Drsná matematika III – 7. týden
Závěrečné poznámky k diferenciálním rovnicím; přehled popisné statistiky

Jan Slovák

Masarykova univerzita
Fakulta informatiky

31. – 4. 11. 2016

Obsah přednášky

Plán přednášky

Kde je dobré číst?

- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice `www.math.muni.cz/Matematika_drsne_svizne`
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.
- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.

Plán přednášky

Rovnice vyšších řádů

Obyčejnou diferenciální rovnicí řádu k (vyřešenou vzhledem k nejvyšší derivaci) rozumíme rovnici

$$y^{(k)}(t) = f(t, y(t), y'(t), \ldots, y^{(k-1)}(t)),$$

kde f je známá funkce v $k+1$ proměnných, x je nezávisle proměnná a $y(x)$ je neznámá funkce v jedné proměnné. Ukážeme, že taková rovnice je vždy ekvivalentní systému k rovnic prvního řádu.

Rovnice vyšších řádů

Obyčejnou diferenciální rovnicí řádu k (vyřešenou vzhledem k nejvyšší derivaci) rozumíme rovnici

$$y^{(k)}(t) = f(t, y(t), y'(t), \ldots, y^{(k-1)}(t)),$$

kde f je známá funkce v $k+1$ proměnných, x je nezávisle proměnná a $y(x)$ je neznámá funkce v jedné proměnné. Ukážeme, že taková rovnice je vždy ekvivalentní systému k rovnic prvního řádu. Zavedeme nové neznámé funkce v proměnné t takto: $y_0(t) = y(t)$, $y_1(t) = y_0'(t)$, $\ldots$, $y_{k-1}(t) = y_{k-2}'(t)$.

Nyní je funkce $y(t)$ řešením naší původní rovnice tehdy a jen tehdy, když je první komponentou řešení systému rovnic

$$\begin{aligned} y_0'(t) &= y_1(t) \\ y_1'(t) &= y_2(t) \\ &\vdots \\ y_{n-2}'(t) &= y_{n-1}(t) \\ y_{n-1}'(t) &= f(t, y_0(t), y_1(t), \ldots, y_{n-1}(t)). \end{aligned}$$

Přímým důsledkem vět o systémech ODR 1. řádu je proto následující věta:

Theorem

Nechť funkce $f(t, y_0, \ldots, y_{k-1}) : U \subset \mathbb{R}^{k+1} \to \mathbb{R}$, *má spojité parciální derivace na otevřené množině* U. *Pak pro každý bod* $(t_0, z_0, \ldots, z_{k-1}) \in U$ *existuje maximální interval* $I_{max} = [x_0 - a, x_0 + b]$, *s kladnými* $a, b \in \mathbb{R}$, *a právě jedna funkce* $y(t) : \mathbb{I}_{max} \to \mathbb{R}$, *která je řešením rovnice k-tého řádu*

$$y^{(k)}(t) = f(t, y(t), y'(t), \ldots, y^{(k-1)}(t))$$

s podmínkou $y(t_0) = z_0$, $y'(t_0) = z_1$, $\ldots$, $y^{(k-1)}(t_0) = z_{k-1}$. *Toto řešení navíc závisí diferencovatelně na počáteční podmínce a případných dalších parametrech vstupujících diferencovatelně do funkce* f.

Vidíme tedy, že pro jednoznačné zadání řešení obyčejné diferenciální rovnice k–tého řádu musíme zadat v jednom bodě hodnotu a prvních $k - 1$ derivací výsledné funkce. Obdobně lze diskutovat systémy rovnic libovolných řádů.

Operace derivování je lineární zobrazení z (dostatečně) hladkých funkcí do funkcí. Pokud derivace $(\frac{d}{dx})^j$ jednotlivých řádů j vynásobíme pevnými funkcemi $a_j(x)$ a výrazy sečteme, dostaneme tzv. *lineární diferenciální operátor*:

$$y(x) \mapsto D(y)(x) = a_k(x)y^{(k)}(x) + \cdots + a_1(x)y'(x) + a_0y(x).$$

Operace derivování je lineární zobrazení z (dostatečně) hladkých funkcí do funkcí. Pokud derivace $(\frac{d}{dx})^j$ jednotlivých řádů j vynásobíme pevnými funkcemi $a_j(x)$ a výrazy sečteme, dostaneme tzv. *lineární diferenciální operátor*:

$$y(x) \mapsto D(y)(x) = a_k(x)y^{(k)}(x) + \cdots + a_1(x)y'(x) + a_0y(x).$$

Řešit příslušnou *homogenní lineární diferenciální rovnici* pak znamená najít funkci y splňující $D(y) = 0$, tj. obrazem je identicky nulová funkce.
Ze samotné definice je zřejmé, že součet dvou řešení bude opět řešením, protože pro libovolné funkce y_1 a y_2 platí

$$D(y_1 + y_2)(x) = D(y_1)(x) + D(y_2)(x).$$

Obdobně je také konstantní násobek řešení opět řešením. Celá mmnožina všech řešení lineární diferenciální rovnice k-tého řádu je tedy vektorovým prostorem.

Přímou aplikací předchozí věty o jednoznačnosti a existenci řešení rovnic dostáváme:

Corollary

Vektorový prostor všech řešení homogenní lineární diferenciální rovnice k–tého řádu je vždy dimenze k. Proto můžeme vždy řešení zadat jako lineární kombinaci libovolné množiny k lineárně nezávislých řešení. Taková řešení jsou zadána jednoznačně lineárně nezávislými počátečními podmínkami na hodnotu funkce $y(x)$ jejích prvních $(k-1)$ derivací.

Připomeňme homogenní lineární diferenční rovnice.
Analogie jde i dále v okamžiku, kdy jsou všechny koeficienty a_j diferenciálního operátoru D konstantní. Už jsme viděli u takové rovnice prvního řádu, že řešením je exponenciála s vhodnou konstantou u argumentu. Stejně jako u diferenčních rovnic se podbízí vyzkoušet, zda takový tvar řešení $y(x) = \mathrm{e}^{\lambda x}$ s neznámým parametrem λ může splnit rovnici k–tého řádu. Dosazením dostaneme

$$D(\mathrm{e}^{\lambda x}) = \left(a_k\lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0\right)\mathrm{e}^{\lambda x}\,.$$

Parametr λ tedy vede na řešení lineární diferenciální rovnice s konstantními koeficienty tehdy a jen tehdy, když je λ kořenem tzv. *charakteristického polynomu* $a_k\lambda^k + \cdots + a_1\lambda + a_0$.

Pokud má charakteristický polynom k různých kořenů, dostáváme bázi celého vektorového prostoru řešení. Pokud je λ násobný kořen, přímým výpočtem s využitím toho, že je pak také kořenem derivace charakteristického polynomu, dostaneme, že je řešením i funkce $x\,\mathrm{e}^{\lambda x}$. Podobně pak pro vyšší násobnost ℓ dostáváme ℓ různých řešení $\mathrm{e}^{\lambda}, x\,\mathrm{e}^{\lambda x}, \ldots, x^{\ell-1}\,\mathrm{e}^{\lambda x}$.

U obecné lineární diferenciální rovnice předepisujeme nenulovou hodnotu diferenciálního operátoru D. Opět úplně analogicky k úvahám o systémech lineárních rovnic nebo u lineárních diferenčních rovnic přímo vidíme, že obecné řešení takovéto (nehomogenní) rovnice

$$D(y)(x) = b(x)$$

pro nějakou pevně zadanou funkci $b(x)$ je součtem jednoho jakéhokoliv řešení této rovnice a množiny všech možných řešení příslušné homogenní rovnice $D(y)(x) = 0$. Celý prostor řešení je tedy opět pěkný konečněrozměrný afinní prostor, byť ukrytý v obrovském prostoru funkcí.

Plán přednášky

V praxi se setkáváme s postupy, jak přibližně spočíst řešení rovnice, se kterou pracujeme (protože exaktní řešení jsou vzácná).
Už jsme podobné úvahy dělali všude tam, kde jsme se zabývali aproximacemi (tj. zejména lze doporučit porovnání s dřivějšími úvahami o splajnech, Taylorových polynomech a Fourierových řadách).
S trochou odvahy můžeme také považovat diferenční a diferenciální rovnice za vzájemné aproximace. V jednom směru nahrazujeme diference diferenciály (např. u ekonomických nebo populačních modelů), ve druhém pak naopak.
Zastavíme se na chvilku u nahrazování derivací diferencemi.
Nejdříve si však připomeneme obvyklé značení pro zápis odhadů chyb.

Definition

Pro funkci $f(x)$ v proměnné x řekneme, že je v okolí hromadného bodu x_0 svého definičního oboru **řádu velikosti** $O(\varphi(x))$ pro nějakou funkci $\varphi(x)$, jestliže existuje okolí U bodu x_0 a konstanta C taková, že

$$|f(x)| \leq C \cdot |\varphi(x)|$$

pro všechny $x \in U$. Limitní bod x_0 bývá často i nevlastní hodnota $\pm\infty$.

Nejobvyklejší příklady jsou $O(x^p)$ pro **polynomiální řád velikosti** a to v nule nebo v nekonečnu, $O(\ln x)$ pro **logaritmický řád velikosti** v nekonečnu atd. Všimněme si, že logaritmický řád velikosti nezávisí na volbě základu.
Dobrým příkladem je aproximace funkce jejím Taylorovým polynomem řádu k v bodě x_0. Taylorova věta pro funkce jedné proměnné říká, že chyba této aproximace je $O(h^{k+1})$, kde h je přírůstek argumentu $x - x_0 = h$.
Podobné úvahy jsme dělali i u Fourierových řad.

V případě obyčejných diferenciálních rovnic je nejjednodušším schématem aproximace tzv. Eulerovými polygony. Budeme ji prezentovat pro jednu obyčejnou rovnici s jednou nezávislou a jednou závislou veličinou. Úplně stejně ale funguje pro systémy rovnic, když skalární veličiny a jejich derivace v čase t nahradíme vektory závislé na času a jejich derivacemi.
Uvažujme tedy opět rovnici (pro jednoduchost a bez újmy na obecnosti prvního řádu)

$$y'(t) = f(t, y(t)).$$

Označme si diskrétní přírůstek času h, tj. $t_n = t_0 + nh$, a $y_n = y(t_n)$. Z Taylorovy věty (se zbytkem druhého řádu) a naší rovnice vyplývá, že

$$y_{n+1} = y_n + y'(t_n)h + O(h^2) = y_n + f(t_n, y_n)h + O(h^2).$$

Jestliže tedy od t_0 do t_n uděláme n takových kroků o přírůstek h, bude očekávaný odhad celkové chyby vyplývající z lokálních nepřesností naší lineární aproximace nejvýše $hO(h^2)$, tj. chyba bude v řádu velikosti $O(h)$. Ve skutečnosti vstupují při výpočtu do hry ještě zaokrouhlovací chyby.
Při numerickém řešení Eulerovou metodou postupujeme tak, že za přibližné řešení považujeme po částech lineární polygon definovaný výše.

Plán přednášky

Statistika v širším slova smyslu = **jakékoliv zpracování číselných dat o nějakém souboru objektů a jejich (více či méně přehledná) prezentace.**

Statistika v širším slova smyslu = **jakékoliv zpracování číselných dat o nějakém souboru objektů a jejich (více či méně přehledná) prezentace.**
Podstatou **matematické statistiky** je pro daná data zjišťovat:

- vlastnosti objektů
- věrohodnost odvozených výsledků.

Statistika v širším slova smyslu = **jakékoliv zpracování číselných dat o nějakém souboru objektů a jejich (více či méně přehledná) prezentace.**
Podstatou **matematické statistiky** je pro daná data zjišťovat:

- vlastnosti objektů
- věrohodnost odvozených výsledků.

Zpravidla jde o data (cíleně nebo náhodně vybrané) části souboru objektů, jejich následnou analýzu a konečně o vyslovení důsledků pozorování pro celý soubor.

Statistika v širším slova smyslu = **jakékoliv zpracování číselných dat o nějakém souboru objektů a jejich (více či méně přehledná) prezentace.**
Podstatou **matematické statistiky** je pro daná data zjišťovat:

- vlastnosti objektů
- věrohodnost odvozených výsledků.

Zpravidla jde o data (cíleně nebo náhodně vybrané) části souboru objektů, jejich následnou analýzu a konečně o vyslovení důsledků pozorování pro celý soubor.
Teorie pravděpodobnosti studuje modely popisující chování abstraktních souborů prostřednictvím **pravděpodobnosti jevů z jevového pole**, matematická statistika studuje skutečné náhodné výběry z nějakého základního souboru a zdůvodňuje **výběr teoretického pravděpodobnostního modelu a kvalitativní informace o jeho parametrech.**

Example

Za soubor objektů vezměme všechny studenty této přednášky, jako číselný údaj můžeme uvažovat

1. „průměrný počet bodů“ dosažený při hodnocení tohoto předmětu v poslední písemce,
2. průměrnou známku dosaženou u zkoušky z tohoto a z jiných pevně vybraných předmětů,
3. číslená data vypovídající o historii dřívějšího studia,
4. počet pracovních hodin týdně odpracovaných mimo fakultu.

Example

Za soubor objektů vezměme všechny studenty této přednášky, jako číselný údaj můžeme uvažovat

1. „průměrný počet bodů“ dosažený při hodnocení tohoto předmětu v poslední písemce,
2. průměrnou známku dosaženou u zkoušky z tohoto a z jiných pevně vybraných předmětů,
3. číslená data vypovídající o historii dřívějšího studia,
4. počet pracovních hodin týdně odpracovaných mimo fakultu.

Samotný aritmetický průměr bodů nám mnoho neřekne ani o kvalitě přednášky ani o kvalitě přednášejícího ani o samotném hodnocení. Zajímá nás např. hodnota, která bude „uprostřed souboru“, tj. počet bodů, pro které je stejně studentů pod ní a nad ní.

Example

Za soubor objektů vezměme všechny studenty této přednášky, jako číselný údaj můžeme uvažovat

1. „průměrný počet bodů“ dosažený při hodnocení tohoto předmětu v poslední písemce,
2. průměrnou známku dosaženou u zkoušky z tohoto a z jiných pevně vybraných předmětů,
3. číslená data vypovídající o historii dřívějšího studia,
4. počet pracovních hodin týdně odpracovaných mimo fakultu.

Samotný aritmetický průměr bodů nám mnoho neřekne ani o kvalitě přednášky ani o kvalitě přednášejícího ani o samotném hodnocení. Zajímá nás např. hodnota, která bude „uprostřed souboru“, tj. počet bodů, pro které je stejně studentů pod ní a nad ní. Obdobně první a poslední čtvrtina, desetina apod. Všem takovým údajům říkáme **statistiky** posuzované veličiny. V uvedených příkladech se jim říká **medián**, **kvartil**, **decil** apod.

Z obecné zkušenosti nebo jako výsledek úvah mimo matematiku víme, jakou „strukturu“ by měla mít sledovaná data. Např. víme, že rozumné hodnocení studentů by mělo mít tzv. **normální rozdělení**. Tento pojem patří do teorie pravděpodobnosti.

Z obecné zkušenosti nebo jako výsledek úvah mimo matematiku víme, jakou „strukturu“ by měla mít sledovaná data. Např. víme, že rozumné hodnocení studentů by mělo mít tzv. **normální rozdělení**. Tento pojem patří do teorie pravděpodobnosti.
Pokud je naše představa oprávněná, pak porovnáním výsledku třeba i docela malého náhodného výběru studentů s teoretickým modelem můžeme zjistit odhad parametrů takového rozdělení a činit závěry, zda je hodnocení „skutečně rozumné“. Zároveň budeme umět popsat věrohodnost našich závěrů.

Daleko zajímavější vývody ovšem můžeme činit, když porovnáním statistik pro různé veličiny budeme moci dovozovat informace o souvislostech. Pokud např. neexistuje žádná doložitelná souvislost mezi historií předchozího studia a výsledky v dané přednášce, je jedním z možných vysvětlení vývod, že je přednáška prostě špatná.

Daleko zajímavější vývody ovšem můžeme činit, když porovnáním statistik pro různé veličiny budeme moci dovozovat informace o souvislostech. Pokud např. neexistuje žádná doložitelná souvislost mezi historií předchozího studia a výsledky v dané přednášce, je jedním z možných vysvětlení vývod, že je přednáška prostě špatná.

Závěr úvodních úvah:

- V matematice pracujeme s abstraktním matematickým popisem pravděpodobnosti.
- Vývody pro konktrétní soubory dat, pro které je zvolený model relevantní dává matematická statistika.
- Názor, zda je takový popis adekvátní pro konkrétní výběr dat, je také možné podpořit nebo zavrhnout pomocí metod matematické statistiky.

Plán přednášky

Popisná statistika není matematická disciplína ...
Jde o dlouho řadu zvyklostí/postupů, jak zpracovávat a prezentovat data, a názvů pro jednotlivé typy sestav dat.

Popisná statistika není matematická disciplína ...
Jde o dlouho řadu zvyklostí/postupů, jak zpracovávat a prezentovat data, a názvů pro jednotlivé typy sestav dat.
Zpravidla pracujeme se **statistickým souborem**, který je sestaven ze **statistických jednotek**. Na statistických jednotkách se pak měří (zjišťují) jednotlivé **statistické znaky**.

Popisná statistika není matematická disciplína ...
Jde o dlouho řadu zvyklostí/postupů, jak zpracovávat a prezentovat data, a názvů pro jednotlivé typy sestav dat.
Zpravidla pracujeme se **statistickým souborem**, který je sestaven ze **statistických jednotek**. Na statistických jednotkách se pak měří (zjišťují) jednotlivé **statistické znaky**.
Např. souborem mohou být všichni studenti MU, každý zvlášť je pak **statistickou jednotkou**. O těchto jednotkách pak můžeme schraňovat mnoho znaků – např. všechny číselné hodnoty zjistitelné z ISu, jakou mají nejraději barvu, co snědli večer před poslední písemkou, atd.

Popisná statistika není matematická disciplína ...
Jde o dlouho řadu zvyklostí/postupů, jak zpracovávat a prezentovat data, a názvů pro jednotlivé typy sestav dat.
Zpravidla pracujeme se **statistickým souborem**, který je sestaven ze **statistických jednotek**. Na statistických jednotkách se pak měří (zjišťují) jednotlivé **statistické znaky**.
Např. souborem mohou být všichni studenti MU, každý zvlášť je pak **statistickou jednotkou**. O těchto jednotkách pak můžeme schraňovat mnoho znaků – např. všechny číselné hodnoty zjistitelné z ISu, jakou mají nejraději barvu, co snědli večer před poslední písemkou, atd.
Základním objektem pro zkoumání jednotlivých znaků je pak **soubor hodnot**. Zpravidla jej máme ve formě uspořádaných hodnot. Uspořádání je buď dáno přirozeně (když jsou hodnotami např. reálná čísla) nebo je můžeme zavést pro určitost (třeba když budeme sledovat barvy, tak je můžeme vyjdřovat v RGB standardu a řadit podle tohoto příznaku).

Statistický popis chce srozumitelně a přehledně sdělit něco o celém souboru. Musíme proto umět jednotlivé hodnoty nějak porovnovávat a poměřovat. Potřebujeme tedy nějaké **měřítko**. Podle toho jakého charakteru jsou hodnoty, hovoříme o měřítku:

- **nominálním** (mezi hodnotami není žádný vztah, jde pouze o četnosti možných hodnot, např. politická strana v ČR nebo učitelé MU při zkoumání obliby);
- **ordinální** (totéž jako předchozí, ale s přidaným uspořádáním, např. počet hvězdiček u hotelu v bedekrech);
- **intervalové** (jde o číselné hodnoty, ale jde o porovnání velikostí, nikoliv absolutní hodnotu, např. u měření teplot je poloha nuly dohodnuta, ale není podstatná);
- **poměrové** (máme pevně stanovené měřítko a nulu, např. většina fyzikálních veličin).

V dalším budeme pracovat se **souborem hodnot** $x_1, x_2, \ldots, x_n$ (které vznikly měřením na n statistických jednotkách) a uspořádáme je do **uspořádaného souboru hodnot**

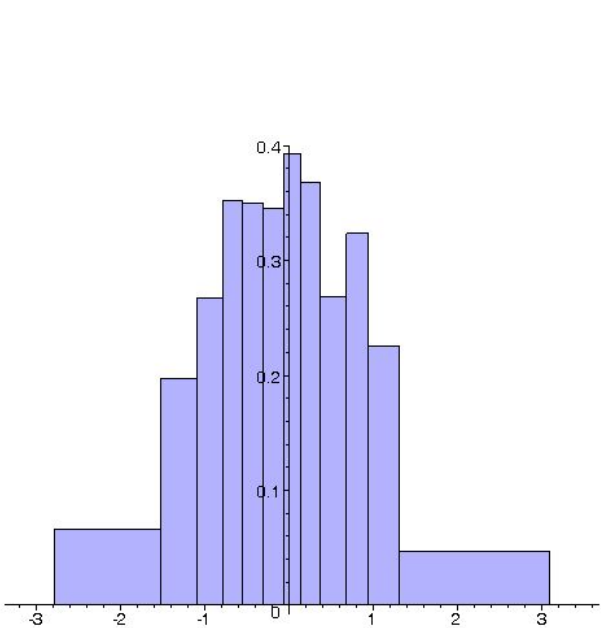
$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

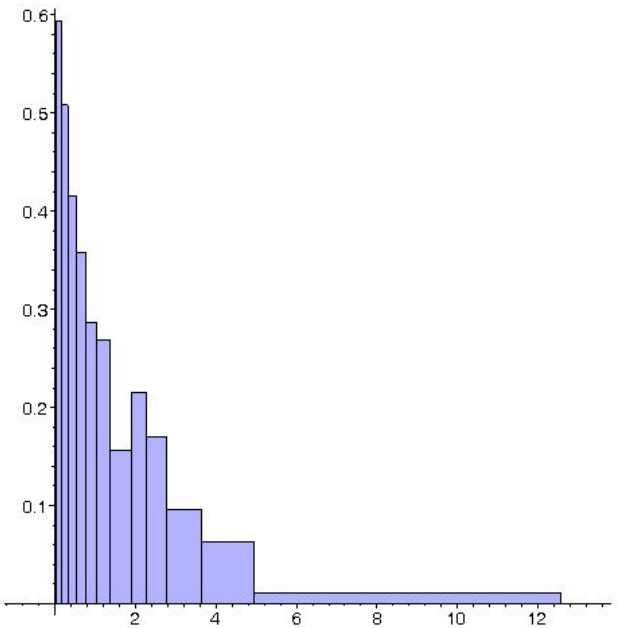
Číslo n nazýváme **rozsah souboru**.
Nejjednodušší je u rozsáhlých souborů znaků, které ale připouští jen málo hodnot uvádět pouze četnosti. Např. při průzkumu preferencí politických stran nebo u prezentace kvality hotelové sítě uvádíme u každé možné hodnoty počet jejích výskytů.

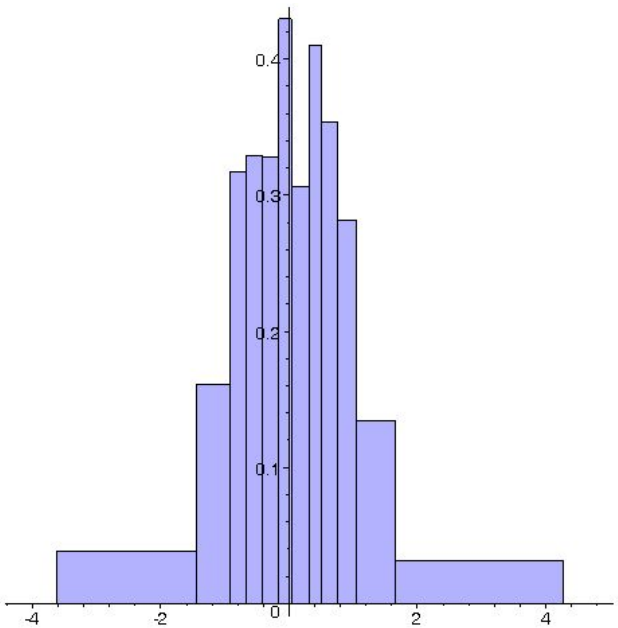
Pokud je i možných hodnot více (nebo dokonce připouštíme kontinuální reálné hodnoty), dělíme často možný rozsah hodnot na vhodný počet intervalů a o statistickém znaku uvádíme četnost hodnot v daných intervalech. Intervalům se často říká **třídy** a počtu znaku ve třídě pak **třídní četnost**.
Používáme také **kumulativní třídní četnosti**, které vznikají prostým součtem třídních četností s hodnotami nejvýše jako má daná třída.

Pokud je i možných hodnot více (nebo dokonce připouštíme kontinuální reálné hodnoty), dělíme často možný rozsah hodnot na vhodný počet intervalů a o statistickém znaku uvádíme četnost hodnot v daných intervalech. Intervalům se často říká **třídy** a počtu znaku ve třídě pak **třídní četnost**.
Používáme také **kumulativní třídní četnosti**, které vznikají prostým součtem třídních četností s hodnotami nejvýše jako má daná třída.
Nejčastěji pak uvažujeme střed a_i dané třídy za hodnotu, která ji reprezentuje a hodnota $a_i n_i$, kde n_i je četnost výskytu této třídy představuje celkový příspěvek této třídy. Velmi často také místo četností zobrazujeme relativní četnosti a_i/n, resp. relativní kumulativní četnosti.

Graf, který na jedné ose vynáší intervaly jednotlivých tříd a nad nimi obdélníky s výškou rovnou četnosti se nazývá **histogram**.
Obdobně se znázorňuje kumulativní četnost.
Na obrázku jsou histogramy souborů o rozsahu $n = 500$, které vznikly náhodným generováním dat s rozdělením normálním, χ^2 a studentovým







Míry polohy statistických znaků

Chceme-li velikost hodnot, kolem kterých se jednotlivá pozorování znaků shromažďují používáme většinou následující:

Definition

Nechť $(x_1, \ldots, x_n)$ je soubor hodnot měřeného znaku.

- **Průměr** (nebo také výběrový průměr) je dán

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{m} n_j a_j;$$

- **Geometrický průměr** je dán

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

a má smysl pouze u kladných hodnot znaků.

Definition (pokračování ...)

- Harmonický průměr je dán

$$\bar{x}_H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

a je také definován jen pro kladné hodnoty znaků.

Definition (pokračování ...)

- Harmonický průměr je dán

$$\bar{x}_H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

a je také definován jen pro kladné hodnoty znaků.

Výběrový průměr je jediný invariantní vůči afinním transormacím, tj. pro libovolné skaláry a, b platí $\overline{(a + b \cdot x)} = a + b \cdot \bar{x}$. Ostatní průměry jsou proto nevhodné pro intervalová měřítka.

Definition (pokračování ...)

- Harmonický průměr je dán

$$\bar{x}_H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

a je také definován jen pro kladné hodnoty znaků.

Výběrový průměr je jediný invariantní vůči afinním transormacím, tj. pro libovolné skaláry a, b platí $\overline{(a + b \cdot x)} = a + b \cdot \bar{x}$. Ostatní průměry jsou proto nevhodné pro intervalová měřítka.
Logaritmus geometrického průměru je obyčejný průměr logaritmů znaků. Je obzvlášť vhodný pro znaky, které se kumulují multiplikativně, např. úrokové míry. Je-li totiž úroková míra v jednotlivých časových jednotkách $x_i\%$, bude za celé období výsledek takový, jakoby byla konstatní úroková míra $\bar{x}\%$.

Definition (pokračování ...)

- Harmonický průměr je dán

$$\bar{x}_H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

a je také definován jen pro kladné hodnoty znaků.

Výběrový průměr je jediný invariantní vůči afinním transormacím, tj. pro libovolné skaláry a, b platí $\overline{(a + b \cdot x)} = a + b \cdot \bar{x}$. Ostatní průměry jsou proto nevhodné pro intervalová měřítka.
Logaritmus geometrického průměru je obyčejný průměr logaritmů znaků. Je obzvlášť vhodný pro znaky, které se kumulují multiplikativně, např. úrokové míry. Je-li totiž úroková míra v jednotlivých časových jednotkách $x_i\%$, bude za celé období výsledek takový, jakoby byla konstatní úroková míra $\bar{x}\%$.
Platí $\bar{x}_H \leq \bar{x}_G \leq \bar{x}$.

Medián, kvartil, decil, percentil, ...

Jiný způsob vyjádření míry, jakou hodnotu nabývají znaky je najít pro číslo α mezi nulou a jedničkou takovou hodnotu x_α, aby $100\alpha\%$ hodnot znaku bylo nejvýše x_α a zbylé byly alespoň x_α. Pokud takový znak není určen jednoznačně, volíme zpravidla průměr mezi dvěmi možnými hodnotami. Nejobvyklejší jsou:

- **medián** (často také výběrový medián) definovaný vztahem $\tilde{x} = x_{(\frac{n+1}{2})}$ pro liché n a $\tilde{x} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$;
- **dolní a horní kvartil** $Q_1 = x_{0,25}$ a $Q_3 = x_{0,75}$;
- **p-tý kvantil** (též výběrový kvantil nebo percentil) x_p, kde $0 < p < 1$ (zpravidla zadaný na dvě desetinná místa).

Lze se setkat také s hodnotou **modus**, která udává hodnotu znaku s největší četností.

Míry variability statistických znaků

Rozumným požadavkem na jakoukoliv míru variability je její invariance vůči konstantním posunutím.

Definition

- **Rozptyl** souboru znaků x je definován vztahem

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 = \frac{1}{n}\sum_{j=1}^{m} n_j(a_j - \bar{x})^2$$

  případně v jmenovateli zlomku používáme $(n-1)$.
- **Směrodatná odchylka** je dána jako odmocnina z výběrového rozptylu.
- **Rozpětí výběru** je $R = x_{(n)} - x_{(1)}$, **kvartilové rozpětí** je $Q = Q_3 - Q_1$.

Rozptyl

je „zprůměrovaný kvadrát“ standardní euklidovské vzdálenosti vektoru výběrových hodnot od jejich střední hodnoty. Díky této definici se chová velice přirozeně a budeme se s ním často potkávat. Používá se také tzv. **průměrná odchylka**

$$d_x = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|.$$

Všimněme si, že tady jde o skutečný průměr vzdáleností hodnot znaků, ovšem od mediánu!

Rozptyl

je „zprůměrovaný kvadrát“ standardní euklidovské vzdálenosti vektoru výběrových hodnot od jejich střední hodnoty. Díky této definici se chová velice přirozeně a budeme se s ním často potkávat. Používá se také tzv. **průměrná odchylka**

$$d_x = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}|.$$

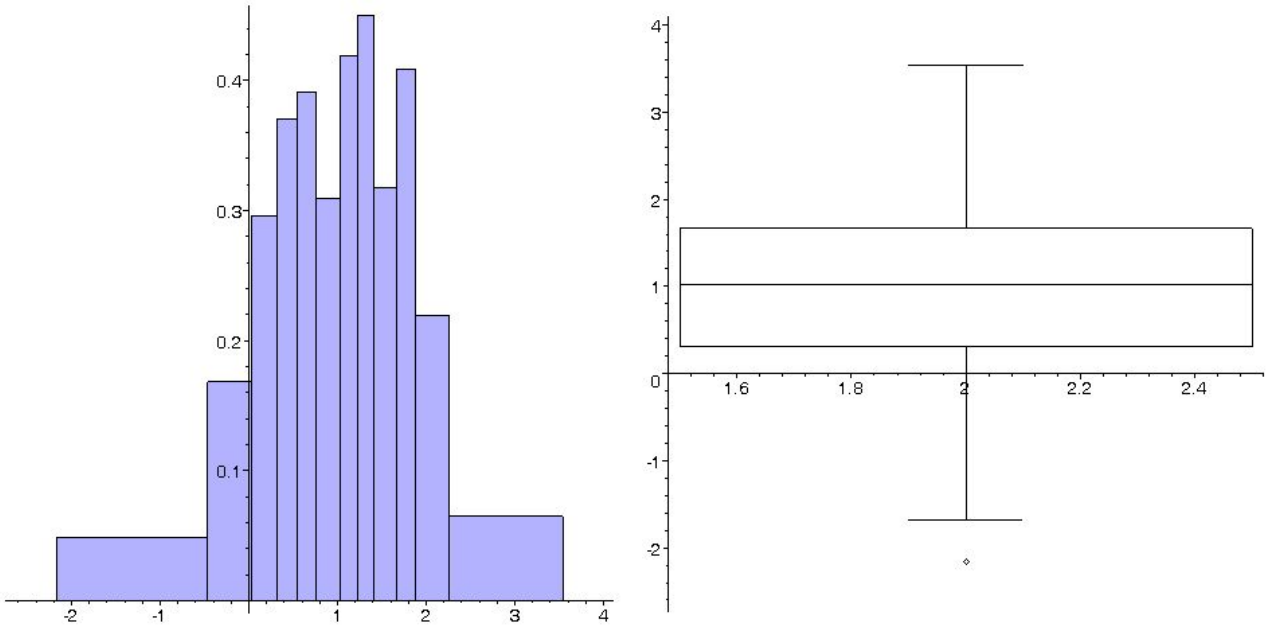
Všimněme si, že tady jde o skutečný průměr vzdáleností hodnot znaků, ovšem od mediánu!
Následující věta říká, proč zrovna tyto míry volíme:

Theorem

- *Funkce* $S(t) = (1/n)\sum_{i=1}^{n}(x_i - t)^2$ *nabývá svého minima pro* $t = \bar{x}$, *tj. pro výběrový průměr.*
- *Funkce* $D(t) = (1/n)\sum_{i=1}^{n}|x_i - t|$ *nabývá svého minima pro* $t = \tilde{x}$, *tj. pro medián.*

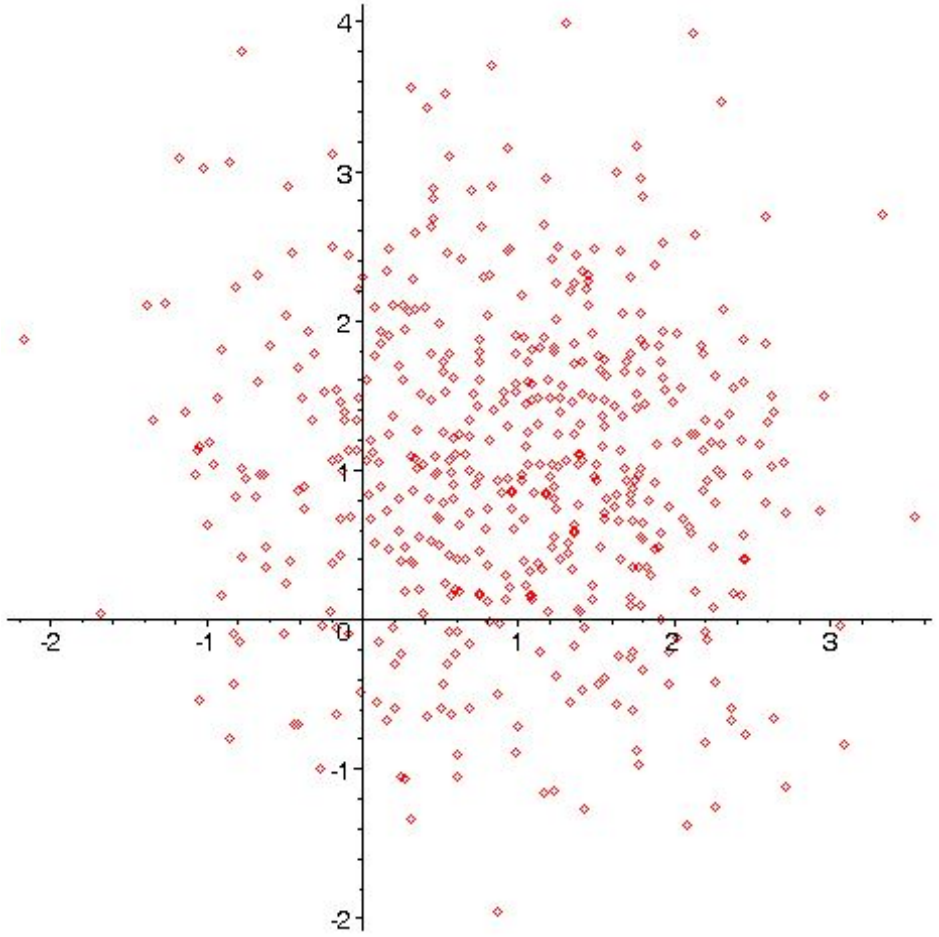
Diagramy

Pro rychlé vstřebávání složitěji strukturovaných informací je člověk skvěle vybaven zrakově. Proto se pro zobrazení statistiky jednotlivých znaků nebo jejich korelací používá mnoho standardizovaných nástrojů. Jedním z nich jsou tzv. **krabicové diagramy**.



Střední linka je medián, kraje boxu jsou kvartily, "packy"ukazují 1,5 kvartilového rozsahu, ne však víc než kraje rozsahu výběru, případné hodnoty mimo jsou přímo naznačeny body.

Běžné zobrazovací nástroje nám umožnějí dobře vidět případné závislosti dvou výběrů zjištěných znaků. Např. na obrázku jsou za souřadnice voleny hodnoty ze dvou nezávislých výběrů z normálních rozdělení se střední hodnotou 1 a rozptylem 1.



Entropie

Variabilitu chceme postihnout i u nominálních typů znaků. K dispozici máme jen třídní četnosti a můžeme tedy relativní četnost i-té třídy, $p_i = \frac{n_i}{n}$, vnímat jako pravděpodobnost, že náhodně vybraný prvek bude v této třídě.
Podbízí se pro datový soubor x definovat

$$H_X = \sum_{i=1}^{n} p_i F(p_i),$$

kde F je zatím neznámá funkce.
Je-li $p_k = 1$a ostaní $p_j = 0$, pak je variabilita je nulová. chceme proto $F(1) = 0$.

Celkem přirozeně chceme pro soubor znaků Z tvořený dvojicemi znaků ze souborů X a Y (např. můžeme na statistických jednotkách-osobách sledovat barvu očí a barvu vlasů), aby variabilita znaků z byla součtem variabilit jednotlivých znaků, tj. požadujeme $H_Z = H_X + H_Y$.

Celkem přirozeně chceme pro soubor znaků Z tvořený dvojicemi znaků ze souborů X a Y (např. můžeme na statistických jednotkách-osobách sledovat barvu očí a barvu vlasů), aby variabilita znaků z byla součtem variabilit jednotlivých znaků, tj. požadujeme $H_Z = H_X + H_Y$.
Známe relativní třídní četnosti p_i pro znaky v souboru X a q_j pro znaky souboru Y. Relativní třídní četnosti pro Z jsou

$$r_{ij} = \frac{n_i m_j}{nm} = p_i q_j$$

a požadujeme tedy rovnost

$$\sum_{i,j} p_i q_j F(p_i q_j) = \sum_i p_i F(p_i) + \sum_j q_j F(q_j).$$

Díky tomu, že p_i a q_j jsou relativní četnosti a tedy dávají v součtu 1, můžeme pravou stranu rovnosti přepsat jako

$$\left(\sum_j q_j\right)\left(\sum_i p_i F(p_i)\right) + \left(\sum_i p_i\right)\left(\sum_j q_j F(q_j)\right).$$

$$\sum_{i,j} p_i q_j F(p_i q_j) = \sum_{i,j} p_i q_j \big(F(p_i) + F(q_j)\big).$$

Tomuto požadavku vyhovuje jakýkoliv konstantní násobek logaritmu při kterémkoliv pevně zvoleném základu $a > 1$ (a lze ukázat, že jiná spojitá řešení F neexistují).

Poněvadž je $p_i \leq 1$, je jistě $\ln p_i \leq 0$. My však chceme variabilitu nezápornou, zvolíme proto za funkci F logaritmickou funkci s násobkem -1. Taková volba také automaticky splňuje náš požadavek $F(1) = 0$.

Definition (Entropie)

Míru variability znaků v nominálním měřítku vyjadřujeme pomocí **entropie**. Je dána vztahem

$$H_X = -\sum_{i=1}^{k} \frac{n_i}{n} \ln\left(\frac{n_i}{n}\right),$$

kde k je počet tříd ve výběru. Kromě přirozeného logaritmu se často také setkáváme (např. teorii informace) se stejným vztahem ale s logaritmem při základu 2.

Často se také místo H_X pracuje s veličinou

$$e^{H_X} = \prod_i p_i^{-p_i},$$

případně totéž s jiným zvoleným základem pro logaritmus.
Pro výběr X s k stejně velkými třídními četnostmi je
$e^{H_X} = \left(\left(\frac{1}{k}\right)^{-\frac{1}{k}}\right)^k = k$, nezávisle na velikosti výběru.