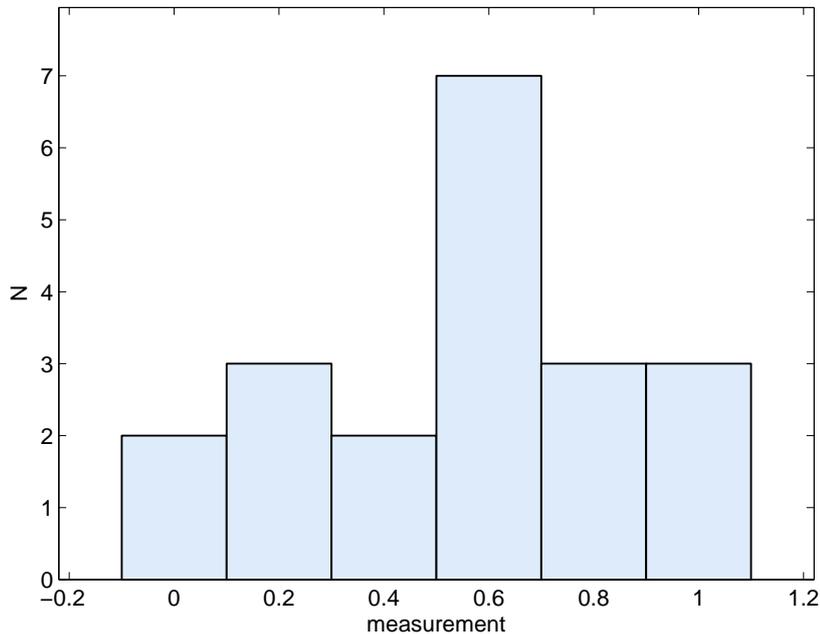


Fig. 1. Histogram měření extinkčního koeficientu.

Nejinstruktivnějším vyjádřením distribuční funkce je u diskretních veličin tzv. tyčkový graf, v případě spojité veličiny pak histogram (histogram). Celý interval pokrytý daty se rozdělí na vhodný počet n_h ekvidistantních intervalů a počítá se počet (četnost), respektive suma vah dat k nim příslušejících. Graficky se potom distribuční funkce znázorní sloupcovým diagramem. Doporučený počet sloupců pro n měření udává Sturgesovo pravidlo: $n_h = 1 + 3,3 \log n$.

Podrobněji lze měření popsat pomocí *kvantilů* (quantile). Kvantil odpovídající hodnotě p z intervalu $0 < p < 1$, je takové číslo y z intervalu $\langle y_1, y_n \rangle$, pro nějž platí, že $p \times n$ hodnot souboru je menších než y a $(1-p) \times n$ větších. *Vážený kvantil* (weighted quantile) se zavádí u vzorků s měřeními o nesteranné váze. Pokud je zkoumaný soubor vzorkem nějakého většího souboru, pak kvantil $p(y)$ je odhadem pravděpodobnosti, že nějaké náhodně vybrané číslo ze souboru bude menší než zvolená hodnota y . Rozdíl $p(y_a) - p(y_b)$ pak udává odhad pravděpodobnosti, že se takové číslo vyskytne v intervalu $\langle y_b, y_a \rangle$. Je-li p vyjádřeno v procentech, pak se kvantilu říká *percentil* (percentile).

Zvláštní význam má kvantil pro $p = 0,5$ (50%), nazývaný *median*, *první kvartil* (first quartile) — $p = 0,25$ (25%) a *třetí kvartil* (third quartile) — $p = 0,75$ (75%).

Výše naznačený předpis je jen rámcový, pro algoritmus výpočtu kvantilů je nutno být konkrétnější. Výhodné je proto definovat si tzv. *kumulativní distribuční funkci*, případně *váhouvanou kumulativní distribuční funkci* $\Phi(y)$, která vyjadřuje závislost kvantilu p na měřené veličině x . Kumulativní distribuční funkce $\Phi(y)$ je představována lomenou čarou s uzlovými body v $\{y_i, p_i\}$.

Pro p_i platí: $p_1 = 1/(2n)$, $p_i = p_{i-1} + 1/n \Rightarrow p_i = (1 + 2i)/(2n)$ pro $y < y_1$ je hodnota p rovna nule, pro $y > y_n$ je funkce rovna 1. Obdobně pak váhovaná kumulativní

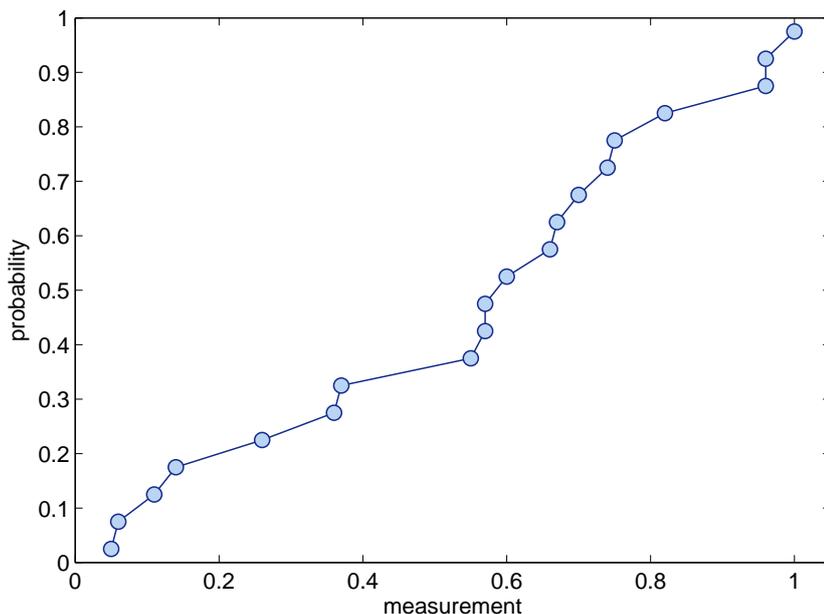


Fig. 2. Kumulativní distribuční funkce pro měření extinkčního koeficientu - viz Úloha. Dosti se odlišuje od ideálu normálního rozdělení.

distribuční funkci $\Phi(y)$ je představována lomenou čarou s uzlovými body v $\{y_i, p_i\}$. Pro p_i platí: $p_1 = w_1/(2S_w)$, $p_i = p_{i-1} + (w_{i-1} + w_i)/(2S_w)$, pro $y < y_1$ je hodnota p rovna nule, pro $y > y_n$ je funkce rovna 1.

Ořezaný medián (trimmed mean) jakožto robustní odhad polohy centra je jistým kompromisem mezi aritmetickým průměrem a mediánem. Jako parametr se používá veličina p vyjádřená zpravidla v procentech (nejčastěji $p = 10\%$). Ze seřazeného souboru dat odstraníme $\text{round}(p/2)$ nejvyšších a stejný počet nejnižších hodnot a ze zbytku vypočteme aritmetický průměr. Pro $p = 0$ jde o aritmetický průměr, blíží-li p 100%, pak jde o medián. U váhovaných veličin je definice ořezaného průměru poněkud vágní a proto se běžně nepoužívá.

Modus je nejčetněji zastoupená hodnota měření (nebo hodnota s největší vahou). Význam má u diskretních výsledků měření, nebo v určitých intervalech – nejpohodlněji ji lze odečíst z histogramu.

3 Míry rozptýlení

Nejčastější mírou rozptýlení dat kolem centra je takzvaný *rozptyl* (variance) s^2 nebo *směrodatná odchylka* (standard deviation) s .

$$s^2 = \frac{1}{S_w} \sum_{i=1}^n (y_i - \bar{y})^2 w_i = \overline{y^2} - \bar{y}^2 \quad (5)$$

Centrem rozptýlení je zde aritmetický průměr. Dokažte, že právě pro něj nabývá funkcionál $S(a) = \sum (y_i - a)^2 w_i$ svého minima.

Směrodatná odchylka je vynikající položkou popisující rozptýlení dat kolem aritmetického průměru, pokud se v datech nenacházejí tzv. odlehlé body, případně pokud nebyla napozorovaná data ještě nějakým neodborným způsobem upravována, například neoprávněným vypouštěním bodů - viz. Fig. 4.

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

$$\text{md} = \overline{|y - a|}; \quad \text{wmd} = \frac{1}{S_w} \sum_{i=1}^n |y_i - a| w_i; \quad (6)$$

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

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

$$\text{mad} = \text{median}(|y - \tilde{y}|); \quad \text{wmad} = \text{medianw}(|y - \tilde{y}|); \quad (7)$$

V případě, že máte data s množstvím odlehlých bodů a potřebujete dobrý odhad disperze, pak můžete tyto vztahy s výhodou použít.

Jednoduchým robustním odhadem rozptýlení, který lze v případě nekvalitních dat aplikovat je tzv. *mezikvartilní rozpětí* (interquartile range) Δ_{13} . Jde o rozdíl mezi 3. a 1. kvantilem, takže se vztahuje jen na vnitřní část rozdělovací křivky.

4 Normální rozdělení

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 :

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y - \mu)^2}{2\sigma^2}\right]. \quad (8)$$

Gaussovský „Říp“ je přísně symetrický podle osy $y = \mu = \bar{y}$, kterážto hodnota je současně aritmetickým průměrem, mediánem i modem souboru podřizujícímu se normálnímu rozdělení. Lze ukázat, že směrodatná odchylka s je právě rovna parametru popisujícímu šířku normálního rozdělení σ (disperze), tedy:

$$s^2 = \overline{(y - \mu)^2} = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (y - \mu)^2 \exp\left[-\frac{(y - \mu)^2}{2\sigma^2}\right] dy = \sigma^2. \quad (9)$$

4.1. Odhad μ a σ

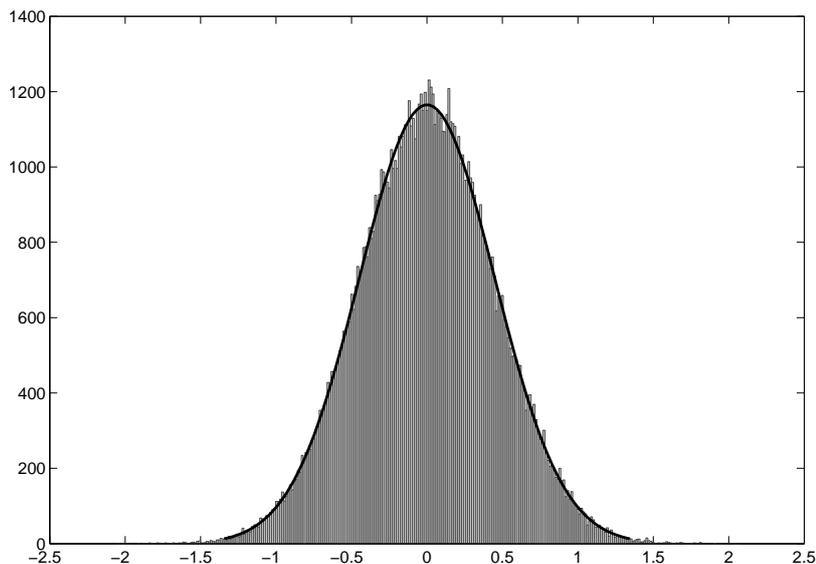


Fig. 3. Simulace výsledků odhadů centra v aritmetickém průměru (vlevo) a standardní odchylky (vpravo) pro normální rozdělení s centrem v 0 a standardní odchylkou 1 při výběru pouhých 5 bodů. Tento výběr byl ovšem opakován 100 000krát..

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. Za předpokladu, že zkoumaný soubor má normální rozdělení, pak lze ukázat, že parametry μ, σ a jejich nejistoty $\delta\mu$ a $\delta\sigma$ lze odhadnout pomocí vztahů:

$$\mu \cong \bar{y}; \quad \delta\mu = \frac{\sigma}{\sqrt{n}}, \quad \sigma \cong s \sqrt{\frac{n}{n-1}} = \sqrt{\frac{n}{n-1}(\bar{y}^2 - \bar{y}^2)}, \quad \delta\sigma = \frac{\sigma}{\sqrt{2n}}. \quad (10)$$

Pokud je distribuční funkce narušena třeba výskytem odlehlých bodů, je vhodné místo aritmetického průměru použít raději medián a disperzi odhadnout pomocí robustnějších indikátorů, jako váhovaná střední (absolutní) odchylka $\text{mad}(y)$, medián střední odchylky $\text{md}(y)$ nebo mezikvartilní rozpětí Δ_{13} :

$$\sigma(y) \cong 1.482 \text{mad}(y); \quad \sigma(y) \cong 1.253 \text{md}(y); \quad \sigma(y) \cong 0.741 \Delta_{13}; \quad (11)$$

To, proč zde hovoříme jen o odhadech příslušných veličin, dostatečně ilustrují obrázky 4 a 5 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 $\rho = 1/\sqrt{2n}$ 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 podřídít počet míst a způsob zaokrouhlování (v těchto případech se přimlouvám zaokrouhlovat vždy spíše nahoru).

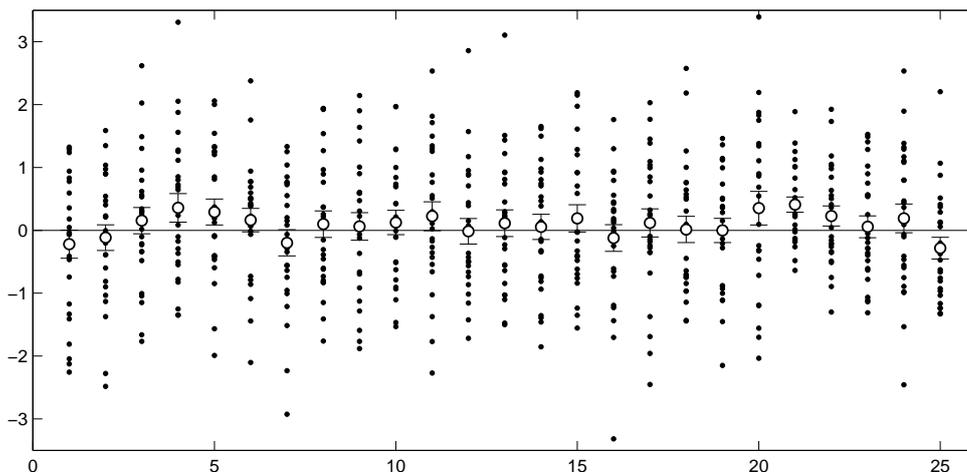


Fig. 4. Simulace výsledků 26 měření pro normální rozdělení s centrem v 0 a standardní odchylkou 1. Jednotlivá měření jednotlivých sad jsou znázorněna nad sebou plnými kotoučky, průměr s jeho nejistotou je naznačen větším prázdným kroužkem a chybovou úsečkou. Povšimněte si jak odlišné může být rozložení těchto bodů v jednotlivých sadách, rovněž tak, že body s odchylkou 3σ jsou zcela běžné - v tomto případě tedy nejde o odlehle body. Prostudujte i obr. 5, který je dalším zpracování této simulace.

5 Úloha

Výsledkem měření atmosférické extinkce z pozorování komet na observatoři Skalnaté Pleso jsou tyto hodnoty extinkčních koeficientů ve vlnové délce 416 nm (mag/vzdušnou hmotu):

$$\begin{array}{cccccc}
 0.82 \pm 0.07 & 0.39 \pm 0.03 & 0.54 \pm 0.05 & 0.57 \pm 0.03 & 0.42 \pm 0.04 & \\
 0.39 \pm 0.07 & 0.69 \pm 0.05 & 0.81 \pm 0.05 & 0.33 \pm 0.05 & 0.41 \pm 0.04 & \\
 0.11 \pm 0.07 & 0.23 \pm 0.04 & 0.39 \pm 0.04 & 0.43 \pm 0.04 & 0.97 \pm 0.03 & \\
 0.26 \pm 0.05 & 0.47 \pm 0.04 & 0.41 \pm 0.05 & 0.52 \pm 0.04 & 0.45 \pm 0.03 &
 \end{array} \tag{12}$$

Instrumentářem popisné statistiky charakterizujte tento soubor, speciálně pak uveďte:

- počet měření a jejich charakter (spojité, diskrétní?)
- váhy jednotlivých měření a diskutujte, zda je v tomto případě případné tyto váhy použít. Bez ohledu na výsledek úvahy počítejte všechny další úlohy ve dvou variantách – s vahami a bez nich.
- odhad aritmetického průměru a jeho nejistotu za předpokladu normálního rozdělení, harmonický, geometrický, kvadratický průměr a medián, ořezaný průměr pro 10% a 20% (jen pro případ bez vah)
- minimální a maximální hodnotu extinkce a celkové rozpětí
- rozptyl s , odhad disperze σ , střední velikost odchylky s centrem v aritmetickém průměru a v mediánu

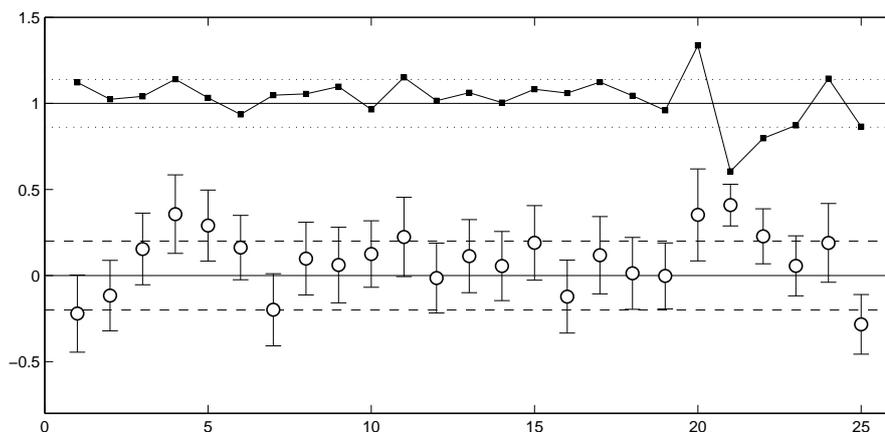


Fig. 5. Simulace výsledků měření pro normální rozdělení s centrem v 0 a standardní odchylkou 1. Každý z 25 výsledků byl zkonstruován z 26 individuálních měření. Je patrné jak kolísání polohy aritmetického průměru kolem nuly, tak i kolísání hodnoty naměřené směrodatné odchylky.

f) graf kumulativních distribuční funkce a pomocí ní stanovte hodnoty kvartilů a mezikvartilního rozpětí

g) Porovnejte odhady μ a σ pro normální rozdělení získané různými metodami

i) pomocí stanovte optimální počet sloupců v histogramu a sestrojte jej. Doporučuji sloupce v histogramu centrovat na násobky 0.2

j) odhadněte modus rozdělení

k) diskutujte tvar rozdělovací funkce s vědomím, že konstantní složka extinkčního koeficientu ve 416 nm způsobená Rayleighovým rozptylem na náhodných shlucích molekul vzduchu činí 0.262 mag/vzdušnou hmotu.

Reference