Matematika III – 8. týden
Jak na statistiku?

Jan Slovák

Masarykova univerzita
Fakulta informatiky

9. 11. – 13. 11. 2015

Obsah přednášky

Plán přednášky

Kde je dobré číst?

- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- 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`
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, Teorie pravděpodobnosti a matematická statistika (sbírka příkladů), Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- 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.

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.

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**

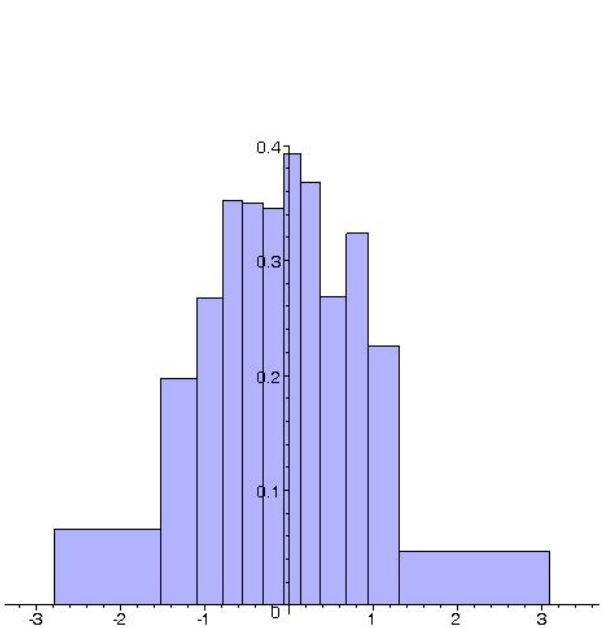
$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

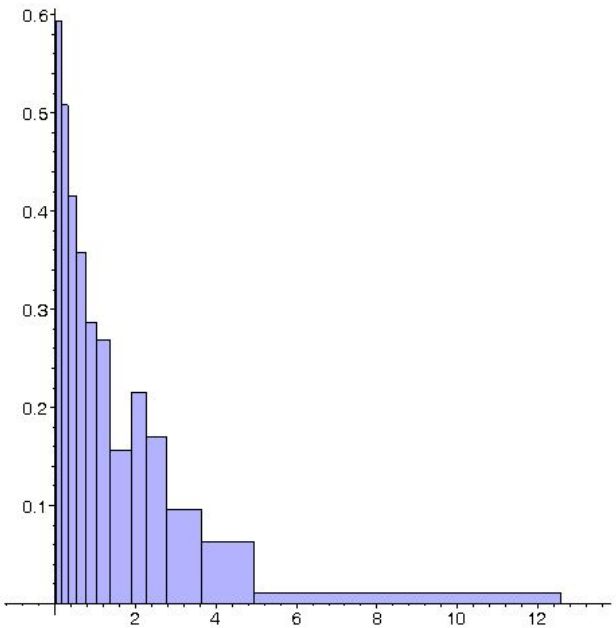
Číslo n nazýváme **rozsah souboru**.
Nejjednodušší je u rozsáhlých souborů znaků, které ale připouští jen málo hodnot uvádět pouze četnosti. Např. při průzkumu preferencí politických stran nebo u prezentace kvality hotelové sítě uvádíme u každé možné hodnoty počet jejích výskytů.

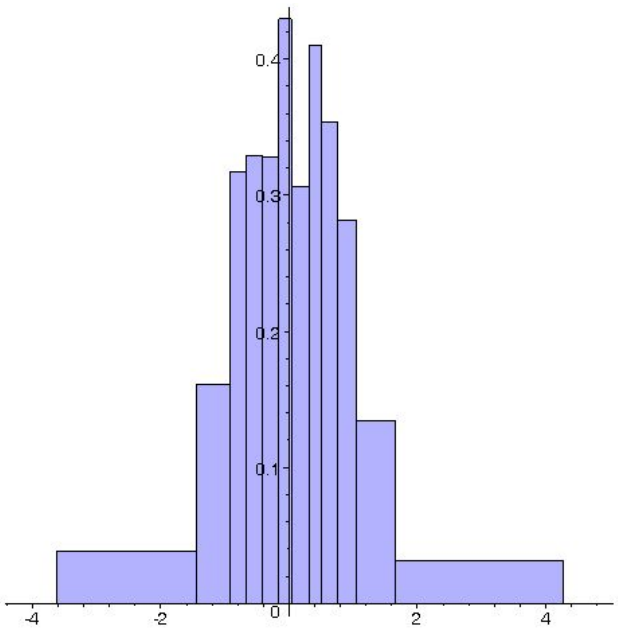
Pokud je i možných hodnot více (nebo dokonce připouštíme kontinuální reálné hodnoty), dělíme často možný rozsah hodnot na vhodný počet intervalů a o statistickém znaku uvádíme četnost hodnot v daných intervalech. Intervalům se často říká **třídy** a počtu znaku ve třídě pak **třídní četnost**.
Používáme 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**

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

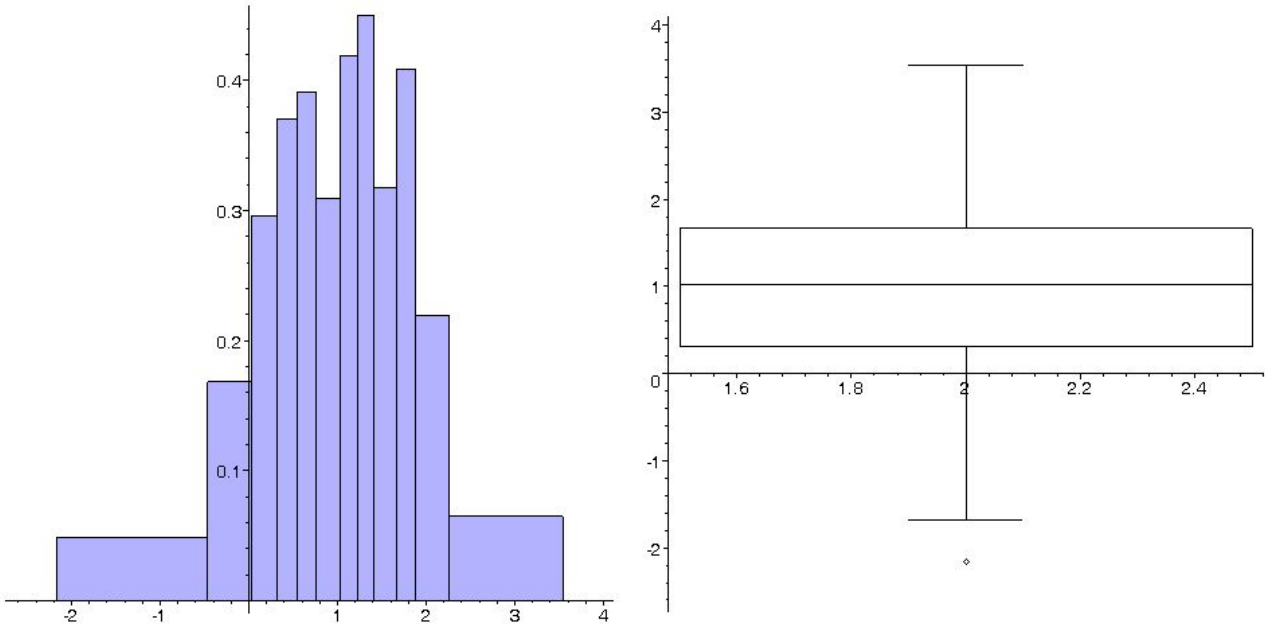
Všimněme si, že tady jde o skutečný průměr vzdáleností hodnot znaků, ovšem od mediánu!
Následující věta říká, proč zrovna tyto míry volíme:

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.*

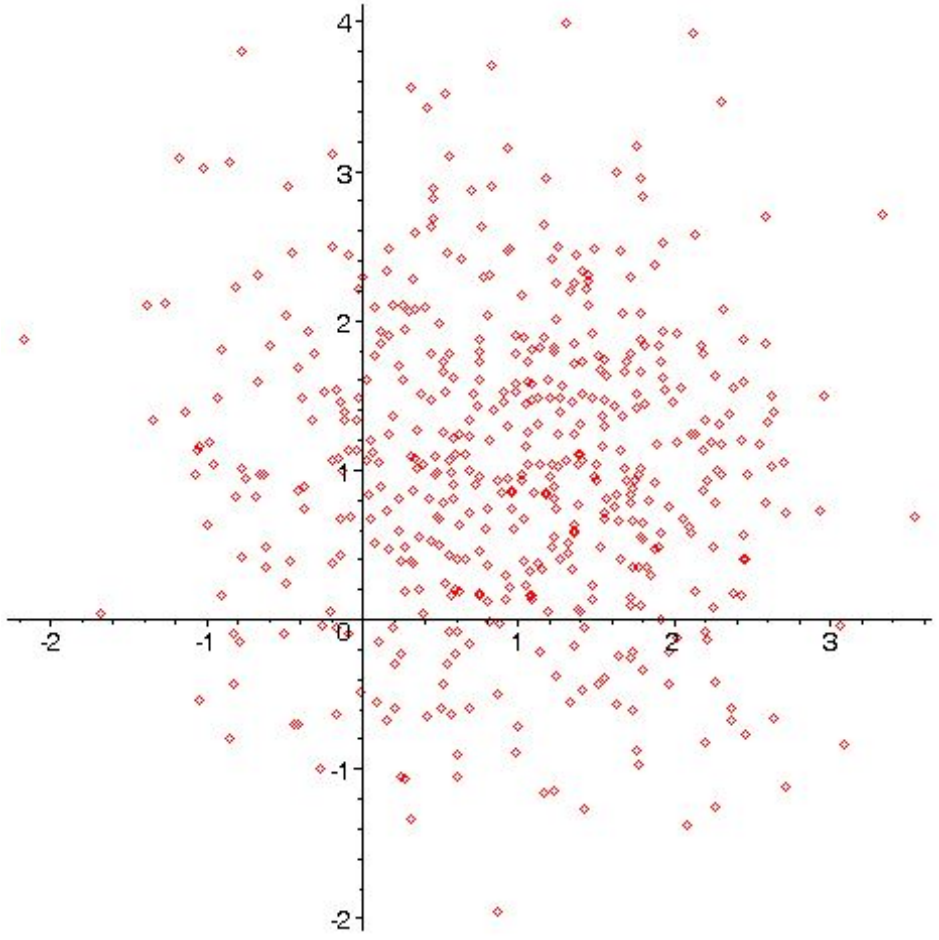
Diagramy

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 **entropii**

$$H_X = -\sum_{i=1}^{n} p_i \ln(p_i).$$

Je-li $p_k = 1$ a ostaní $p_j = 0$, pak je variabilita je nulová tomu odpovídá $H_X = 0$.

Entropie je chrakterizovaná následující vlastností. Pro soubor znaků Z tvořený dvojicemi znaků ze souborů X a Y (např. můžeme na statistických jednotkách-osobách sledovat barvu očí a barvu vlasů), je variabilita znaků z součtem variabilit jednotlivých znaků, tj. $H_Z = H_X + H_Y$.
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.

Plán přednášky

Připomeneme (a trochu zobecníme) pojmy a výsledky z druhé přednášky prvního semestru.

Definition (Náhodné jevy)

Budeme pracovat s neprázdnou pevně zvolenou množinou Ω všech možných výsledků, kterou nazýváme **základní prostor**.

Připomeneme (a trochu zobecníme) pojmy a výsledky z druhé přednášky prvního semestru.

Definition (Náhodné jevy)

Budeme pracovat s neprázdnou pevně zvolenou množinou Ω všech možných výsledků, kterou nazýváme **základní prostor**.
Prvky $\omega \in \Omega$ představují jednotlivé **možné výsledky**.

Připomeneme (a trochu zobecníme) pojmy a výsledky z druhé přednášky prvního semestru.

Definition (Náhodné jevy)

Budeme pracovat s neprázdnou pevně zvolenou množinou Ω všech možných výsledků, kterou nazýváme **základní prostor**.
Prvky $\omega \in \Omega$ představují jednotlivé **možné výsledky**.
Systém podmnožin $\mathcal{A}$ základního prostoru se nazývá **jevové pole** a jeho prvky se nazývají **jevy**, jestliže

- $\Omega \in \mathcal{A}$, tj. základní prostor, je jevem,
- je-li $A, B \in \mathcal{A}$, pak $A \setminus B \in \mathcal{A}$, tj. pro každé dva jevy je jevem i jejich množinový rozdíl,
- je-li $A_i \in \mathcal{A}$, $i \in I$ nejvýše spočetný systém jevů, pak také jejich sjednocení je jevem, tj. $\cup_{i \in I} A_i \in \mathcal{A}$.

- Komplement $A^c = \Omega \setminus A$ jevu A je jevem, který nazýváme *opačný jev* k jevu A.

- Komplement $A^c = \Omega \setminus A$ jevu A je jevem, který nazýváme *opačný jev* k jevu A.
- Průnik dvou jevů opět jevem, protože pro každé dvě podmnožiny A, $B \subset \Omega$ platí

$$A \setminus (\Omega \setminus B) = A \cap B.$$

- Komplement $A^c = \Omega \setminus A$ jevu A je jevem, který nazýváme *opačný jev* k jevu A.
- Průnik dvou jevů opět jevem, protože pro každé dvě podmnožiny A, $B \subset \Omega$ platí

$$A \setminus (\Omega \setminus B) = A \cap B.$$

Jevové pole je tedy systém podmnožin základního prostoru uzavřený na konečné průniky, spočetná sjednocení a množinové rozdíly. Jednotlivé množiny $A \in \mathcal{A}$ nazýváme **náhodné jevy** (vzhledem k $\mathcal{A}$).

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají **elementární jevy**,

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají **elementární jevy**,
- **společné nastoupení jevů** A_i, $i \in I$, odpovídá jevu $\cap_{i \in I} A_i$, **nastoupení alespoň jednoho z jevů** A_i, $i \in I$, odpovídá jevu $\cup_{i \in I} A_i$,

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají **elementární jevy**,
- **společné nastoupení jevů** A_i, $i \in I$, odpovídá jevu $\cap_{i \in I} A_i$, **nastoupení alespoň jednoho z jevů** A_i, $i \in I$, odpovídá jevu $\cup_{i \in I} A_i$,
- $A, B \in \mathcal{A}$ jsou **neslučitelné jevy**, je-li $A \cap B = \emptyset$,

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají **elementární jevy**,
- **společné nastoupení jevů** A_i, $i \in I$, odpovídá jevu $\cap_{i \in I} A_i$, **nastoupení alespoň jednoho z jevů** A_i, $i \in I$, odpovídá jevu $\cup_{i \in I} A_i$,
- $A, B \in \mathcal{A}$ jsou **neslučitelné jevy**, je-li $A \cap B = \emptyset$,
- jev A má za **důsledek** jev B, když $A \subset B$,

Terminologie připomíná souvislosti s popisem skutečných jevů a jejich statistickým popisem:

- celý základní prostor Ω se nazývá **jistý jev**, prázdná podmnožina $\emptyset \in \mathcal{A}$ se nazývá **nemožný jev**,
- jednoprvkové podmnožiny $\{\omega\} \in \Omega$ se nazývají **elementární jevy**,
- **společné nastoupení jevů** A_i, $i \in I$, odpovídá jevu $\cap_{i \in I} A_i$, **nastoupení alespoň jednoho z jevů** A_i, $i \in I$, odpovídá jevu $\cup_{i \in I} A_i$,
- $A, B \in \mathcal{A}$ jsou **neslučitelné jevy**, je-li $A \cap B = \emptyset$,
- jev A má za **důsledek** jev B, když $A \subset B$,
- je-li $A \in \mathcal{A}$, pak se jev $B = \Omega \setminus A$ nazývá **opačný jev k jevu** A, píšeme $B = A^c$.

Definition (Pravděpodobnost)

Pravděpodobnostní prostor je jevové pole $\mathcal{A}$ podmnožin (konečného) základního prostoru Ω, na kterém je definována skalární funkce $P : \mathcal{A} \to \mathbb{R}$ s následujícími vlastnosti:

- je nezáporná, tj. $P(A) \geq 0$ pro všechny jevy A,
- je aditivní, tj. $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$, pro každý nejvýše spočetný systém po dvou disjunktních jevů,
- pravděpodobnost jistého jevu je 1.

Funkci P nazýváme **pravděpodobností** na jevovém poli $(\Omega, \mathcal{A})$.

Definition (Pravděpodobnost)

Pravděpodobnostní prostor je jevové pole $\mathcal{A}$ podmnožin (konečného) základního prostoru Ω, na kterém je definována skalární funkce $P : \mathcal{A} \to \mathbb{R}$ s následujícími vlastnosti:

- je nezáporná, tj. $P(A) \geq 0$ pro všechny jevy A,
- je aditivní, tj. $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$, pro každý nejvýše spočetný systém po dvou disjunktních jevů,
- pravděpodobnost jistého jevu je 1.

Funkci P nazýváme **pravděpodobností** na jevovém poli $(\Omega, \mathcal{A})$.

Důsledky

Pro všechny jevy platí $P(A^c) = 1 - P(A)$.

Definition (Pravděpodobnost)

Pravděpodobnostní prostor je jevové pole $\mathcal{A}$ podmnožin (konečného) základního prostoru Ω, na kterém je definována skalární funkce $P : \mathcal{A} \to \mathbb{R}$ s následujícími vlastnosti:

- je nezáporná, tj. $P(A) \geq 0$ pro všechny jevy A,
- je aditivní, tj. $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$, pro každý nejvýše spočetný systém po dvou disjunktních jevů,
- pravděpodobnost jistého jevu je 1.

Funkci P nazýváme **pravděpodobností** na jevovém poli $(\Omega, \mathcal{A})$.

Důsledky

Pro všechny jevy platí $P(A^c) = 1 - P(A)$.
Additivnost platí pro jakýkoliv spočetný počet neslučitelných jevů $A_i \subset \Omega$, $i \in I$, tj.

$$P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i), \text{ kdykoliv je } A_i \cap A_j = \emptyset,\ i \neq j,\ i, j \in I.$$

Připomeňme si klasickou konečnou pravděpodobnost.

Připomeňme si klasickou konečnou pravděpodobnost.

Definition

Nechť Ω je konečný základní prostor a nechť jevové pole $\mathcal{A}$ je právě systém všech podmnožin v Ω. **Klasická pravděpodobnost** je pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ s pravděpodobnostní funkcí $P : \mathcal{A} \to \mathbb{R}$,

$$P(A) = \frac{|A|}{|\Omega|}.$$

Zjevně takto zadaná funkce skutečně definuje pravděpodobnost.

Peterburgský paradox (Bernoulli, 1738)

Typický příklad klasické pravděpodobnosti jsou jevy související s házením mincí. Představme si následující pravidla kasina:

Peterburgský paradox (Bernoulli, 1738)

Typický příklad klasické pravděpodobnosti jsou jevy související s házením mincí. Představme si následující pravidla kasina: Návštěvník zaplatí vklad C a poté hází mincí. Je-li T počet hodů potřebných k první hlavě, pak obdrží výhru 2^T. Jaká je „fér hodnota“ pro vklad C?

Peterburgský paradox (Bernoulli, 1738)

Typický příklad klasické pravděpodobnosti jsou jevy související s házením mincí. Představme si následující pravidla kasina: Návštěvník zaplatí vklad C a poté hází mincí. Je-li T počet hodů potřebných k první hlavě, pak obdrží výhru 2^T. Jaká je „fér hodnota“ pro vklad C?

Pravděpodobnost, že padne hlava je u férové mince $1/2$, je proto $P(T = k) = 2^{-k}$. Pravděpodobnost, že po nějakém konečném počtu hodů hra skončí je dána součtem $\sum_{k=1}^{\infty} 2^{-k} = 1$. Proto je úpravděpodobnost jevu, že stále padá orel nulová.

Sečteme-li všechny pravděpodobnosti výsledků vynásobených výhrami 2^k, dostaneme $\sum_1^{\infty} 1 = \infty$. Zdá se proto, že se vyplatí vložit i velký vklad...

Peterburgský paradox (Bernoulli, 1738)

Typický příklad klasické pravděpodobnosti jsou jevy související s házením mincí. Představme si následující pravidla kasina: Návštěvník zaplatí vklad C a poté hází mincí. Je-li T počet hodů potřebných k první hlavě, pak obdrží výhru 2^T. Jaká je „fér hodnota“ pro vklad C?

Pravděpodobnost, že padne hlava je u férové mince $1/2$, je proto $P(T = k) = 2^{-k}$. Pravděpodobnost, že po nějakém konečném počtu hodů hra skončí je dána součtem $\sum_{k=1}^{\infty} 2^{-k} = 1$. Proto je úpravděpodobnost jevu, že stále padá orel nulová.

Sečteme-li všechny pravděpodobnosti výsledků vynásobených výhrami 2^k, dostaneme $\sum_1^{\infty} 1 = \infty$. Zdá se proto, že se vyplatí vložit i velký vklad...

Ve skutečnosti simulací hry zjistíme, že nezávisle na počtu pokusů se prakticky všechny výhry budou pohybovat v rozmezí T do 6. Důvodem je, že vysoké výhry jsou velice nepravděpodobné a proto je při reálných úvahách nelze brát vážně.

Podmíněná pravděpodobnost

Obvyklé je také klást dotazy s dodatečnou podmínkou. Např. „jaká je pravděpodobnost, že při hodu dvěmi kostkami padly dvě pětky, je-li součet hodnot deset?“. Připomeneme, že formalizovat takové úvahy umíme následovně.

Definition

Nechť H je jev s nenulovou pravděpodobností v jevovém poli $\mathcal{A}$ v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. **Podmíněná pravděpodobnost** $P(A|H)$ jevu $A \in \mathcal{A}$ vzhledem k hypotéze H je definována vztahem

$$P(A|H) = \frac{P(A \cap H)}{P(H)}.$$

Definice odpovídá požadavku, že jevy A a H nastanou zároveň, za předpokladu, že A nastal s pravděpodobností $P(A \cap H)/P(A)$.

Podmíněná pravděpodobnost

Obvyklé je také klást dotazy s dodatečnou podmínkou. Např. „jaká je pravděpodobnost, že při hodu dvěmi kostkami padly dvě pětky, je-li součet hodnot deset?“. Připomeneme, že formalizovat takové úvahy umíme následovně.

Definition

Nechť H je jev s nenulovou pravděpodobností v jevovém poli $\mathcal{A}$ v pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. **Podmíněná pravděpodobnost** $P(A|H)$ jevu $A \in \mathcal{A}$ vzhledem k hypotéze H je definována vztahem

$$P(A|H) = \frac{P(A \cap H)}{P(H)}.$$

Definice odpovídá požadavku, že jevy A a H nastanou zároveň, za předpokladu, že A nastal s pravděpodobností $P(A \cap H)/P(A)$.
Je také vidět přímo z definice, hypotéza H a jev A jsou nezávislé tehdy a jen tehdy, je-li $P(A) = P(A|H)$.

Bayesovy věty

Přepsáním formule pro podmíněnou pravděpodobnost dostáváme

$$P(A \cap B) = P(B \cap A) = P(A)P(B|A) = P(B)P(A|B).$$

Theorem (Bayesovy věty)

Pro pravděpodobnost jevů A a B platí

1. $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$.
2. $P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A')P(B|A')}$.

Bayesovy věty

Přepsáním formule pro podmíněnou pravděpodobnost dostáváme

$$P(A \cap B) = P(B \cap A) = P(A)P(B|A) = P(B)P(A|B).$$

Theorem (Bayesovy věty)

Pro pravděpodobnost jevů A a B platí

1. $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$.
2. $P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A')P(B|A')}$.

Důkaz.

První tvrzení je přepsáním předchozí formule, druhé z prvého plyne doszením $P(B) = P(A)P(B|A) + P(A')P(B|A')$. □

Příklad – testování

Předpokládejme, že předpokladem přijetí studentů na univerzitu jsou testy způsobilosti ke studiu. Inteligentní osoba v něm má 99% úspěšnost. Zároveň předpokládejme, že úspěšnost neinteligentních osob je 0.5%.

Příklad – testování

Předpokládejme, že předpokladem přijetí studentů na univerzitu jsou testy způsobilosti ke studiu. Inteligentní osoba v něm má 99% úspěšnost. Zároveň předpokládejme, že úspěšnost neinteligentních osob je 0.5%.
S jakou pravděpodobností je náhodně vybraný student/ka univerzity inteligentní, jestliže je v populaci je p promile inteligentních osob (tj. p osob z tisíce považujeme za inteligentní).

Příklad – testování

Předpokládejme, že předpokladem přijetí studentů na univerzitu jsou testy způsobilosti ke studiu. Inteligentní osoba v něm má 99% úspěšnost. Zároveň předpokládejme, že úspěšnost neinteligentních osob je 0.5%.
S jakou pravděpodobností je náhodně vybraný student/ka univerzity inteligentní, jestliže je v populaci je p promile inteligentních osob (tj. p osob z tisíce považujeme za inteligentní).
Označme A jev, že je daná osoba je inteligentní, a B jev, že prošla testem. Dle Bayesovy věty je hledaná pravděpodobnost

$$P(A|B) = \frac{p/1000 \cdot 99/100}{p/1000 \cdot 99/100 + (1000 - p)/1000 \cdot 5/1000}$$

Jestliže zvolíme za p nějaké konkrétní četnosti, dostaneme příslušné očekávatelné spolehlivosti testu. V následující tabulce je spočten výsledek pro několik p:

p	500	100	10	1	0.1
$P(A\|B)$	0.99	0,96	0.67	0.17	0.02

Pokud stejné číselné zadání použijeme pro screening některé nemoci, řekněme HIV pozitivity, dostáváme hrozné výsledky!

Výsledek asi neodpovídá naší intuici a může se zdát šokující ve vztahu k použití takovýchto testů.

Výsledek asi neodpovídá naší intuici a může se zdát šokující ve vztahu k použití takovýchto testů.
Evidentně prostý výběr náhodné osoby a použití jediného testu, byť velmi citlivého, specifického a účinného, nejsou vhodné ani na otestování skutečného stavu populace, ani na preventivní vyšetření jednotlivců, pokud nemáme další podpůrné informace a lepší nástroje.

Výsledek asi neodpovídá naší intuici a může se zdát šokující ve vztahu k použití takovýchto testů.
Evidentně prostý výběr náhodné osoby a použití jediného testu, byť velmi citlivého, specifického a účinného, nejsou vhodné ani na otestování skutečného stavu populace, ani na preventivní vyšetření jednotlivců, pokud nemáme další podpůrné informace a lepší nástroje.
Právě matematická statistika dává nástroje na kvalifikovanější postupy v medicínské i průmyslové diagnostice, ekonomických modelech, vyhodnocování experimentálních dat atd.

Plán přednášky

Vraťme se k jednoduchému a názornému příkladu statistik kolem výsledků studentů v daném předmětu. Je a není podobný klasické pravděpodobnosti a s ní související statistice při házení kostkou.

Vraťme se k jednoduchému a názornému příkladu statistik kolem výsledků studentů v daném předmětu. Je a není podobný klasické pravděpodobnosti a s ní související statistice při házení kostkou. Na jedné straně jsme připustili pouze konečný počet možných bodových hodnocení (celá čísla od 0 do 20), zároveň ale není patrně vhodné představovat si výsledky jednotlivých studentů jako analogii nezávislého házení kostkou (to by byla skutečně divně vedená přednáška).

Vraťme se k jednoduchému a názornému příkladu statistik kolem výsledků studentů v daném předmětu. Je a není podobný klasické pravděpodobnosti a s ní související statistice při házení kostkou. Na jedné straně jsme připustili pouze konečný počet možných bodových hodnocení (celá čísla od 0 do 20), zároveň ale není patrně vhodné představovat si výsledky jednotlivých studentů jako analogii nezávislého házení kostkou (to by byla skutečně divně vedená přednáška).
Místo toho máme na základním prostoru Ω všech studentů definovánu funkci bodového ohodnocení $X : \Omega \to \mathbb{R}$. Je to typický příklad **náhodné veličiny**.
S každou náhodnou veličinou potřebujeme umět pracovat s vhodnou množinou jevů. Zpravidla požadujeme, abychom mohli pracovat s pravděpodobnostmi příslušnosti hodnoty X do předem zadaného intervalu.

Na prostoru $\mathbb{R}^k$ uvažujme nejmenší jevové pole $\mathcal{B}$ obsahující všechny k–rozměrné intervaly. Množinám v $\mathcal{B}$ říkáme **Borelovské množiny** na $\mathbb{R}^k$.

Na prostoru $\mathbb{R}^k$ uvažujme nejmenší jevové pole $\mathcal{B}$ obsahující všechny k–rozměrné intervaly. Množinám v $\mathcal{B}$ říkáme **Borelovské množiny** na $\mathbb{R}^k$.

Definition (Náhodné veličiny a distribuční funkce)

Náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ je taková funkce $X : \Omega \to \mathbb{R}$, že vzor $X^{-1}(B)$ patří do $\mathcal{A}$ pro každou Borelovskou množinu $B \in \mathcal{B}$ na $\mathbb{R}$.

Na prostoru $\mathbb{R}^k$ uvažujme nejmenší jevové pole $\mathcal{B}$ obsahující všechny k–rozměrné intervaly. Množinám v $\mathcal{B}$ říkáme **Borelovské množiny** na $\mathbb{R}^k$.

Definition (Náhodné veličiny a distribuční funkce)

Náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ je taková funkce $X : \Omega \to \mathbb{R}$, že vzor $X^{-1}(B)$ patří do $\mathcal{A}$ pro každou Borelovskou množinu $B \in \mathcal{B}$ na $\mathbb{R}$.
Náhodný vektor $(X_1, \ldots, X_k)$ na $(\Omega, \mathcal{A}, P)$ je k–tice náhodných veličin.

Definice náhodné veličiny zajišťuje, že pro všechny $-\infty \leq a \leq b \leq \infty$ existuje pravděpodobnost $P(a \leq X < b)$, kde používáme stručné značení pro jev $A = (\omega \in \Omega;\ a \leq X(\omega) < b))$.

Definition

Distribuční funkcí náhodné veličiny X je funkce $F : \mathbb{R} \to \mathbb{R}$ definovaná pro všechny $x \in \mathbb{R}$ vztahem

$$F(x) = P(X < x).$$

Definice náhodné veličiny zajišťuje, že pro všechny $-\infty \leq a \leq b \leq \infty$ existuje pravděpodobnost $P(a \leq X < b)$, kde používáme stručné značení pro jev $A = (\omega \in \Omega;\ a \leq X(\omega) < b))$.

Definition

Distribuční funkcí náhodné veličiny X je funkce $F : \mathbb{R} \to \mathbb{R}$ definovaná pro všechny $x \in \mathbb{R}$ vztahem

$$F(x) = P(X < x).$$

Distribuční funkcí náhodného vektoru $(X_1, \ldots, X_k)$ je funkce $F : \mathbb{R}^k \to \mathbb{R}$ definovaná pro všechny $(x_1, \ldots, x_k) \in \mathbb{R}^k$ vztahem

$$F(x) = P(X_1 < x_1 \wedge \cdots \wedge X_k < x_k).$$

Diskrétní náhodné veličiny

Předpokládejme, že pro náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$. Pak existuje tzv. **pravděpodobnostní funkce** $f(x)$ taková, že

$$f(x) = \begin{cases} P(X = x_i) & x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Evidentně $\sum_1^n f(x_i) = 1$.

Diskrétní náhodné veličiny

Předpokládejme, že pro náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$. Pak existuje tzv. **pravděpodobnostní funkce** $f(x)$ taková, že

$$f(x) = \begin{cases} P(X = x_i) & x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Evidentně $\sum_1^n f(x_i) = 1$.
Takové náhodné veličině se říká **diskrétní**.

Diskrétní náhodné veličiny

Předpokládejme, že pro náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$. Pak existuje tzv. **pravděpodobnostní funkce** $f(x)$ taková, že

$$f(x) = \begin{cases} P(X = x_i) & x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Evidentně $\sum_1^n f(x_i) = 1$.
Takové náhodné veličině se říká **diskrétní**.
Každá náhodná veličina definovaná pro klasickou pravděpodobnost je diskrétní.

Diskrétní náhodné veličiny

Předpokládejme, že pro náhodná veličina X na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ nabývá jen konečně mnoha hodnot $x_1, x_2, \ldots, x_n \in \mathbb{R}$. Pak existuje tzv. **pravděpodobnostní funkce** $f(x)$ taková, že

$$f(x) = \begin{cases} P(X = x_i) & x = x_i \\ 0 & \text{jinak.} \end{cases}$$

Evidentně $\sum_1^n f(x_i) = 1$.
Takové náhodné veličině se říká **diskrétní**.
Každá náhodná veličina definovaná pro klasickou pravděpodobnost je diskrétní. Obdobně lze definici pravděpodobnostní funkce rozšířit na veličiny se spočetně mnoha hodnotami (pracujeme pak s absolutně konvergentními nekonečnými řadami :-)

Spojité náhodné veličiny

I když hodnoty náhodné veličiny X nejsou diskrétní, můžeme postupovat podobně s užitím nástrojů diferenciálního a integrálního počtu. Intuitivně lze uvažovat takto: **hustotu** $f(x)$ **pravděpodobnosti** pro X si představíme jako

$$P(x \leq X < x + dx) = f(x)dx.$$

Spojité náhodné veličiny

I když hodnoty náhodné veličiny X nejsou diskrétní, můžeme postupovat podobně s užitím nástrojů diferenciálního a integrálního počtu. Intuitivně lze uvažovat takto: **hustotu** $f(x)$ **pravděpodobnosti** pro X si představíme jako

$$P(x \leq X < x + dx) = f(x)dx.$$

To znamená, že chceme pro $-\infty \leq a \leq b \leq \infty$

$$P(a \leq X < b) = \int_a^b f(x)dx. \qquad (*)$$

Spojité náhodné veličiny

I když hodnoty náhodné veličiny X nejsou diskrétní, můžeme postupovat podobně s užitím nástrojů diferenciálního a integrálního počtu. Intuitivně lze uvažovat takto: **hustotu** $f(x)$ **pravděpodobnosti** pro X si představíme jako

$$P(x \leq X < x + dx) = f(x)dx.$$

To znamená, že chceme pro $-\infty \leq a \leq b \leq \infty$

$$P(a \leq X < b) = \int_a^b f(x)dx. \qquad (*)$$

Definition

Náhodná veličina X, pro kterou existuje její **hustota pravděpodobnosti** splňující $(*)$, se nazývá **spojitá**.

Theorem

Pro každou náhodnou veličinu X má její dstribuční funkce $F : \mathbb{R} \to [0, 1]$ následující vlastnosti

1. *F je neklesající funkce;*
2. *F má v každém bodě $x \in \mathbb{R}$ limitu zleva i limitu zprava;*
3. *F je zleva spojitá;*
4. *v nevlastních bodech má F limity*

$$\lim_{x \to \infty} F(x) = 1, \qquad \lim_{x \to -\infty} = 0; \tag{1}$$

5. *pravděpodobnost, že X nabývá právě hodnotu x je dána*

$$P(X = x) = \lim_{y \to x_+} F(y) - F(x). \tag{2}$$

6. *Distribuční funkce náhodné veličiny má vždy nejvýše spočetně mnoho bodů nespojitosti.*

Důkaz věty je založený na pozorování vyplývajícím vcelku jednoduše z axiomů pravděpodobnosti:

Theorem

Uvažme pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ *a neklesající řetězec jevů* $A_1 \subset A_2 \subset \dots$. *Pak platí*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i).$$

Pokud je naopak $A_1 \supset A_2 \supset A_3 \supset \dots$, *potom platí*

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i).$$