STATISTIKA A PRAVDĚPODOBNOST

OJ205

statistický soubor

- Statistický soubor je posloupnost údajů o nějakých objektech.
- Typy těchto údajů nazýváme statistické znaky.
- Jejich počet pak určuje rozměr statistického souboru.
- Základní soubor (též populace) uvažuje všechny objekty daného typu.
- Statistický soubor je omezený výběr objektů ze základního souboru.

jednorozměrný statistický soubor

- **Rozsah** statistického souboru je počet jeho prvků.
- **Absolutní četnost** hodnoty (někdy též pouze četnost) v souboru je počet jejích výskytů.
- **Relativní četnost** je absolutní četnost podělená rozsahem souboru a udává se obvykle v procentech.
- **Kumulativní četnost** hodnoty je četnost hodnoty souboru plus četnost všech menších hodnot. Rozeznáváme opět absolutní a relativní kumulativní četnost.

charakteristiky

- Pro jednorozměrný statistický soubor zavádíme tzv. **charakteristiky polohy** a **charakteristiky variability**.

- **Charakteristiky polohy** shrnují potenciálně velké množství dat do několika málo čísel, které lze snadno interpretovat a vytvořit si tak hrubý úsudek o celém vzorku dat.

- **Charakteristiky variability** ukazují, jak je statistický soubor vnitřně konzistentní, čili jak moc se od sebe vzájemně liší hodnoty obsažené v souboru.

charakteristiky polohy

- **Modus**: je hodnota či třída s největší četností.
- **Aritmetický průměr** (značený avg) je součet hodnot ve statistickém souboru, podělený velikostí souboru.
- **Medián** je „prostřední” hodnota v souboru po jeho setřídění. V případě, že datový soubor má sudý počet prvků, je to průměr ze dvou prostředních.

charakteristiky variability

Hlavní charakteristikou variability statistického souboru je **rozptyl** (též **disperze** nebo **variance**) značený s^2. Je definován jako

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - avg)^2 = ((x_1 - avg)^2 + (x_2 - avg)^2 + ... + (x_n - avg)^2)/n$$

dvourozměrný statistický soubor

- **Dvourozměrný statistický soubor** lze chápat jako dva jednorozměrné soubory, vzájemně provázané. Formálně jej můžeme reprezentovat jako posloupnost uspořádaných dvojic, $((x1, y1), (x2, y2), ..., (xn, yn))$.

- Důležitou vlastností dvourozměrného statistického souboru je **korelace statistických znaků**. Pojmem korelace rozumíme stupeň lineární závislosti znaků x a y, tedy to, do jaké míry hodnoty znaku x lineárně závisí na hodnotách znaku y.
- (Jinými slovy, to, jak dobře lze grafem závislosti x na y proložit přímku.)

korelace znaků

- Formální zápis – ($s(x)$, $s(y)$ jsou směrodatné odchylky jednorozměrných datových souborů pro znaky x a y):

$$\frac{\sum_{i=1}^{n}(x_i - avg(x))(y_i - avg(y))}{n * s(x) * s(y)}$$

korelace znaků

- Hodnoty korelace se pohybují od -1 do 1.
- Pokud je korelace 0, jsou hodnoty znaků dokonale nezávislé.
- Pokud je korelace 1, jedná se o přímou úměrnost
- (čím větší je x, tím větší je y a hodnoty y lze z hodnot x získat jednoduše vynásobením nějakou kladnou konstantou).
- Pokud je korelace -1, jedná se o nepřímou úměrnost (čím větší je x, tím menší je y a hodnoty y lze z hodnot x získat jednoduše vynásobením nějakou zápornou konstantou).

pravděpodobnostní rozložení

- **Náhodná proměnná**, A je vlastnost, jejíž hodnotu neznáme, protože o ní nemáme dost informací nebo protože dosud žádné hodnoty nenabyla.

- Většinou máme nějaké informace o dané vlastnosti, které nám mohou vyloučit nebo téměř vyloučit některé hodnoty, například minulá pozorování.

pravděpodobnostní rozdělení

- Pravděpodobnostní rozdělení (pravděpodobnostní rozložení, pravděpodobnostní distribuce) jevu či vlastnosti A, je funkce, která pro jednotlivé možné hodnoty ukazuje pravděpodobnost, s jakou vlastnost A nabude této hodnoty.

- Formálně se jedná o funkci

$$p : X \rightarrow [0, 1]$$

kde X je množina možných hodnot příslušné vlastnosti a [0, 1] je uzavřený interval od nuly do jedné, tedy

$$\forall x \in X(p(x) \leq 1 \wedge p(x) \geq 0).$$

pravděpodobnostní rozdělení

- Musí platit, že součet hodnot funkce pro všechny možné hodnoty je 1, tedy

$$\sum_{x \in X} p(x) = 1$$

- Zároveň platí:

- $p(x) = P(A = x)$

(Hodnota pravděpodobnostního rozložení (malé p) je rovna pravděpodobnosti (velké P, tedy obecná pravděpodobnost), s jakou vlastnost A nabude hodnoty x. Dvojice (X, p), tedy množina všech možných hodnot vlastnosti spolu s pravděpodobnostním rozložením, se nazývá pravděpodobnostní prostor.)

určení pravděpodobnostního rozložení

- První možnost: použití nějaké „ideální” funkce, které vychází z našich předpokladů o dané vlastnosti. (například hod kostkou)

- Druhá možnost: určovat pravděpodobnostní rozložení na základě měření provedeného v minulosti, které bylo zachyceno ve statistickém souboru. (například pravděpodobnostní rozložení slov, sousloví, slovních druhů v jazyku)

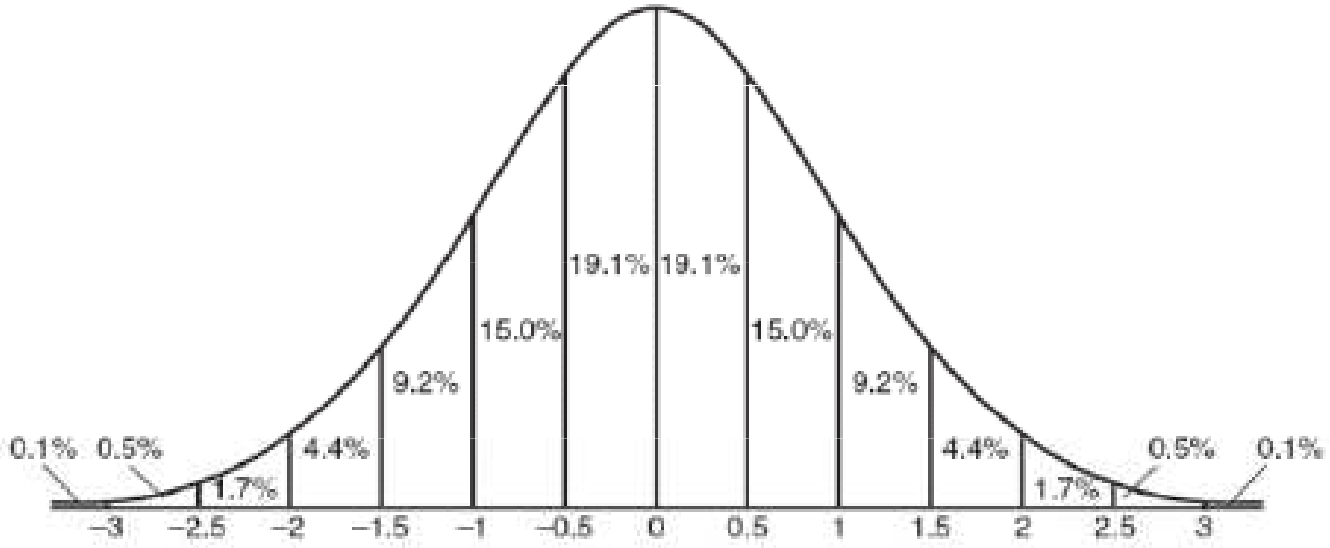
uniformní pravděpodobnostní rozložení

- Uniformní rozložení je takové, v němž všechny hodnoty mají přibližně stejnou pravděpodobnost. Grafem jsou tedy body uspořádané přibližně do přímky vodorovné s osou x.

- Příkladem může být pravděpodobnostní rozložení výsledků hodu vyváženou kostkou.

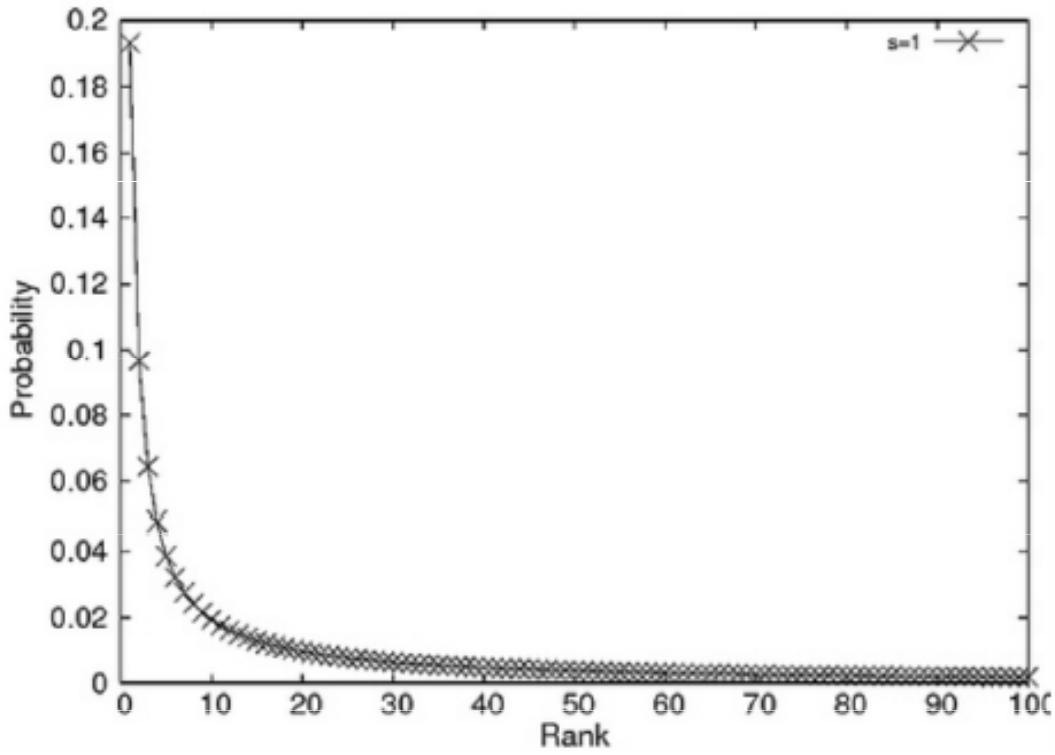
normální pravděpodobnostní rozložení

- Normální rozložení se vyznačuje tím, že nejpravděpodobnější hodnoty jsou blízké průměru a s větší odchylkou od průměru pravděpodobnost klesá.
- Graf takového rozložení má tvar zvonu, např.:



Zipfovo pravděpodobnostní rozložení

- Zipfovo rozložení: několik málo nejčastějších hodnot má velkou pravděpodobnost, s každou další hodnotou (setřídění od nejčastější) tato pravděpodobnost prudce klesá.



Zipfův zákon

- Zipfovo rozložení výstižně popisuje velké množství jevů, proto se někdy označuje jako „Zipfův zákon”.
- Zejména v přirozeném jazyce tento zákon platí téměř všude: téměř vždy je frekvence (nebo ekvivalentně – pravděpodobnost výskytu) zhruba nepřímo úměrná pořadí podle této frekvence; to platí pro slova, dvojice slov, slovní druhy, syntaktické vztahy, sémantické kategorie a mnohá další.
- Frekvence nejčastějších slovních tvarů v angličtině: „the” má relativní četnost 7 %, druhé „of” má 3,5 % a více než polovina anglických korpusů je pokryta 135 nejčastějšími slovy (stoplist)

distribuční funkce

- **Pravděpodobnostní rozložení** je pravděpodobnost, že náhodná veličina nabude určité hodnoty (resp. zda patří do dané třídy), čili $p(x) = P(A = x)$, a její hodnoty odpovídají relativním četnostem ve statistickém souboru.

- **Distribuční funkce** (cumulative distribution function) F, je pravděpodobnost, že náhodná veličina nabude určité hodnoty nebo menší, čili $F(x) = P(A \leq x)$.
- Její hodnoty odpovídají kumulativním relativním četnostem ve statistickém souboru. Hodnoty distribuční funkce jsou také dobře známé jako tzv. percentil. Hodnota distribuční funkce (percentil) mediánu statistického souboru je 0,5.

náhodný vektor

- Náhodný vektor je posloupnost náhodných veličin (počasí). Jeho pravděpodobnostní rozložení můžeme modelovat s využitím vícerozměrného statistického souboru.
- Pro dvourozměrný náhodný vektor (A,B) je hodnota pravděpodobnostního rozložení $p(x, y) = P(A = x \wedge B = y)$.
- Lze definovat i distribuční funkci pro náhodný vektor, např. pro dvourozměrný náhodný vektor je distribuční funkce analogicky definována jako $F(x, y) = P(A \leq x \wedge B \leq y)$.

podmíněná pravděpodobnost

- **Podmíněná pravděpodobnost** je motivována potřebou formalizovat to, že často máme kromě pravděpodobnostního rozložení daného jevu další informace o jiném jevu, který s původním může, ale nemusí souviset.
- Podmíněnou pravděpodobnost zapisujeme $P(A|B)$

  a čteme „pravděpodobnost jevu A za předpokladu,

  že nastal jev B”.
- Podmíněnou pravděpodobnost lze vypočítat:

  $P(A|B) = P(A,B) \,|P(B)$

  kde $P(A,B)$ je pravděpodobnost, že jevy A a B

  nastaly současně.

nezávislé jevy

- Skrze podmíněnou pravděpodobnost definujeme i tzv. **nezávislé jevy**. Intuitivně platí, že pokud jsou jevy nezávislé, pak by nám informace o jednom z nich neměla dát žádnou informaci o druhém z nich. Jevy A a B jsou nezávislé, pokud

$$P(A|B) = P(A) \wedge P(B|A) = P(B)$$

čili jsou nezávislé, pokud to, jestli nastal jev B,

nijak neovlivní pravděpodobnost jevu A a naopak.

- Jen a pouze pro nezávislé jevy pak platí vzorec, který se snadno odvodí z nezávislosti jevů a z definice podmíněné pravděpodobnosti:

$$P(A,B) = P(A) * P(B)$$