Popisná statistika

(Descriptive statistics)

Výsledkem měření je soubor n naměřených hodnot vytvářející datový soubor $D = \{x_i\}$. V datovém souboru se mohou vyskytovat tytéž hodnoty i vícekrát, zejména tehdy, mají-li veličiny diskrétní (nespojitou) povahu (počet rohlíků).

Pokud chceme tento soubor dat blíže popsat, použijeme některý z instrumentů tzv. popisné statistiky.

1 Váha

Pokud není kvalita jednotlivých pozorování stejná, je užitečné ji popsat nezáporným číslem tzv. *vahou* - w_i. Váha se vztahuje vždy k jednomu, konkrétnímu měření, proto ji nezaměňujte s četností příslušného výsledku. Váha většinou souvisí s odhadem tzv. vnitřní nejistoty určení hodnoty konkrétního měření - δx_i:

$$w_i \sim (\delta x_i)^{-2}.$$

Zkušenost ukazuje, že zavedením vah se globální charakteristiky souboru obvykle změní jen nevýznamně, a proto je třeba si předem rozmyslet, zda váhy při výpočtech vůbec použijeme.

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. Naopak jsme je *povinni* použít pokud jsou deklarovány, tedy zejména při transformaci měřených veličin nějakou nelineární funkcí ($\log x$, $1/x$) nebo při některých robustních metodách zpracování výsledků.

Zaveďme si sumu vah S_w a střední váhu w_s:

$$S_w = \sum_{i=1}^{n} w_i, \quad w_s = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{S_w}{n}.$$

2 Míra polohy

Nejznámější a nejpoužívanější mírou vztahující se ke středu studovaného datového souboru je tzv. *aritmetický průměr*, často jen *průměr* (arithmetic mean, mean), případně váhovaný průměr (weighted mean):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{x} = \frac{1}{S_w}\sum_{i=1}^{n} x_i\, w_i.$$

Důležitou vlastností průměru je fakt, že: $\sum (x_i - \bar{x}) = 0$, resp. $\sum (x_i - \bar{x})\, w_i = 0$.

Geometrický průměr (geometric mean):

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n} \qquad \bar{x}_G = \sqrt[S_w]{x_1^{w_1} x_2^{w_2} \ldots x_n^{w_n}}\ .$$

Harmonický průměr (harmonic mean):

$$\bar{x}_H^{-1} = \frac{1}{n}\sum_{i=1}^{n} x_i^{-1}, \qquad \bar{x}_H^{-1} = \frac{1}{S_w}\sum_{i=1}^{n} x_i^{-1} w_i\ .$$

Kvadratický průměr (quadratic mean):

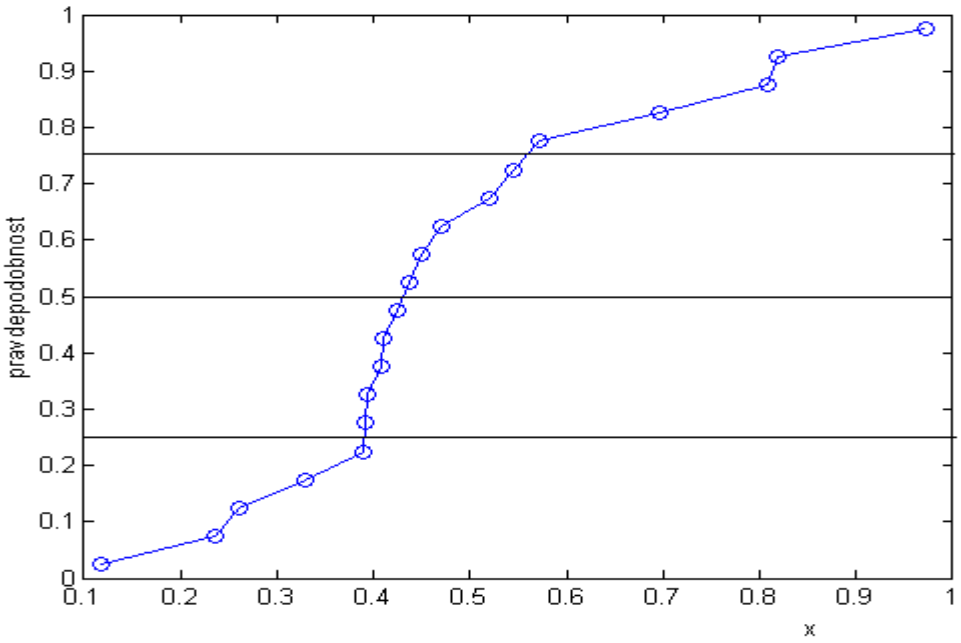
$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \overline{x^2} = \frac{1}{S_w}\sum_{i=1}^{n} x_i^2 w_i \ .$$

Pro další charakteristiky je vhodné soubor $\{x_i\}$ případně $\{x_i, w_i\}$ seřadit podle velikosti x_i.

Kvantil (quantile) určený číslem p, $0<p<1$ je číslo x z intervalu $<x_1, x_n>$, pro nějž platí, že pn hodnot souboru je menších než x a $(1-p)\,n$ větších. *Vážený kvantil* (weighted quantile) se vztahuje k vahám. Pokud je zkoumaný soubor vzorkem nějakého většího souboru, pak kvantil $p(x)$ je odhadem pravděpodobnosti, že nějaké náhodně vybrané číslo ze souboru bude menší než zvolená hodnota x. Rozdíl $p(x_a)$ - $p(x_b)$ pak udává odhad pravděpodobnosti, že se takové číslo vyskytne v intervalu $<x_b, x_a>$. Je-li p vyjádřeno v procentech, pak se kvantilu říká *percentil* (percentile). Zvláštní význam má kvantil pro p = 0,5 (50 %), nazývaný *medián*, *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 k tomu definovat si tzv. *kumulativní distribuční funkci*, případně *váhovanou kumulativní distribuční funkci* $\Phi(x)$, která vyjadřuje závislost kvantilu p na měřené veličině x. Kumulativní distribuční funkce $\Phi(x)$ je představována lomenou čarou s uzlovými body v $\{x_i, p_i\}$.

Pro p_i platí: $p_1 = 1/(2\,n)$, $p_i = p_{i-1} + 1/n \Rightarrow p_i = (1+2\,i)/(2n)$ pro $x < x_1$ je hodnota p rovna nule, pro $x > x_n$ je funkce rovna 1. Obdobně pak váhovaná kumulativní distribuční funkci $\Phi(x)$ je představována lomenou čarou s uzlovými body v $\{x_i, p_i\}$. Pro p_i platí: $p_1 = w_1/(2\,S_w)$, $p_i = p_{i-1} + (w_{i-1}+w_i)/(2\,S_w)$, pro $x < x_1$ je hodnota p rovna nule, pro $x > x_n$ je funkce rovna 1.



Medián (median) $\tilde{x}$ nebo váhovaný medián – je oblíbená robustní míra polohy centra souboru, jež prakticky nezávisí na výskytu „odlehlých" bodů. Z výše uvedené definice funkce $\Phi(x)$ plyne, že je-li n liché číslo (n=2m+1), pak $\tilde{x} = x_m$, je-li sudé číslo (n=2m), pak $\tilde{x} = (x_{m-1} + x_m)/2$.

Ořezaný průměr (trimmed mean) $\bar{x}_T(D, p)$ – robustní odhad polohy centra – je jistým kompromisem mezi aritmetickým průměrem a mediánem. Jako parametr se používá veličina p vyjádřená zpravidla v procentech (nejčastěji 10 %). Ze seřazený soubor dat odstraníme round(p/2) nejvyšších a stejný počet nejnižších hodnot a ze zbytku vypočteme aritmetický průměr. Pro p = 0 jde o prů-

měr, pro $p \Rightarrow 100\%$ o medián. U váhovaných veličin je definice ořezaného průměru poněkud vágní a proto se běžně nepoužívá.

Modus – je-li nejčetněji zastoupená hodnota (nebo hodnota s největší vahou) – bývá u diskrétních výsledků měření, nebo v určitých intervalech – nejpohodlněji ji lze odečíst z histogramu (viz 1.2)

3 Míry rozptýlení, distribuční funkce

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

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \qquad s^2 = \frac{1}{S_w}\sum_{i=1}^{n}(x_i - \bar{x})^2 w_i = \overline{x^2} - \bar{x}^2.$$

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

Robustní třídou měr rozptýlení je tzv. *střední velikost odchylky* (mean absolute deviation – MAD), respektive vážená střední velikost odchylky (weighted mean absolute deviation – WMAD), centrovaná k a, nejčastěji pak aritmetickému průměru nebo k mediánu:

$$mad(a) = \frac{1}{n}\sum_{i=1}^{n}|x_i - a| \qquad wmad(a) = \frac{1}{S_w}\sum_{i=1}^{n}|x_i - a|\, w_i.$$

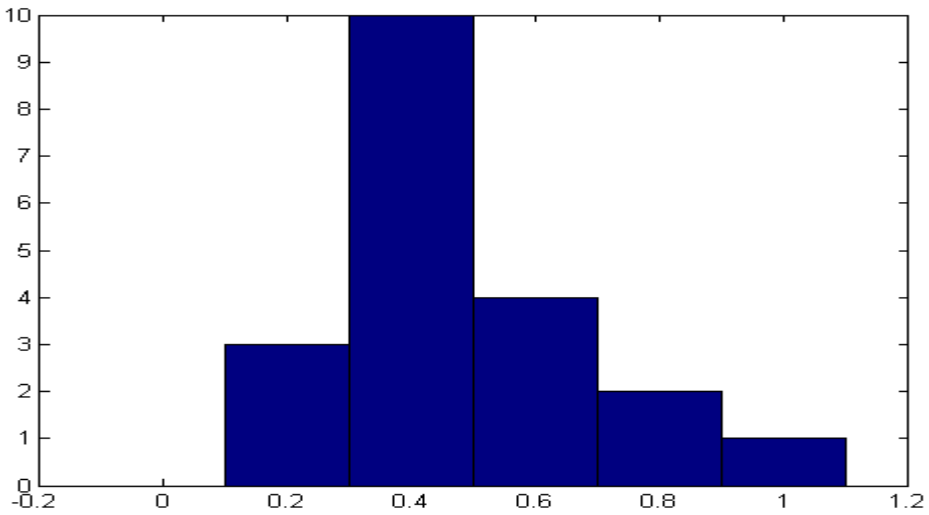
Lze ukázat, že pro $a = \tilde{x}$ je hodnota $mad(a)$, resp. $wmad(a)$, minimální.

Celkové rozpětí (total range) daný rozdílem mezi největším a nejmenší naměřenou hodnotou.

Mezikvartilní rozpětí (interquartile range), což je rozdíl mezi 3. a 1. kvartilem slouží jako robustní odhad rozptýlení, neboť se vztahuje na vnitřní část rozdělovací křivky.

Nejinstruktivnějším vyjádřením distribuční funkce je u diskrétních veličin tzv. tyčkový graf, v případě spojitých veličin 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*:

$$n_h = 1 + 3{,}3 \log n\,.$$



4 Normální rozdělení

Výjimečné postavení mezi rozdělovacími funkcemi má tzv. *normální rozdělovací funkce*, zvaná též Gaussova funkce, odpovídají rozdělení zcela náhodných veličin. Funkce hustoty pravděpodobnosti $f(x)$ je normovaná na 1 a je popsána dvojicí parametrů μ a σ:

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

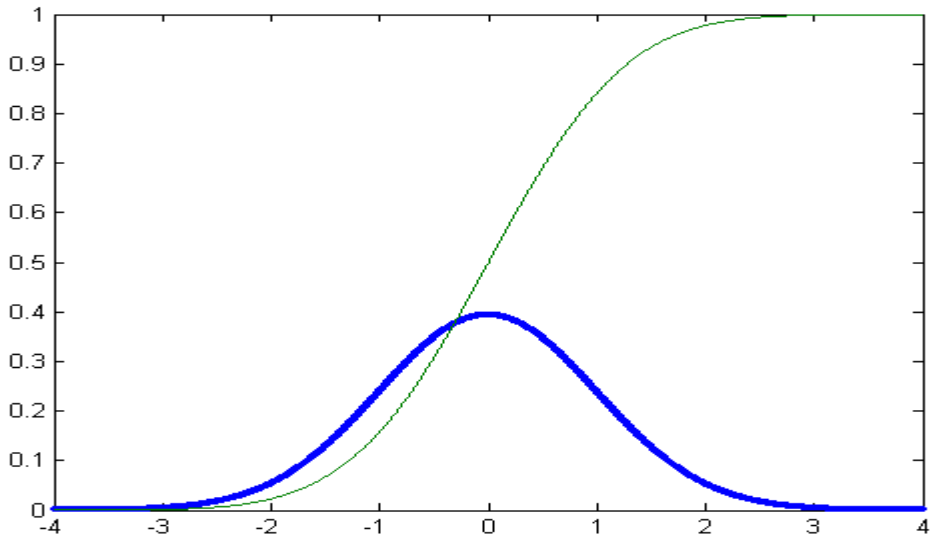
„Gaussův Říp“ je přísně symetrický podle osy $x = \mu$, kterážto hodnota je současně aritmetickým průměrem, mediánem i modem souboru podřizujícímu se normálnímu rozdělení. Lze ukázat, že směrodatná odchylka s je právě rovna parametru popisujícímu šířku normálního rozdělení σ (disperze), tedy:

$$s^2 = \overline{(x-\mu)^2} = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] dx = \sigma^2.$$

Kumulativní distribuční funkci lze s výhodou popsat pomocí speciální tabelované funkce erf(x) odpovídající Gaussovu rozdělení s $\mu = 0$ a $\sigma = 1/2$:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \Rightarrow \Phi\left(\frac{x-\mu}{\sigma} \right) = \frac{1}{2}\left[erf\left(\frac{x-\mu}{\sqrt{2}\,\sigma} \right) + 1 \right]$$

Několik charakteristik: v rozmezí $\pm\sigma$ se nachází 68% případů, $\pm 2\sigma$ 95%, $\pm 3\sigma$ 99,7%. 1. kvartil se nachází ve vzdálenosti 0.6745 σ od centra, mezikvartilní rozpětí tak odpovídá 1,349 σ. σ = 1/0.6745 *mad* = 1.483 *mad*.



4.1 Odhad μ a σ

K tomu, abychom dokonale mohli zjistit oba parametry normálního rozdělení σ a μ, bychom museli mít k dispozici nekonečně mnoho bodů. Ve skutečnosti máme k dispozici jen omezený vzorek celého souboru, a pomocí dat tohoto vzorku můžeme nanejvýš stanovit odhad obou parametrů, který

je zatížen jistou neurčitostí. Za předpokladu, že zkoumaný soubor má normální rozdělení, pak lze ukázat, že nejlepší nezávislý odhad parametru σ je dán vztahem:

$$\sigma_{\text{odh}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{n}{n-1}(\overline{x^2} - \bar{x}^2)}, \quad \sigma_{\text{odh}} = \sqrt{\frac{\sum (x_i - \bar{x})^2 w_i}{w_s (n-1)}} = \sqrt{\frac{n}{n-1}(\overline{x^2} - \bar{x}^2)}\,.$$

Pomocí tohoto odhadu střední kvadratické odchylky lze odhadnout i neurčitost stanovení parametru μ (vlastně aritmetického průměru):

$$\mu_{\text{odh}} = \bar{x};\ \delta(\mu_{\text{odh}}) = \frac{\sigma_{\text{odh}}}{\sqrt{n}} = \sqrt{\frac{(\overline{x^2} - \bar{x}^2)}{n-1}}\,.$$

4.2 Odchylky od normálního rozdělení, šikmost a špičatost

K popisu rozdělovací křivky se občas používá ještě jemnějšího popisu, který využívá

Obecný moment k-tého řádu (moment of *k*-th order):

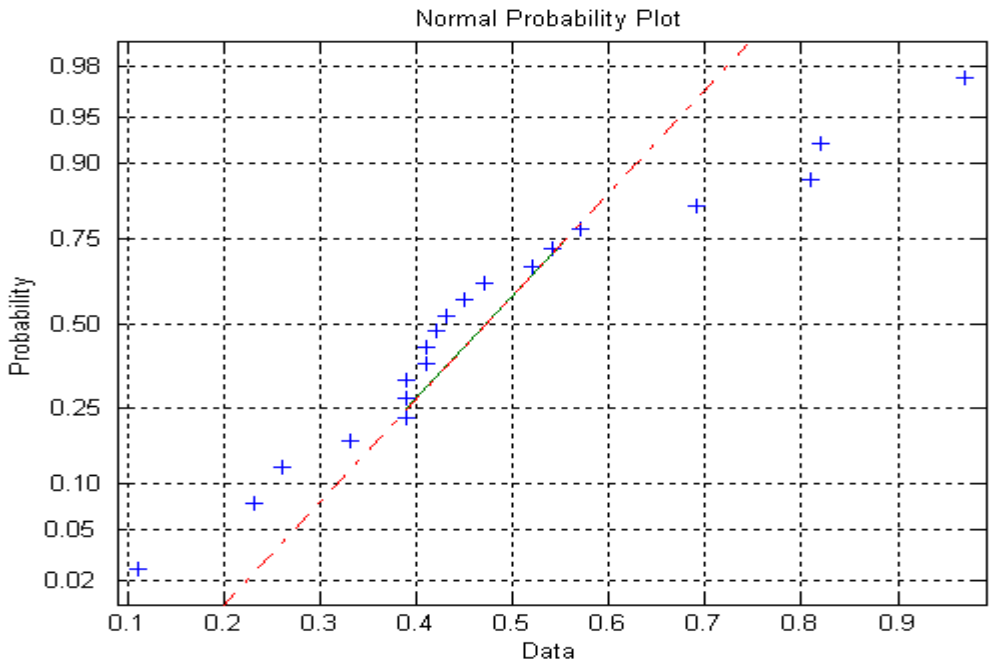
$$\overline{x^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k\,, \qquad \overline{x^k} = \frac{1}{S_w}\sum_{i=1}^{n} x_i^k w_i\,.$$

Obecný centrální moment k-tého řádu kolem bodu *a* (centred moment of *k*-th order):

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - a)^k\,, \qquad m_k = \frac{1}{S_w}\sum_{i=1}^{n} (x_i - a)^k w_i\,.$$

Centrem bývá nejčastěji aritmetický průměr, resp. váhovaný aritmetický průměr $\bar{x}$. Vidíme, že pro tento případ $m_0 = m_1 = 0$, $m_2 = s^2$.

Zavádíme teď ještě dvě bezrozměrné charakteristiky: tzv. *šikmost* (skewness): $a_3 = m_3/s^3$ a *špičatost* (kurtosis) $a_4 = m_4/s^4$ funkce. Šikmost symetrických funkcí je nulová (tedy i normálního rozdělení), charakteristiky tedy popisuje míru asymetrie funkce. Charakteristika a_4 přináší informaci o tom, jak se vlastně body koncentrují kolem průměru. Je-li a_4 blízké 3, pak mluvíme o souborech s normální špičatostí, při $a_4 < 3$, hovoříme o souborech plochých a při $a_4 > 3$ se mluví o souborech špičatých.



Za nejsdělnější nástroj k posouzení odchylek pozorovaného rozdělení od normálního rozdělení považuji *graf normální pravděpodobnosti* (normal probability plot), do něhož vynášíme kumulativní distribuční funkci, přičemž osa pravděpodobností (kvantilová) je transformována tak, aby se tam soubory s normálním rozdělením zobrazily jako přímky. Je vhodné si přitom body odpovídající 1. a 3. kvartilu proložit přímkou a diskutovat pak odchylky reálného rozložení bodů od ní. V Matlabu je pro tuto úlohu příkaz: *normplot*.

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):

0.82±0.07	0.39±0.03	0.54±0.05	0.57±0.03	0.42±0.04
0.39±0.07	0.69±0.05	0.81±0.05	0.33±0.05	0.41±0.04
0.11±0.07	0.23±0.04	0.39±0.04	0.43±0.04	0.97±0.03
0.26±0.05	0.47±0.04	0.41±0.05	0.52±0.04	0.45±0.03

Instrumentářem popisné statistiky charakterizujte tento soubor, speciálně pak uveďte:

a) počet měření a jejich charakter (spojité, diskrétní?)

b) stanovte 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.

c) 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)

d) minimální a maximální hodnotu extinkce a celkové rozpětí

e) rozptyl s^2, směrodatnou odchylku s, odhad rozptylu σ_{odh}, střední velikost odchylky s centrem v aritmetickém průměru a v mediánu

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

h) Vypočtěte šikmost a špičatost rozdělovací funkce a porovnejte s normálním rozdělením. Jaký je to typ souboru? Sestrojte graf normálního rozdělení a diskutujte (řešte bez vah).

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.

Instrumentářem popisné statistiky charakterizujte tento soubor, speciálně pak uveďte:

a) počet měření a jejich charakter (spojité, diskrétní?) – 20, spojité
b) stanovte 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. – není případné použití, standardní odchylka je mnohem větší, než nejistota jednoho měření
c) odhad aritmetického průměru a jeho nejistotu za předpokladu normálního rozdělení (*mean* = 0,480±0,047; *meanw* = 0,501±0,045), harmonický (0,382), geometrický (0,435), kvadratický průměr (0,552) a medián (0,425), ořezaný průměr pro 10% a 20% (jen pro případ bez vah: 0,474; 0,468)
d) minimální a maximální hodnotu extinkce a celkové rozpětí (0,11 až 0,97; 0,86)
e) rozptyl s^2, směrodatnou odchylku *s*, odhad rozptylu σ_{odh}, střední velikost odchylky s centrem v aritmetickém průměru a v mediánu (v aritmetickém průměru: 0,0417; 0,204; 0,0439; 0,210; se středem v mediánu: 0,448; 0,212; 0,0471; 0,217)
f) graf kumulativní distribuční funkce a pomocí ní stanovte hodnoty kvartilů a mezikvartilního rozpětí (interkv = 0,165)
g) Porovnejte odhady μ a σ pro normální rozdělení získané různými metodami; (σ_{odh} = 0,210; *mad* = 0,156, madmed = 0,146
h) Vypočtěte šikmost a špičatost rozdělovací funkce a porovnejte s normálním rozdělením. Jaký je to typ souboru? Sestrojte graf normálního rozdělení a diskutujte (řešte bez vah).
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.