**Matematika a statistika**

*1. Logika a její aplikace na přirozený jazyk.*

*2. Matematické objekty – množiny, relace, uspořádání, posloupnosti, …*

*3. Statistika a pravděpodobnost – statistický soubor a jeho charakteristiky, pravděpodobnostní prostor.*

*4. Aplikace statistiky na zpracování jazyka.*

*5. Statistický jazykový (n-gramový) model – konstrukce, využití.*

*6. Vyhodnocování aplikací zpracování jazyka – přesnost, pokrytí, F-míry a zlatý standard, křížová validace.*

# 1. Logika a její aplikace na přirozený jazyk

**Logika** je věda, která se zabývá usuzováním, pravdivostí, dokazatelností a vyvratitelností. V logice jde pouze o formu sdělení. Nezajímá nás, co konkrétně je sdělováno, stejně jako nás nezajímají různé psychologické interpretace apod.

**Výroková logika**

* formální odvozovací systém, ve kterém atomické formule tvoří výrokové proměnné (na rozdíl od [predikátové logiky](https://cs.wikipedia.org/wiki/Predik%C3%A1tov%C3%A1_logika))
	+ *proměnná:* způsob [symbolické](https://cs.wikipedia.org/wiki/Symbol) reprezentace objektů, který umožňuje zcela abstraktní manipulaci s nimi. Proměnná zastupuje libovolný myslitelný objekt z dané třídy. Manipulace s proměnnými a vztahy pro ně platné mohou být chápány jako manipulace s libovolnými objekty resp. vztahy platné pro všechny objekty.
* základní jednotkou je ***výrok***, kterému lze přiřadit pravdivostní hodnotu 0 (nepravda) nebo 1 (pravda)
* ke konstrukci složitějších výroků se používají *logické funkce* – ***negace, implikace, konjunkce, disjunkce, ekvivalence***
* pokud je na levé straně implikace nepravdivý výrok, pak vždy platí celý výrok
* pravdivostní tabulky



* *tautologie* – výrok, který je za všech okolností pravdivý
* *kontradikce* – výrok, který neplatí nikdy
* *axiomy a odvozovací pravidla, modus ponens*
	+ **modus ponens:**
		- odvozovací pravidlo, základ argumentace a dokazování nejen ve výrokové logice
		- říká „Jestliže platí A a platí zároveň „z A do vyplývá B“, pak platí i B.“
		- např. P -> Q Jestliže prší, je mokro.
			* P: Prší.
			* Poté tedy platí i Q: Je mokro.
* **důkaz sporem:** oblíbená technika dokazování. Pokud chceme dokázat, že nějaký [výrok](http://www.matematika.cz/vyroky) platí, předpokládáme, že je pravdivý, pak ho znegujeme a pokusíme se dojít k nějakému sporu. Cílem je dokázat, že námi zvolený předpoklad vede k nějakému nesmyslu.

**Predikátová logika prvního řádu**

* rozšíření výrokové logiky
* formální systém používaný v matematice, filozofii, lingvistice a informatice. Často se pro její označení používá kratší a méně přesný termín predikátová logika
* predikátová logika prvního řádu se odlišuje od výrokové logiky **zavedením kvantifikovaných proměnných**
* na rozdíl od výrokové logiky si predikátová logika (PL) **všímá struktury vět samotných**. Rozlišuje v každé větě individuum, resp. individua, o němž, resp. o nichž, se něco predikuje
	+ **predikát** intuitivně chápeme jako **vlastnost nebo vztah**
* ***proměnné*** – symboly, které mohou nabývat různých hodnot, malá písmena
* ***kvantifikátory***
	+ *univerzální* (následující formule platí pro všechny možná ohodnocení proměnné za kvantifikátorem)
	+ *existenční* *(částečný) predikátor* (existuje alespoň jedno ohodnocení takové, že následující formule platí)
* ***predikáty***
	+ veškeré symboly, které nepatří do samotného jazyka logiky (např. *Prime(x)*)
	+ každý predikát má definovanou *aritu* – počet parametrů
	+ predikáty s aritou 0 nazýváme **konstantami**
	+ **predikátový symbol** je výraz označující **predikát**, tedy vlastnost nebo vztah, který lze predikovat (vypovídat) o individuu, nebo individuích (např. "být dívka" je predikát)
* **jazykově nezávislá**
* teorie o určitém tématu bývá obvykle právě predikátová logika prvního řádu společně se specifickou univerzální množinou (též. **univerzem**), ze které jsou brány proměnné, dále pak konečně mnoha funkcemi a predikáty nad touto množinou, a konečně množinou rekurzivních axiomů, jež jsou v rámci teorie pokládány za platné. Někdy pojmem *teorie* formálně rozumíme množinu vět (sentencí) zapsaných v predikátové logice
* predikátová logika prvního řádu je nesmírně **důležitá pro samotné základy matematiky**, protože je standardní logikou pro axiomatické systémy.
* žádná teorie prvního řádu nemá sílu plně a kategoricky popsat struktury s nekonečnou doménou, např. celá čísla nebo reálná čísla. K tomu jsou zapotřebí logiky vyšších řádů
* s logikou výrokovou a predikátovou pracuje **důkazová technika matematické indukce**
	+ matem. indukce = základním principem je, že dané tvrzení dokážeme pro nějaký první prvek, v přirozených číslech to nejčastěji je n = 1. To dokážeme prostým dosazením. V dalším, indukčním, kroku dokážeme implikaci „pokud tvrzení platí pro n = a, pak platí i pro n = a + 1“. Z těchto dvou kroků můžeme odvodit, že daný výraz platí pro všechna n (z nějaké množiny, se kterou zrovna pracujeme).

**PL v praxi:**

Uvažujme následující věty: “Sokrates je filozof.” a “Platón je filozof.”

Ve *výrokové logice* se bude jednat *o dva čistě nezávislé výroky*, označeny např. *p* a *q*.

V *predikátové logice* s nimi bude zacházeno daleko provázaněji za použití predikátu *P(x)*. Predikát *P* říká, že objekt, reprezentovaný proměnnou *x*, je filozof. Tedy, pokud bude *x* reprezentovat Sokrata, bude *P* říkat to samé, co výrok *p*. Obdobně tomu bude u Platóna a druhého výroku.

Na tomto příkladu je patrný klíčový aspekt predikátové logiky prvního řádu: *P* jako symbol je syntaktickou entitou, které byl dán **sémantický význam deklarováním**, že *P(x)* platí právě tehdy, když *x* je filozofem.

Takovémuto přiřazení sémantického významu říkáme **interpretace**.

**Omezenost predikátové logiky 1. řádu**

* ne všechny konstrukce v PJ jsou propozice (*Dobrý den. Děkuji Vám*)
* ne všechny propozice jsou prvního řádu (*Všichni lidé mají společné vlastnosti*)
* v PJ mnohem víc kvantifikátorů (*většina, velká část, kdekdo, pár lidí, skoro nikdo*)
* implicitní existence
* dva omezující rysy:
	+ nedostatečná expresivita
	+ extenzionalismus

**Logická analýza přirozeného jazyka**

* *Logická sémantika:* strukturní sémantika pěstovaná metodami formální logiky; je pro ni typické ztotožňovat významy vět s matematicky formalizovanými podmínkami, za nichž je věta pravdivá
* převod jazyka do formální logiky – formální odvozování, interlingua pro strojový překlad, přesné vyjádření faktů
* jaký formalismus? Predikátová logika, modální logiky, TIL
* ***predikátová logika***
	+ problém – různé typy pravdivosti a nedostatečně granulární popis, obcházení limitů může být složité
	+ výhody – jednoduchá, dobře prozkoumaná, existuje inferenční stroj
* ***TIL***
	+ *Fregeho model sémantiky (Sinn a Bedeutung; vizte Lexikologii)*
		- Článek [*O smyslu a významu*](https://cs.wikipedia.org/w/index.php?title=O_smyslu_a_v%C3%BDznamu&action=edit&redlink=1) (*Über Sinn und Bedeutung*) se zabývá problematikou [vlastních jmen](https://cs.wikipedia.org/wiki/Vlastn%C3%AD_jm%C3%A9no). Frege vychází z toho, že věty formy „a = a“ a „a = b“ mají odlišnou poznávací hodnotu. Například věta „Večernice je Večernice“ nepřináší žádnou zajímavou informaci, zatímco věta „Večernice je Jitřenka“ vyjadřuje nesamozřejmý astronomický fakt.
	+ *koncepce možných světů*
		- možný svět – množina bezesporných výroků o univerzu, aktuální svět je jeden z možných světů
		- pravdivost výpovědi závisí na světě, význam je na něm nezávislý
	+ *intenzionální logiky*
		- intenze (nezávisí na světě)
		- extenze (denotáty v nějakém světě)
	+ *transparentní intenzionální logika*
		- kromě aktuálního světa zavádí také aktuální čas
		- význam je konstrukce
		- algoritmus, který dosazením aktuálního světa a času vytvoří objekt (extenzi)
		- typovaná logika – proměnné pravda/nepravda, množina individuí, reálných čísel (časových okamžiků) a možných světu
		- propozice a vlastnosti propozice
	+ *algoritmus normální translace* – konverze přirozeného jazyka do TIL
	+ korektní a jemná analýza jazyka, je ale složitá a málo rozšiřitelná

**Transparentní intenzionální logika**

* **Pavel Materna, Pavel Tichý, Richard Montague [montegy]**
* typovaná logika; funguje jako lambda kalkul, je jazykově nezávislý
* autor systému pro TILovou analýzu na FI: **Aleš Horák**
* logický systém založený na modifikaci typovaného lambda kalkulu = logický aparát umožňující manipulaci s funkcemi
* založen na principu teorie typů. Je **vhodný pro sémantickou analýzu výrazů PJ**. Byl navržen speciálně pro zachycení významu výrazů PJ
* pro TIL je sémantika výrazu dána tím, že způsob, jakým je tento výraz strukturován, zobrazuje strukturu konstrukce, jejímiž složkami nejsou složky jazykového výrazu, nýbrž objekty těmito složkami označené. (Např. ve formálním pojetí je sémantikou výrazu *3+5* číslo 8, v pojetí TIL je to určitý způsob, jakým uvedené složky spolupracují na vytvoření objektu.)
* neobsahuje žádná „logická slova“ jako jsou výrokové spojky nebo kvantifikátory
* TIL aplikována na analýzu přirozeného jazyka se stává *sémantikou založenou na pojmu* ***možných světů***. Univerzum je chápáno jako množina společná všem možným světům. Vztah označování a vztah vyjadřování je nahrazen jiným schématem.

**Základními rysy TIL systému jsou:**

* schopnost systematicky překračovat omezení platná v predikátové logice 1. řádu
* intezionalizmus a z něho vyplývající schopnost přesného definování intenzí[[1]](#footnote-1) a zacházení s nimi
* **větší expresivní síla**
* má rozvětvenou typovou hierarchii
* temporální
* transparentní – nositel významu (konstrukce) není prvek formálního aparátu, tento aparát pouze studuje konstrukce, zachycení intenzionality je přesně popsáno z matematického hlediska

**Typy v TILu**

* základní typy – typová báze = {ο, ι, τ, ω}, umožňují přiřadit typ objektům z intenzionální báze jazyka – třída zákl. vlastností (barvy, rozměry, postoje, …) popisujících stav světa
	+ **ο (omikron)** – pravdivostní hodnoty Pravda (true, T) a Nepravda (false, F), odpovídají běžným logikám
	+ **ι (jota)** – třída individuí, individua jako numerická identifikace nestrukturované entity
	+ **τ (tau)** – třída časových okamžiků (časová kontinua), zachycení závislosti na čase, třída reálných čísel
	+ **ω (omega)** – třída možných světů, zachycení empirické závislosti na stavu světa
* funkcionální typy – funkce nad typovou bází
* typy vyšších řádů

**Možné světy**

* termín možný svět – **Gottfried Wilhelm von Leibniz** (1646–1716, ﬁlozof a matematik)
* požadavky na deﬁnici “možného světa:”
	+ soubor myslitelných faktů
	+ je konzistentní a maximální ze všech takových souborů
	+ je objektivní (nezávislý na individuálním názoru)
* mezi možnými světy existuje právě jeden aktuální svět – jeho znalost ≡ vševědoucnost
* **možný svět v TILu** = rozhodovací systém, pro každý prvek intenzionální báze obsahuje konzistentní přiřazení hodnot
	+ **** např. realita s 2 objekty a 2 vlastnostmi (9 možných světů):

# 2. Matematické objekty – množiny, relace, uspořádání, posloupnosti, …

Matematický objekt je objekt, který zkoumá matematika; jedná se o prvek materiálu matematického myšlení. Příkladem matematického objektu je například číslo, čtverec, vektor, funkce, matice, tranzitivnost.

**Množiny**

* množina se dá chápat jako soubor prvků **neobsahující duplicity**, **není uspořádaná**
* každá množina obsahuje určitý počet prvků (konečný, nekonečný),
	+ výjimkou je tzv. **prázdná množina** (∅ nebo {}), která neobsahuje žádný prvek
* množinu značíme velkým písmenem (např. **M**), prvek množiny malým písmenem (např. **x**)
* pro zápis množinových objektů používáme:
	+ predikát (∈) „je prvkem“, např. a ∈ b
	+ prázdná množina (∅)
	+ pro zápis množinových objektů používáme složené závorky **{}**
* prvky množiny mohou být i jiné množiny
* množiny mohou být **nekonečné**
* téměř všechny matematické objekty lze definovat jako množiny

**Množinové operace**

* **je prvkem**
	+ o ∅ ∈ **{**∅**}** (= prázdná množina **je prvkem** množiny obsahující prázdnou množinu)
* **je podmnožinou**
	+ o {1,2} ⊆ {1,2,3} (= množina {1,2} je podmnožinou množiny {1,2,3})
	+ o {0,1} ⊆ {0,1}
* **potenční množina**
	+ o Množina všech podmnožin dané množiny
	+ o 𝒫 ({∅}) = {∅,{∅}}
	+ o 𝒫 ({1,2,3}) = {{1,2},{2,3},{1,3},{1},{2},{3},{1,2,3}}
* **průnik, sjednocení**
	+ o A ∩ 𝐵 = {𝑥 | 𝑥 ∈ 𝐴 ∧ 𝑥 ∈ 𝐵 };
	+ o A ∪ 𝐵 = {𝑥 | 𝑥 ∈ 𝐴 ∨ 𝑥 ∈ 𝐵 }

**Čísla**

Přirozeným číslem (ℕ) se v matematice rozumí celé kladné číslo (např. 1, 2, 391, 7 039…). V teorii množin či v informatice se mezi přirozená čísla počítá i nula. Používají se k označení počtu nějakých objektů nebo pro vyjadřování pořadí.

Přirozená čísla jsou definována jako objekty, které splňují jisté formální axiomy (axiom = tvrzení, které se předem pokládá za platné, a tudíž se nedokazuje; základní věta, která se přijímá a bez důkazu se považuje za pravdivou)

Jazyk přirozených čísel zavádí do **predikátové logiky** dva nové symboly:

* ***nulu*** (0)
* ***unární funkční symbol*** S, který interpretujeme jako ***následníka***

Přirozená čísla jsou pak určena množinou axiomů, které nazýváme **Peanova aritmetika**. Tyto axiomy jsou:

* ⴺ*x(x=0) …* existuje nula
* ∀*x(*ⴺ*y(y=S(x))) …* každé číslo *x* má následníka *S(x)*
* ∀*x(*¬*(0=S(x)))* … nula není následníkem žádného čísla
* ∀*x(∀y(S(x)=S(y)⇒x=y))* … různá čísla mají různé následníky

Množinový systém, který splňuje výše uvedené axiomy, nám dá přirozená čísla:

* *0≡*∅… nula je definována jako prázdná množina
* *S(x)≡x*∅*{x}* … následník libovolného čísla je definován jako sjednocení toho čísla (resp. množiny odpovídající tomuto číslu) s jednoprvkovou množinu, která toto číslo (resp. množinu) obsahuje.

Např.:



Pro každé přirozené číslo *n* platí, že *n={0, 1,…,n-1}.*

**Základní operace** na množině přirozených čísel jsou sčítání (+) a násobení (\*):

1. sčítání: a+0=a
a+S(b)=S(a+b)
2. násobení: a\*0=0
a\*S(b)=(a\*b)+a

(Obě definice jsou induktivní)

Tyto operace jsou nezávislé na konkrétní (množinové) realizaci a pracují pouze s operací následníka. Množinovou realizaci přirozených čísel již nemusíme používat, lze pracovat pouze s abstrakcí následníka.

Např. 1+2=1+S(1)=S(1+1)=S(1+S(0))=S(S(1+0))=S(S(1))=S(S(S(0)))=3

Jelikož jsme zatím nedokázali komutativitu sčítání, nemůžeme zaměňovat 1+2 a 2+1, i když teď víme, že výsledek bude stejný.

**Funkce a relace, jejich uzávěry**

Relace

Pojem relace zavádíme proto, že potřebujeme nějakým způsobem vyjádřit, že dvě (případně více) hodnoty jsou spolu svázány, chceme vyjádřit, že některé hodnoty z obou (nebo více) množin mají něco společného.

**Relace** také **tvoří základ pro definici funkcí**.

**Matematický pojem zobecňující vlastnosti vztahů**, jako je „rovnost“, „rovnoběžnost“, „být hlavní město“ nebo „být větší než“.

Jako **relaci** nebo **n-ární relaci** nazveme v [matematice](https://cs.wikipedia.org/wiki/Matematika) **libovolný vztah mezi skupinou prvků jedné nebo více množin.**

Relace je **matematický protějšek pojmu „vztah“.** Prvky v množinách jsou mezi sebou v relaci**.**

**Typy relací**

* **Unární**
	+ **„být mužem“** (operujeme s jedním prvkem)
* **Binární**
	+ **„menší než“** (porovnáváme dva prvky)
* **Ternární**
	+ **rovnice** (**4+5=9** -> součet dvou prvků se rovná třetímu)

Příklady relací:

* **relace identity na množině A**:
	+ Id(A) … binární relace
	+ Id(A) = {(a,a) ∈ A × A | a ∈ A}
* **relace větší nebo rovno na přirozených číslech**:
	+ ≥(N) … binární relace
	+ ≥(N) = {(a,b) ∈ N × N | b ⊆ a}
* **relace plus na přirozených číslech**:
	+ +(N) … ternární relace
	+ +(N) = {(a,b,c) ∈ N × N× N | a+b=c}
	+ a+b=c pak můžeme prohlásit jen za jiný zápis pro (a,b,c) ∈ +(N) a všechny operace na číslech považovat za relace

Relace na malých konečných množinách můžeme přehledně zapisovat tabulkou, kde prvky množiny jsou v záhlaví řádků a sloupců a příslušnost dvojice v relaci značíme jedničkou a nulou v příslušné buňce tabulky. Přehledný je rovněž zápis grafem, kdy mezi prvky množiny kreslíme orientované šipky podle toho, které dvojice jsou v relaci.

**Vlastnosti relací**

* **reflexivita** – vyjadřuje, že **každý prvek je v relaci sám se sebou**. Relace *R* na množině *A* je reflexivní, pokud platí *∀a ∈ A((a,a) ∈ R).*
* **symetrie** – tato vlastnost vynucuje, že ke každé dvojici v relaci existuje symetrická dvojice, tedy taková, kde první a druhý prvek jsou prohozeny. Relace R na množině A je symetrická, pokud platí *∀a,b ∈ A((a,b) ∈ R => (b,a) ∈ R).* Tedy symetrická je každá relace, pokud pro každé *a* a *b* z *A* platí, že pokud *a* je v relaci s *b*, je i *b* v relaci s *a*.
	+ Např. “být narozen ve stejný rok” je symetrická relace, kdežto třeba “menší než” není symetrická
* **antisymetrie** – opak symetrie, říká, že relace neobsahuje žádné dvě symterické dvojice (s výjimkou takových, kde první a druhý prvek jsou stejné). Formálně, relace *R* na množině *A* je asntisymentrická, pokud platí *∀a,b ∈ A((a,b) ∈ R ∧ (b,a) ∈ R => a=b)*
* **tranzitivita** – relace je tranzitivní, pokud ke každým dvěma dvojicím (a,b) a (b,c), které jsou v relaci, existuje v relaci i dvojice (a,c); tranzitivita tedy vyjadřuje jistý typ souvislosti relace. V případě zápisu grafem si tuto vlastnost lze představit tak, že dojdeme-li z nějakého bodu do jiného podle šipek, pak musí existovat zkratka, tedy šipka přímo ze startu do cíle. Formálně, relace *R* na množině *A* je tranzitivní, pokud platí *∀a,b,c ∈ A((a,b) ∈ R ∧ (b,c) ∈ R=>(a,c) ∈ R)*
	+ např. relace “je větší než” a “je rovno” (pokud *a=b* a *b=c*, platí i *a=c*) jsou tranzitivní, rovněž i relace “je podmnožinou”, “je menší nebo rovno”, “dělí”…
* **ekvivalence** je taková relace, která je současně reflexivní, symetrická i tranzitivní
* **uspořádání** je taková relace, která je současně reflexivní, antisymetrická i tranzitivní
* Např. ***identita*** na libovolné množině splňuje všechny uvedené vlastnosti, je tedy ekvivalencí i uspořádáním.
* Např. ***menší nebo rovno*** na přirozených číslech je reflexivní, antisymetrická a tranzitivní, je tedy uspořádáním
* např. ***sedí vedle*** na množině studentů v přednáškové místnosti není tranzitivní, není tedy ekvivalencí ani uspořádáním

**Binární relace**

* podmnožina kartézského součinu dvou množin (A x B)
* její vlastnosti:
	+ **Reflexivní** (**x R x**), *(např. rovnost, dělitelnost)*
	+ **Symetrická** (pokud platí (**x R y**), pak (**y R x**), *(např. sourozenci)*
	+ **Antisymetrická** (pokud (**x R y**) a současně (**y R x**), pak platí **x = y** *(např. „<=, >=“)*
	+ **Tranzitivní** (pokud (**x R y**) a současně (**y R z**), pak (**x R z**), *(např. „menší než“)*

**Binární relace** mezi množinami A a B je podmnožina kartézského součinu A × B (tedy množina uspořádaných dvojic, ve kterých je první položka prvkem množiny A a druhá položka prvkem množiny B. Kartézský součin obsahuje všechny takové kombinace těchto prvků.). Často říkáme „binární relace na množině A“, čímž rozumíme relaci množiny s ní samotnou, tedy podmnožinu kartézského součinu A × A (analogicky můžeme mluvit o *n*-ární relaci na množině A)

Funkce

**Funkce** je v [matematice](https://cs.wikipedia.org/wiki/Matematika) název pro [zobrazení](https://cs.wikipedia.org/wiki/Zobrazen%C3%AD_%28matematika%29) z nějaké množiny M do množiny [čísel](https://cs.wikipedia.org/wiki/%C4%8C%C3%ADslo) (většinou reálných nebo komplexních), nebo do [vektorového prostoru](https://cs.wikipedia.org/wiki/Vektorov%C3%BD_prostor) (pak se mluví o *vektorové funkci*).

Je to tedy předpis, který každému prvku z množiny M jednoznačně přiřadí nějaké číslo nebo vektor (hodnotu funkce). Někdy se však slovo *funkce* používá pro libovolné [zobrazení](https://cs.wikipedia.org/wiki/Zobrazen%C3%AD_%28matematika%29).

Funkce je speciální typ relace, dalo by se říct, že „býti funkci“ je vlastnost relací podobně jako např. tranzitivita. Funkce je taková (*n*-ární) relace, kde prvních *n-1* hodnot v *n*-tici jednoznačně určuje poslední hodnotu.

**Funkce** je předpis, který každému číslu *x* z definičního oboru *M* přiřadí právě jedno *y* z oboru hodnot *N*. Funkci obvykle zapisujeme ve tvaru *y* = *f*(*x*), či ji můžeme vyjádřit explicitně *f*:*y* = *x* kde proměnná *x* je argument funkce.

* U každé funkce musíme také určit její [definiční obor](http://www.matematika.cz/definicni-obor), což je množina všech přípustných hodnot argumentu *x*, tedy všechny hodnoty, kterých může proměnná *x* nabývat. Definiční obor funkce *f* značíme *D*(*f*).
* **Obor hodnot** je poté analogicky množina všech přípustných *y*, tedy množiny všech prvků, kam může ukazovat funkce *f*.

Funkce je vstupně-výstupní mechanismus: prvních *n-1* hodnot v *n*-tici můžeme pokládat za argumenty relace (vstup), poslední hodnotu za její výsledek (výstup). Pokud má být taková relace funkcí, musí být jednoznačně určena argumenty. Z tohoto pohledu vychází i speciální zápis: např. místo *(a,b,c) ∈ +* píšeme často *+(a,b)=c*. Z tohoto pohledu na funkce vychází i konvence ohledně arity funkcí, která se určuje podle počtu argumentů. Tedy unární funkce je binární relace, ternární funkce je binární funkce atd.

Formálně, binární relace *f* na množině *A* je funkce, pokud: *∀a,b,c ∈ A((a,b) ∈ f ∧ (a,c) ∈ f => b=c).*

Obdobně, ternární relace *f* na množině *A* je funkce, pokud: *∀a,b,c,d ∈ A((a,b,c) ∈ f ∧ (a,b,d) ∈ f => c=d).*

Funkcím tedy někdy říkáme **zobrazení**. Pro binární relaci *f* mezi množinami *A* a *B*, která je funkcí, říkáme “funkce z A do B”, případně “zobrazení z A do B” a zapisujeme: *f : A* →*B*. Množinu *A* potom nazýváme **definičním oborem**a někdy značíme *Df* nebo *dom(f)*, množinu *B* nazýváme **oborem hodnot** a značíme *Rf* nebo *f(A)*. Zápis *f(a) = b* čteme jako “*b* je obraz prvku *a*”, případně “*a* je vzor prvku *b*”.

**Funkce**

* speciální typ relace
* zápis: **f: A→ B** (zobrazení A do B)
* A = definiční obor, B = obor hodnot
* **f (a) = b** (b je obraz prvku a, a naopak)

**Vlastnosti funkcí**:

* **injektivita**: Funkce *f: A* →*B* je injektivní (též prostá), pokud platí: *∀a,b ∈ A(f(a) = f(b) =>a=b),* neboli **žádné dva prvky nemají stejný obraz**
* **surjektivita**: Funkce *f: A* →*B* je surjektivní (též na), pokud platí: *∀b ∈ B(∃a ∈ A (b=f(a)),* neboli **každý prvek oboru hodnot má nějaký vzor**, případně můžeme říct, že celý obor hodnot je pokrytý
* **úplnost**: Funkce *f: A* →*B* je úplná, pokud platí: *∀a ∈ A (∃b ∈ B (b=f(a)),* neboli celý definiční obor je pokrytý (často se pojmem *funkce* označuje právě úplná funkce)
* **bijekce**: Funkce je bijekce právě tehdy, je-li současně injektivní, surjektivní a úplná
	+ (úplnost, prosté i zobrazení „na“)
	+ každý prvek z cílové množiny má právě jeden vzor

Posloupnosti a řady

### Definice

* funkce, jejímž definičním oborem **D** jsou buď
	+ **všechna přirozená čísla** () – nekonečná posloupnost
	+ nebo libovolná **podmnožina přirozených čísel** (*n*) – konečná posloupnost
* každému **n ∈ N** přiřazuje číslo **an ∈ R**
* záleží na pořadí prvků
* prvky se mohou v posloupnosti opakovat
* posloupnost se tak vyjadřuje tím, že její prvky jdou nějak **uspořádat**, že každý prvek posloupnosti, kromě prvního a posledního, má svého předchůdce a svého následovníka.
* uspořádaný (konečný nebo nekonečný) soubor matematických objektů
* příklad posloupnosti: 1, 3, 5, 7, 9, 11, …

### Zápis posloupnosti

* posloupnost obvykle pojmenováváme malým písmenem s dolním indexem – *an*
* **výčet všech hodnot posloupnosti** (může být konečný i nekonečný)

*ak = (1, 2, 3, 4, 5)*

*an = (2, 4, 6, 8, 10, …)*

* **vzorec pro n-tý člen posloupnosti**

*an = 2n* (všechna sudá čísla)

* + dolní index n nám určuje, kolikátý prvek posloupnosti zrovna počítáme, např. výpočet sedmého prvku:

*a7 = 2 × 7 = 14*

* **rekurentní definice (induktivní definice)**
	+ **rekurze** = definování objektu pomocí sebe sama; návrat, zpětný běh
	+ umožňuje vypočítat následující člen, pokud známe ten současný (známe-li prvek *an*, umožňuje nám spočítat prvek *an+1*)
	+ vždy zadáváme první prvek (je možné zadat i více prvků) a předpis pro vypočítání *n-*tého prvku pomocí jednoho nebo několika předchozích prvků, např.

*a1 = 2*

*an+1 = an + 2*

* + hodí se pro výpočet větší části posloupnosti
	+ nehodí se pro výpočet konkrétního prvku (pro výpočet 75. členu musíme znát všech 74 předcházejících)
	+ průběh výpočtu

*a1 = 2*

*a2 = a1 + 2 (=4)*

*a3 = a2 + 2 (=6)*

*a4 = a3 + 2 (=8)*

*…*

##### Příklad: Fibonacciho posloupnost

* nekonečná posloupnost přirozených čísel začínající 0, 1, 1, 2, 3, 5, 8, 13, 21, …
* každé číslo je součtem dvou předchozích
* rekurzivní definice:

*a0 = 0*

*a1 = 1*

*an = an−1 + an−2*

### Graf posloupnosti

* grafem posloupnosti je vždy **množina izolovaných bodů**



Obr. 1 Graf posloupnosti an = 1/n

### Speciální druhy posloupností

##### Aritmetická posloupnost

* mezi jednotlivými členy posloupnosti stálý rozdíl – **diference (*d*)**
* každý člen posloupnosti (kromě prvního) je aritmetickým průměrem „svých sousedů“
* obecný zápis: *an + 1 = an + d*
* výpočet *n*-tého členu: *an = a1 + (n - 1)d*
* příkladem je posloupnost sudých čísel

##### Geometrická posloupnost

* dva sousední členy nemají stejný rozdíl, ale podíl – **kvocient (*q*)**
* obecný zápis: *an + 1 = an × q*
* výpočet *n*-tého členu: *an = a1 × qn-1*
* například posloupnost, kde q = 10 a první prvek je 5:

*a1 = 5*

*a2 = a1 × 10 (=50)*

*a3 = a2 × 10 (=500)*

*a4 = a3 × 10 (=5000)*

*…*

### Příklady

1) Induktivně definuj nekonečnou posloupnost 1, 3, 5, 7, 9, …

Řešení: an = an-1 + 2, a1 = 1

2) Induktivně definuj nekonečnou posloupnost 2, 5, 11, 23, 47, …

2) an = 2an-1 + 1, a1 = 2

Opakování

**Množina**

* množina se dá chápat jako soubor prvků **neobsahující duplicity**, **není uspořádaná**
* množina obsahuje určitý počet prvků, popř. může být prázdná nebo nekonečná

**Relace**

* vyjádření toho, že dvě (nebo více) hodnoty jsou spolu svázány; hodnoty z obou množin mají něco společného
* **matematický pojem zobecňující vlastnosti vztahů**, jako je „rovnost“, „rovnoběžnost“, „být hlavní město“ nebo „být větší než“.
* jako **relaci** nebo **n-ární relaci** nazveme v [matematice](https://cs.wikipedia.org/wiki/Matematika) **libovolný vztah mezi skupinou prvků jedné nebo více množin.**
* relace je **matematický protějšek pojmu „vztah“.** Prvky v množinách jsou mezi sebou v relaci**.**
* vlastnosti
	+ *reflexivita*: každý prvek je v relaci sám se sebou
	+ *symetrie:* symetrická dvojice, prvky mezi sebou lze prohodit
	+ *antisymetrie:* relace neobsahuje symetrickou dvojici, nelze je prohodit
	+ *tranzitivita:* vyjadřuje jistý typ souvislosti relace (např. je rovno; a = b, b = c; pak platí i to, že a = c)
* **ekvivalence:** relace, která je současně reflexivní, symetrická a tranzitivní
* **uspořádání:** relace, která je současně reflexivní, antisymetrická a tranzitivní (např. menší nebo rovno)

**Funkce**

* funkce je předpis, který každému číslu *x* z definičního oboru *M* přiřadí právě jedno *y* z oboru hodnot *N*. Funkci obvykle zapisujeme ve tvaru *y* = *f*(*x*) či ji můžeme vyjádřit explicitně *f*:*y* = *x* kde proměnná *x* je argument funkce.
	+ u každé funkce musíme také určit její [**definiční obor**](http://www.matematika.cz/definicni-obor), což je množina všech přípustných hodnot argumentu *x*, tedy všechny hodnoty, kterých může proměnná *x* nabývat. Definiční obor funkce f značíme D(f)
	+ **obor hodnot** je poté analogicky množina všech přípustných *y*, tedy množiny všech prvků, kam může ukazovat funkce *f*

**Posloupnost**

* funkce, jejímž definičním oborem **D** jsou buď
	+ **všechna přirozená čísla** () – nekonečná posloupnost
	+ nebo libovolná **podmnožina přirozených čísel** (*n*) – konečná posloupnost
* **záleží na pořadí prvků**
* prvky se mohou v posloupnosti opakovat
* posloupnost se tak vyjadřuje tím, že její prvky jdou nějak **uspořádat**, že každý prvek posloupnosti, kromě prvního a posledního, má svého předchůdce a svého následovníka.
* uspořádaný (konečný nebo nekonečný) soubor matematických objektů

# 3. Statistika a pravděpodobnost – statistický soubor a jeho charakteristiky, pravděpodobnostní prostor

Statistika

* statistika je věda, která zkoumá, zpracovává a vyhodnocuje data
* cílem statistiky je provést experiment, jehož účelem je zjistit něco zajímavého o dané populaci
	+ **populací** (též **základní soubor**) se myslí obecně jakýkoliv soubor prvků, které chceme zrovna zkoumat. O tomto souboru většinou chceme shromažďovat údaje a vyvozovat z něj různá fakta
	+ jelikož obvykle pracujeme jen s omezeným výběrem objektů ze základního souboru, právě tomuto výběru říkáme **statistický soubor**
* objektem zkoumání matematické statistiky je **statistický soubor**, což chápeme jako posloupnost údajů (většinou číselných nebo vyjádřených číselně) o nějakých objektech
	+ typy těchto údajů někdy nazýváme **statistické znaky** a jejich počet pak určuje **rozměr statistického souboru**
* aby vybraný vzorek správně vypovídal o základním souboru, výběr by měl být reprezentativní. Pro nedostatek informací o základním souboru většinou spoléháme na **náhodný výběr** a na to, že náhoda nám zaručí reprezentativnost vzorku.

**Jednorozměrný statistický soubor:** statistický soubor omezený pouze na jeden statistický znak (např. hmotnost slonů). **Dvourozměrný statistický soubor:** více statistických znaků (např. výška a hmotnost slonů).

Jazykové korpusy jsou soubory dat o přirozeném jazyce. V dnešní době jsou obrovské (až desítky miliard slov). Využití statistiky pro popis jevů v jazyce je vhodný způsob.

Rozložení četnosti statistického souboru

Např. zvážili jsme šest slonů a naměřili hodnoty 2, 4, 4, 4, 5 a 11 tun.

Statistický soubor pak bude šestice (2, 4, 4, 4, 5, 11)

* **rozsah statistického souboru**: počet prvků statistického souboru (v našem případě 6)
* **absolutní četnost hodnoty**: počet výskytu konkrétní hodnoty v daném statistickém souboru (v našem případě je absolutní četnost prvku čtyři 3)
* **relativní četnost hodnoty**: absolutní četnost podělená rozsahem souboru, udává se obvykle v procentech (např. relativní četnost hodnoty 4 je 50 %)
* **kumulativní četnost hodnoty**: četnost hodnoty souboru plus četnost všech menších hodnot. Opět rozeznáváme absolutní i relativní kumulativní četnost hodnoty. (v našem případě kumulativní absolutní četnost hodnoty 4 je 4, kumulativní relativní četnost hodnoty 4 je 66 %)

Četnosti můžeme zaznamenat do sloupcového grafu – graf funkce, který má na ose *x* jména tříd a na ose *y* jejich absolutní nebo relativní četnosti, se nazývá **histrogram**.

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 (např. jak těžký je běžný slon)
* **aritmetický průměr** (*avg*) – součet hodnot ve statistickém souboru podělený velikostí souboru (Pokud bychom uvažovali sloupce v histogramu jako závaží, pak průměr je těžištěm histogramu. Pro náš soubor je průměr 5)
* **modus** – hodnota či třída s největší četnosti (v našem ukázkovém souboru je modus 4)
* **medián** – “prostřední” hodnota v souboru po jeho setřídění. V případě, že datový soubor má sudý počet prvků, je to průměr ze dvou prostředních. (v našem případě je medián 4). Medián není citlivý na extrémní odchylky jako průměr – např. relativně málo hodně vysokých hodnot mže průměr vychýlit i hodně nahoru, zatímco u mediánu se neprojeví vůbec

Charakteristiky variability:

* ukazují, jak je statistický soubor vnitřně konzistentní, čili jak moc se od sebe vzájemně liší hodnoty obsažené v souboru
* **rozptyl** (disperze, variace – značí se *s2*) – je definován jako: Jedná se tedy o aritmetický průměr druhých mocnin odchylek jednotlivých hodnot od průměru. Pokud je rozptyl 0, všechny hodnoty v souboru jsou stejné. Čím je větší rozptyl, tím více se jednotlivé hodnoty od sebe liší.
	+ rozptyl nám udává, jak moc jsou hodnoty v našem statistickém soubory rozptýleny. Rozptylu se někdy též říká variance.
* **směrodatná odchylka** (značí se *s*) – je to odmocnina z rozptylu, vyjadřuje totéž, ale jiným číslem
	+ směrodatná odchylka, podobně jako rozptyl, určuje, jak moc jsou hodnoty rozptýleny či odchýleny od průměru hodnot.

Korelace

* vzájemný vztah mezi dvěma procesy nebo veličinami. Pokud se jedna z nich mění, mění se **korelativně** i druhá a naopak
* míru korelace pak vyjadřuje **korelační koeficient**, který může nabývat hodnot od −1 až po +1
* pojmem korelace rozumíme **stupeň lineární závislosti** znaků *x* a *y*, tedy to, do jaké míry hodnoty znaku *x* lineárně závisí na hodnotách znaku *y*. Tedy to, jak dobře lze grafem závislosti *x* na *y* proložit přímku

Např. data o výšce a hmotnosti slonů je **korelace** statistických 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)
	+ hodnota korelačního koeficientu +1 značí **zcela přímou závislost**, např. vztah mezi rychlostí bicyklu a frekvencí otáček kola bicyklu
* pokud je korelace **-1**, jedná se o **nepřímou úměrnost** (čím větší je *x*, tím menší je *y* a hodnoty *y* lze z hodnoty *x* získat jednoduše vynásobením nějakou zápornou konstantou)
	+ hodnota korelačního koeficientu -1 značí **zcela nepřímou závislost** (**antikorelaci**), tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků, např. vztah mezi uplynulým a zbývajícím časem

**Statistické testování hypotéz**

* chceme zjistit, zda naše statistická dat potvrzují nějakou hypotézu
* vyjádření pravděpodobnosti toho, že dané výsledky vznikly náhodně, je základem metodiky testování hypotéz
* ***nulová hypotéza*** – H0, předpoklad, který obvykle platí a který chceme našimi statistickými daty vyvrátit
* ***alternativní hypotéza*** – H1, tvrzení, pro které hledáme oporu v našich datech, doplňková k H0
* ***chyba typu I*** – pokud potvrdíme H1, neboli vyvrátíme nulovou hypotézu, ale přitom platí H0
* ***chyba typu II*** – pokud nepotvrdíme H1 a ta je přitom platná, není tak závažná
* to, že se nepodařilo prokázat souvislost, neznamená, že souvislost neexistuje

## Pravděpodobnost

* pravděpodobnost nějakého náhodného jevu nám udává, jakou máme šanci, že daný jev nastane
* pravděpodobnost nějakého jevu je definována jako podíl *m/n*, kde *m* je počet možností, kdy daný jev nastal, a *n* je počet všech možností, které potenciálně nastat mohly či mohou

## Pravděpodobnostní rozložení (pravděpodobnostní rozdělení, pravděpodobnostní distribuce, probability distribution, pravděpodobnostní funkce)

* udává, jaká je pravděpodobnost, že náhodná proměnná bude mít zrovna danou hodnotu.
* funkce, která pro jednotlivé možné hodnoty vrací pravděpodobnost, s jakou vlastnost A nabude této hodnoty
	+ *p:X* → [0,1]
		- X – množina možných hodnot příslušné vlastnosti
		- [0,1] – uzavřený interval od nuly do jedné
		- ∀x∈X (p(x) ≤ 1 ∧ p(x) ≥ 0)
		- ∑x∈Xp(x)=1
			* součet hodnot funkce pro všechny možné hodnoty vlastnosti X je 1
		- p(x) = P(A = x)
			* hodnota funkce odpovídá pravděpodobnosti, že vlastnost A nabude hodnoty x
* relativní četnosti ve statistickém souboru odpovídají hodnotám pravděpodobnostního rozložení v pravděpodobnostním prostoru.
* **pravděpodobnostní prostor**
	+ (X,p) – množina všech možných hodnot vlastnosti spolu s pravděpodobnostním rozložením

Typy pravděpodobnostních 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 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.:



„Normální” se jmenuje proto, že dobře vystihuje velké množství vlastností různých přírodních populací – např. rozložení výšek a hmotností zvířat či lidí.

**Zipfovo rozložení** je typické tím, že několik málo nejčastějších hodnot má velkou pravděpodobnost a s každou další hodnotou (při setřídění od nejčastější) tato pravděpodobnost prudce klesá – v ideálním Zipfově rozložení má nejčastější hodnota pravděpodobnost *𝑥*, druhá *𝑥/*2, třetí *𝑥/*3 atd. Graf takového rozložení nejčastěji kreslíme tak, že hodnoty na ose x jsou uspořádány podle jejich četnosti – graf potom tvoří prudce klesající křivku, jako je např. tato:



Toto rozložení velmi často výstižně popisuje nejrůznější frekvenční seznamy – např. pokud vytvoříme pravděpodobnostní rozložení velikostí světových měst, dojdeme ke křivce Zipfova rozložení. Stejnou křivku dostaneme, pokud vytvoříme pravděpodobnostní rozložení slov nebo jejich morfologických kategorií v jazyce.

**Zipfův zákon**

Skutečnost, že Zipfovo rozložení výstižně popisuje opravdu velké množství jevů, 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. Při jakémkoli zpracování jazyka je tedy třeba s ním počítat.

Pro ilustraci uveďme 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 (u těchto slov se můžete setkat s technickým označením *stoplist*). Zipfův zákon platí dokonce i pro „náhodně generovaný” jazyk – pokud budeme náhodně generovat znaky včetně mezery, pravděpodobnostní rozložení takto vygenerovaných „slov” bude odpovídat Zipfovu rozložení. Zipfův zákon platí ale i v mnoha oblastech mimo přirozený jazyk – již jsme zmínili např. počet obyvatel měst; dále jej lze aplikovat např. na velikost platů, počet zaměstnanců společností a mnohá další. Rovněž známé ekonomické pravidlo „80 : 20” – 80 % problému vyřešíme s 20% úsilím – je vlastně

alternativní formulací Zipfova zákona.

**Zákon velkých čísel** říká to, že aby výsledky založené na statistických metodách byly kvalitní a reprezentativní, potřebujeme je odvozovat od velkých statistických souborů, neboli čím větší počet pokusů provedeme, tím větší je pravděpodobnost, že průměr výsledků pokusů bude odpovídat očekávanému vypočtenému průměru.

**Distribuční funkce**

**Distribuční funkce** (cumulative distribution function), často značena *𝐹*, je pravděpodobnost, že náhodná veličina nabude určité hodnoty nebo menší, čili *𝐹*(*𝑥*) = *𝑃*(*𝐴* ≤ *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** chápeme jako posloupnost náhodných veličin, např. náhodný vektor „počasí v Brně zítra v poledne” můžeme chápat jako trojici (*𝑡𝑒𝑝𝑙𝑜𝑡𝑎, 𝑡𝑙𝑎𝑘, 𝑣𝑙lh𝑘𝑜𝑠𝑡*). Jeho pravděpodobnostní rozložení můžeme modelovat s využitím vícerozměrného statistického souboru.

## Podmíněná pravděpodobnost

* často máme kromě pravděpodobnostního rozložení daného jevu další informace, např. o jiném jevu, který s původním může a nemusí souviset
* pravděpodobnost jednoho jevu za předpokladu, že nastal jev jiný
* P(A|B)
	+ pravděpodobnost jevu A za předpokladu, že nastal jev B
* P(A|B)= P(A,B)/P(B)
	+ P(A,B) – pravděpodobnost, že nastaly oba jevy současně
* **nezávislé jevy**
	+ informace o jednom jevu nám nepřinese žádnou informaci o jevu druhém
	+ P(A|B)=P(A) ˄ P(B|A)=P(B)
	+ platí: P(A,B) = P(A)\*P(B)
	+ v reálném životě se opravdu nezávislé jevy vyskytují jen zřídka
	+ dvě mince → co padne na jedné, neovlivní to, co padne na druhé
* závislé jevy spolu mohou, ale nemusí, kauzálně souviset
* *Bayesův vzorec* – někdy nemáme přístup přímo ke statistickému souboru, z něhož vycházíme, ale známe některé pravděpodobnosti z něj odvozené, obecně chceme převádět P(A|B) na P(B|A)

Pokud známe pravděpodobnosti *𝑃*(*𝐴*) a *𝑃*(*𝐵*), můžeme použít tzv. **Bayesův vzorec**:

$$P \left(B\right)= \frac{P\left(A\right)\*P(A)}{P(B)}$$

**Entropie**

* entropie náhodné veličiny je míra informace obsažená v této náhodné veličině
* vyjadřuje, kolik informace získáme, když se dozvíme skutečnou hodnotu náh. veličiny
* je nezáporná a měří se v bitech
* ***nulová entropie*** – hodnotu náhodné veličiny jsme schopni určit se stoprocentní jistotou
* čím vyšší je entropie, tím více informace je v náhodné veličině obsaženo a tím více jednotek informace je třeba k přenesení výsledku náhodného pokusu
* ***podmíněná entropie*** – míra informace obsažená v náhodné veličině za předpokladu, že známe hodnoty jiné náhodné veličiny
* ***mutual information*** – vzájemná informace, určuje míru informace, kterou jedna náhodná proměnná říká o jiné, pokud jsou veličiny nezávislé, je rovna nula, čí je vyšší, tím hodnoty jedné vlastnosti určují hodnoty druhé vlastnosti
	+ ve statistickém zpracování jazyka lze použít k měření síly kolokace – čím vyšší je počet souvýskytů dvou slov v korpusu, tím je vyšší, čím častěji jsou jednotlivá slova, tím více se snižuje

## Kombinatorika

* **kombinatorika** (**kombinatorická matematika**) je část [matematiky](https://cs.wikipedia.org/wiki/Matematika) zabývající se kolekcemi prvků [množin](https://cs.wikipedia.org/wiki/Mno%C5%BEina) s definovanou vnitřní strukturou
* otázky, které kombinatorika řeší, se obvykle týkají počtu nějakých objektů (nebo skupin objektů) s definovanou strukturou, speciálně (pokud počet může být nulový) existencí objektu s definovanou strukturou
* zkoumá skupiny (podmnožiny) prvků vybraných z jisté základní množiny. Podle toho, zda se prvky v jednotlivých skupinách **mohou či nemohou opakovat**, rozdělujeme skupiny prvků na **skupiny s opakováním a skupiny bez opakování**
* **variace:** z nějaké množiny objektů vybíráme určitý počet objektů, přičemž záleží na pořadí, v jakém tyto objekty vybíráme
	+ permutace: e speciálním případem variace, kdy máme množinu o velikosti *n* a chceme zjistit počet všech různých *n*-tic.
* **kombinace:** z nějaké množiny objektů vybíráme určitý počet objektů, přičemž nezáleží na pořadí, v jakém tyto objekty vybíráme. Typickou úlohou je Sportka – v osudí máme 49 čísel a taháme z nich 6 čísel, přičemž je jedno, v jakém pořadí je vytáhneme.
* **pravidlo součtu**
* **pravidlo součinu**

# 4. Aplikace statistiky na zpracování jazyka

* je umožněno rozsáhlými jazykovými korpusy, můžeme modelovat a predikovat chování jazyka
* základním statistickým údajem, který z korpusu můžeme získat je samotná frekvence, lze ji určit globálně (z celého korpusu) nebo lokálně (z konkordance)

**Četnosti/frekvence jevů**

* běžná četnost nemusí být vždy nejlepším vodítkem, problémy tvoří jevy, které se v korpusu vyskytují mnohokrát, ale pouze lokálně (např. v jediném dokumentu)
* ***logaritmická četnost*** – předpokládá, že korpus je rozdělen do malých částí (dokumentů) a četnost je počítána ve vztahu k těmto dokumentům
* ***redukovaná četnost*** – vázanost na dokumenty považuje za nevýhodu, nemusí totiž mít vždy stejné velikosti i u odpovídajících textů
	+ slova, která se vyskytují v korpusu pouze v omezeném intervalu pozic, by měla mít nižší redukovanou četnost než slova se stejnou absolutní četností vyskytující se v korpusu rovnoměrně
	+ pokud se slovo vyskytuje v korpusu rovnoměrně, měla by se redukovaná četnost rovnat absolutní
* obě veličiny se dají použít pro eliminaci nerovnoměrného rozdělení slov korpusu

**Uplatnění Zipfova zákona**

* součet četností a pořadí v seznamu četností (ať už slov, hlásek, …) je zhruba konstantní
* ve frekvenčním seznamu slov korpusu si např. můžeme všimnout, že druhé nejčastější slovo má zhruba poloviční četnost než první, třetí pak třetinovou atd. → četnost dramaticky klesá, až ke konci zůstává velké množství slov s téměř stejnou nízkou frekventovaností
* nejfrekventovanější jevy pokrývají většinu jazyka, zbývajících je sice veliké množství, ale vyskytují se velmi málo

**Vyhledávání kolokací**

* výraz, který se skládá z dvou nebo více slov – jmenné skupiny, slovesné skupiny, ustálené fráze
* **pro identifikaci v korpusu se používají statistické metody**
	+ klíčový parametr – počet výskytů bigramu v korpusu (můžeme zvolit, zda musí stát hned vedle sebe, nebo se slova mohou vyskytovat v nějakém okně)
* využití ***stoplistů*** (znevýhodnění spojení typu "of the" nebo "in the")
* ***T-score***
	+ aplikace statistické metody testování hypotéz pomocí tzv. t-testu
	+ testujeme, jestli zjištěné počty výskytů jednotlivých slov a jejich bigramu odpovídají náhodnému rozložení slov v korpusu
	+ čím vyšší je vypočtená hodnota, tím méně je pravděpodobné náhodné rozložení slov a tedy jde o kolokaci
	+ ***nulová hypotéza***– slova se chovají náhodně podle svých obvyklých pravděpodobnostních rozložení
	+ pokud je nulová hypotéza vyvrácena (slova se vyskytují spolu častěji než náhodně), je dvojice kolokací
	+ nefunguje ovšem dobře pro málo frekventované jevy
	+ T-score často říká, co už známe, má tendenci vyzdvihovat běžné syntaktické vzorce
	+ vyžaduje použití stoplistů na předložky, zájmena apod.
* ***mutual information (MI-score)***
	+ vychází z teorie informace, porovnává míru "překvapení", že slova vidíme zaráz
	+ vyjadřuje množství informace poskytované výskytem slova *y* o výskytu slova *x*
	+ pravděpodobnost odhadujeme z korpusu jako počet výskytů dělený velikostí korpusu
	+ je citlivé na četnost jednotlivých slov, nejvyšších hodnot dosahují dvojice slov, která jsou v korpusu málo častá
		- řešení – nastavují se spodní hranic četnosti slova, slova pod hranicí se vůbec neuvažují
	+ výpočet v konkordanci – první slovo je slovo klíčové, druhé sovo je libovolné slovo nacházející se v zadaném kontextu (okně)
	+ s větší pravděpodobností řekne něco neobvyklého, čeho bychom si nemuseli všimnout
* ***log-likelihood***
	+ modifikace MI-score
	+ porovnává dvě hypotézy, které vysvětlují četnost výskytu určitého bigramu
		- první je formalizací závislosti, která potvrzuje kolokaci
		- druhá je formalizací nezávislosti
	+ je jasněji interpretovatelné a vhodnější pro řídká data
* ***minimum sensitivity***
	+ minimum z relativních četností dvou slov
	+ každý prvek je nějak zajímavý, zjišťuje "zajímavost" slova nejen v rámci kolokace
* ***dice***
	+ vylepšuje minimum sensitivity
	+ závislá pouze na frekvencích slov a na frekvenci bigramu z těchto slov, do výpočtu nevstupuje velikost korpusu
	+ základem je poměřování frekvence bigramu s průměrem frekvencí slov daného bigramu
	+ frekvence bigramu nemůže být nikdy větší než průměr hodnot frekvencí obou jeho konstituentů, hodnoty dice se proto pohybují v intervalu (0,1)
* ***logdice***
	+ normalizovaná varianta *dice*
	+ nabývá hodnot od mínus nekonečna do 14

**Statistika** – umožňuje uchopit a popsat velké množství dat → vyvozovat další informace
**Statistické zpracování přirozeného jazyka** – k dispozici velké soubory jazykových dat (korpusy +10mld slov), které umožňují aplikovat statistické metody, predikovat chování přirozeného jazyka a modelovat jej

**Využití statistiky**:

a) **Vyhledávání kolokací** – kolokace = statisticky významné spojení dvou slov v korpusu, lze určit síla kolokace libovolných dvou slov

Možné způsoby:

* 1. uvažovat při vyhledávání kolokací pouze dvojice složené z určitých slovních druhů
* 2. Statistické metriky, např. T-test – testování hypotéz – Nulová hypotéza – pokud je vyvrácena (slova se spolu vyskytují častěji než náhodně), jde o kolokaci

b) **N-gramové jazykové modely** (viz výše); nedostatky jazykových modelů: jazykové modely mohou být velké → výpočetně náročné X malé *n* zase neposkytuje dostatečný kontext pro kvalitní odhady a *data sparseness* (řídkost dat) – pro slova a n-gramy, které nejsou tak časté → dostaneme nekvalitní odhady pravděpodobností

Ke zpracování přirozeného jazyka existují dva přístupy:

* **pravidlový** – občas kriticky selhává, jestliže jsme na něco nevymysleli pravidlo, je potřeba mnoho pravidel, těžko je možné vymyslet pravidla na všechno – přirozený jazyk je vágní, mnohoznačný a vyvíjí se, při tomto přístupu se používají formální gramatiky
* **statistický** – selhává často, ale jen málo, je potřeba mnoho dat ke statickému zpracování (čím víc dat, tím lepší výsledky)

# 5. Statistický jazykový (n-gramový) model – konstrukce, využití

* ***n-gram*** – posloupnost slov délky *n*, která se vyskytuje v korpusu
* n-gramový jazykový model se skládá **z podmíněných pravděpodobností**, že se vyskytlo slovo wn za předpokladu, že se před ním vyskytla slova w1, …, wn-1
* uplatňují se v rozpoznávání řeči, morfologickém značkování, syntaktické analýze i strojovém překladu
* umožňuje odhadnout nejpravděpodobnější sekvence symbolů, které odpovídají danému (víceznačnému) vstupu
* ***nedostatky jazykových modelů***
	+ mohou být velké – pro větší *n* dostaneme tolik různých pravděpodobností, že počítačová práce s takovým modelem je výpočetně velmi náročná
	+ malé *n* neposkytuje dostatečný kontext pro provádění kvalitních odhadů
	+ *data sparseness* – řídkost dat, pro slova a n-gramy, které se nevyskytují tak často dostaneme nekvalitní odhady pravděpodobností

**Statistický jazykový model** přiřazuje pravděpodobnost posloupnosti slov pomocí rozložení pravděpodobnosti. Přiřazuje každé posloupnosti pravděpodobnost, s jakou je daná posloupnost větou určitého jazyka. Tyto pravděpodobnosti se získávají statistickým zpracováním jazykových dat. Může být unigramový (prosté frekvenční distribuce), bigramový, trigramový, obecně n-gramový (aplikuje podmíněné pravděpodobnosti), cache language model a další.

**N-gramový model** se formálně skládá z podmíněných pravděpodobností P (Wn | W1, …, Wn-1), tzn. pravděpodobností, že se vyskytlo slovo Wn za předpokladu, že před ním se vyskytla slova W1, …, Wn-1.

Souhrn takovýchto pravděpodobností pro všechny možné kombinace slov v korpusu (tento souhrn snadno odvodíme z textu korpusu) se nazývá **n-gramový jazykový model**.

W1, …, Wn-1 nemusí být pouze slova – můžeme vytvořit n‑gramový model znaků, fonémů, pádů, morfologických značek apod. Využití v rozpoznávání řeči, strojovém překladu, word sense disambiguation atd.

S rostoucím *n* rychle přibývá různých n-gramů – větší kontext → vyšší úspěšnost, ale zvyšuje se výpočetní náročnost práce s modelem a náchylnost k data sparseness (řídkost dat, znamená nedostatek dat, týká se zejména okrajových jevů).

**Cache language model** – cache = LIFO, využití například při rozpoznávání mluvené řeči, předpokládáme, že nejpravděpodobnější je to slovo, které bylo vyřčeno před chvílí, např. vyprávíme o slonech, samo slovo není tak časté, podle n-gramového modelu by dostaly přednost častější n-gramy, pokud ale slovo vyslovíme, opravíme na „slon“, bude mu dále přisuzována vyšší pravděpodobnost.

**N-gramové jazykové modely –** hledáme další slovo (značku) na základě předchozích, pravděpodobnostní rozložení P(wn | w1 ,…, wn-1), z dat odvodíme pravděpodobnostní rozložení všech možných wn ; použití – strojový překlad, morfologické značkování, rozpoznávání řeči…; problémy – pro N>4 většinou výpočetně nezvládnutelné, *data sparseness* – pro slova, která se vyskytují méně často, není dost dat –> špatný model

**Bigramy** jsou tak v základu zkoumání kolokací (ačkoli i kolokace mohou být více než dvouslovné). Základní odlišnost bigramu od kolokace je fakt jeho neustálené povahy a neprovázanosti členů (každá kolokace je zároveň bigramem, ne každý bigram je ovšem kolokací, srov. nejčastější bigram *jak se*).

**Trigramy**, tetragramy apod. vznikají zřetězením více slov, jejich frekventovaný souvýskyt značí ustálenou jednotku, jejíž význam může vyjadřovat celou propozici (*zdálo se mi, že*; *jsem si myslel, že*; *podíval jsem se na*; *to je v pořádku* apod.).

Např. P(the/in) = 0,6

 „Pravděpodobnost výskytu *the* za předpokladu, že předchází *in*“

*Máma mele maso. Ema mele maso. Maso se mele.*

Bigramový model na znacích – n=2; text se projíždí po dvojicích znaků

P(a/m) = 6/9 = 2/3 (6x je v textu a, 9x je v textu m)

P(e/m) = 3/9 = 1/3

P(m/a) = 3/6

# 6. Vyhodnocování aplikací zpracování jazyka – přesnost, pokrytí, F-míry a zlatý standard, křížová validace

* ***precision (přesnost) –*** procento položek, které systém určil správně
* ***recall (pokrytí)*** – procento položek, který systém zahrnul do výsledku
* přesnost a pokrytí je vhodné někdy kombinovat do jediné míry → ***F-míry*** – harmonický průměr přesnosti a pokrytí
* ***zlatý standard*** referenční vyšetřovací metoda pro danou diagnózu, ideální vyšetření, které jednoznačně rozdělí populaci (např. na nemocné a zdravé)
* ***křížová validace*** – metoda zjišťování, jak moc bude model statistické analýzy ovlivňovat nezávislé vzorky dat
	+ postup je významný pro predikci neznámých vzorků po předchozí klasifikaci známých vzorků
	+ princip – vstupní množina dat je rozdělena na podmnožiny. Jedna podmnožina slouží jako testovací množina, zbylé podmnožiny slouží jako trénovací množiny. Klasifikátor natrénuje model na trénovací množině a pomocí testovací množiny testuje přesnost a výkonnost tohoto modelu. Tento proces se několikrát opakuje, pokaždé s jinou podmnožinou tvořící trénovací a testovací množinu.
1. **intenze** neboli smysl pojmu je jeho obsah, to, co se pojmem míní, jeho hlavní či ústřední význam [↑](#footnote-ref-1)