

Základy matematiky a statistiky pro humanitní obory II

Pavel Rychlý Vojtěch Kovář

Fakulta informatiky, Masarykova univerzita
Botanická 68a, 602 00 Brno, Czech Republic

{pary, xkovar3}@fi.muni.cz

6. 4. 2011

Obsah přednášky

- 1 Typy pravděpodobnostních rozložení
- 2 Zipfův zákon
- 3 Zákon velkých čísel
- 4 Testování hypotéz

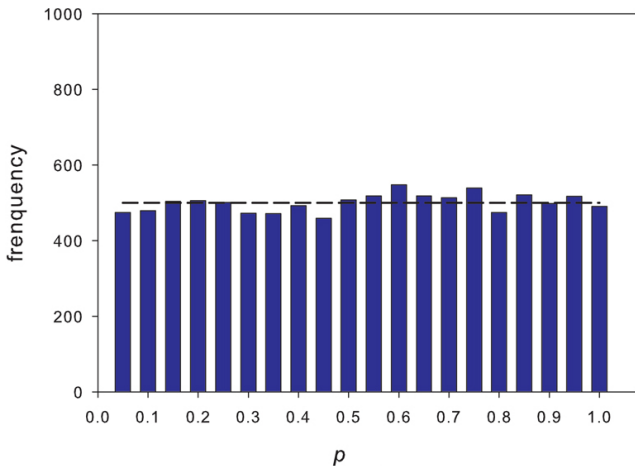
Pravděpodobnostní rozložení

- Funkce, která každé hodnotě přiřadí pravděpodobnost jejího výskytu
 - vyjadřujeme obvykle grafem
 - může být odvozena ze statistického souboru
- Často rozložení aproximujeme "ideální" funkcí
 - vyjádřitelnou vzorcem
 - určíme **typ** pravděpodobnostního rozložení
- Nejčastější typy rozložení
 - využití pro velkou škálu jevů
 - uniformní rozložení, normální rozložení, Zipfovo rozložení

Uniformní rozložení

- Všechny možnosti mají stejnou pravděpodobnost
 - např. házení (vyváženou) kostkou
 - možnosti 1, 2, 3, 4, 5, 6 mají pravděpodobnost $1/6$
 - ostatní mají 0
 - grafem jsou body tvořící úsečku

Uniformní rozložení: příklad



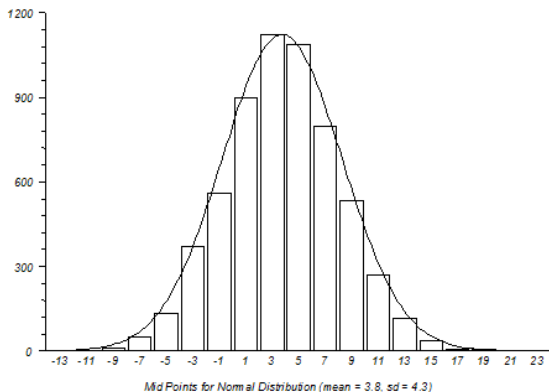
Normální rozložení

■ Normální rozložení

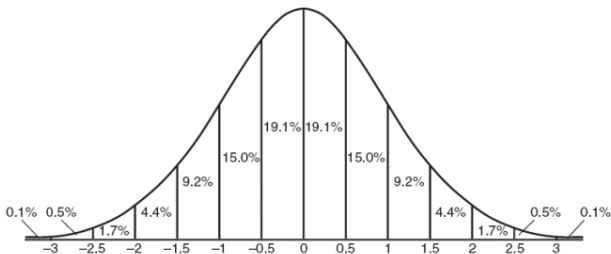
- různé vlastnosti populací
- např. výška, váha (slonů, lidí)
- nejpravděpodobnější hodnoty jsou ty, které jsou blízké průměru
- hodnoty vzdálenější od průměru jsou málo pravděpodobné
- grafem jsou body tvořící „zvon“ s osou v průměrné hodnotě

Normální rozložení: příklad

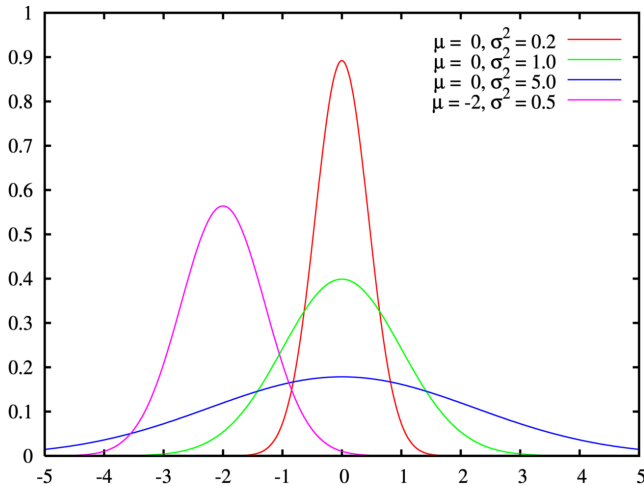
Histogram for Normal Distribution (mean = 3.8, sd = 4.3)



Normální rozložení: příklad



Normální rozložení: příklad

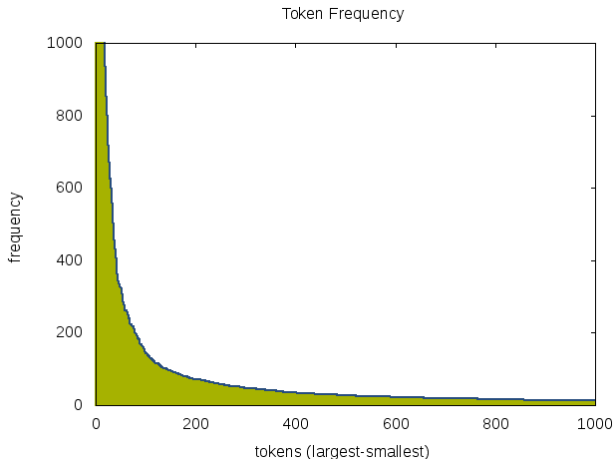


Zipfovo rozložení

■ Zipfovo rozložení

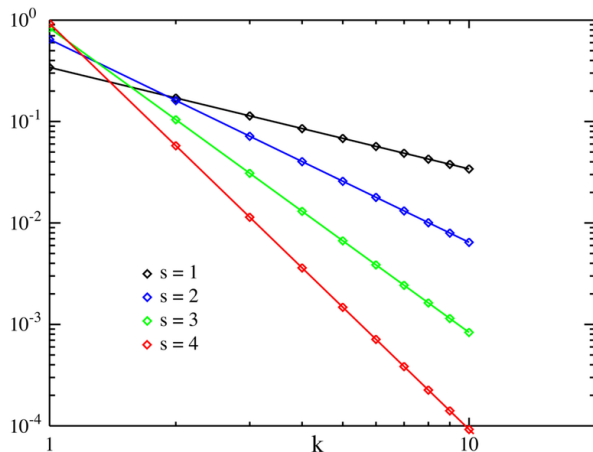
- několik málo hodnot má velkou pravděpodobnost
- pravděpodobnost dalších v pořadí prudce klesá
- např. první hodnota má pravděpodobnost n , druhá $n/2$, třetí $n/3$ atd.
- velmi často výstižně popisuje frekvenční distribuce (měst podle velikosti, slov v přirozeném jazyce podle počtu výskytů, ...)

Zipfovo rozložení: příklad



Zipfovo rozložení: logaritmické osy

V případě logaritmických os tvoří graf Zipfova rozložení přímku



Zipfův zákon

- Zipfovo rozložení dobře popisuje rozložení jazykových jevů
 - nejfrekventovanější jevy pokrývají většinu jazyka
 - frekvence (pravděpodobnost výskytu) je nepřímo úměrná pořadí podle frekvence
- Např. výskyty slov v angličtině
 - „the” tvoří 7 % slovních výskytů
 - „of” tvoří 3,5 % slovních výskytů
 - polovinu anglického korpusu pokrývá 135 nejčastějších slov
- Zipfův zákon v přirozeném jazyce platí, kam se podíváte

Frekvence morf. značek (podst. jm. a slovesa, korp. Brown)

<u>tag</u>	<u>Freq</u>
NN	<u>161881</u>
NP	<u>62669</u>
NNS	<u>56629</u>
VVN	<u>27545</u>
VV	<u>27481</u>
VVD	<u>27391</u>
VVG	<u>16922</u>
VBD	<u>13275</u>
VBZ	<u>11321</u>
VVZ	<u>8254</u>
VVP	<u>7912</u>
VB	<u>6377</u>
VBP	<u>5211</u>
VHD	<u>5190</u>
VHZ	<u>2497</u>
VBN	<u>2470</u>
VHP	<u>2445</u>
VH	<u>1780</u>
NPS	<u>1524</u>
VBG	<u>674</u>
VHG	<u>279</u>
VHN	<u>194</u>

Zipfův zákon – zajímavosti

- Platí pro většinu přirozených jazyků
 - dokonce i pro sekvence náhodně volených znaků (vč. mezery)
 - pro širokou škálu jevů (distribuce významů slov, syntaktických vazeb, ...)
 - je třeba s ním vždy počítat
- „Ekonomické pravidlo“ 80 : 20
 - 80 % problému vyřešíme s 20 % úsilím
 - alternativní formulace Zipfova zákona
- Platí i pro mnoho „nejazykových“ jevů
 - počet obyvatel měst, platy, velikost společností, ...

Zákon velkých čísel

■ Potřeba velkých dat

- čím více pokusů provádíme, tím větší je pravděpodobnost, že průměr výsledků pokusů bude odpovídat očekávané hodnotě (vypočtenému průměru)

■ Alternativně

- čím více dat máme, tím méně náhodných odchylek budou obsahovat

■ Například

- pokud budeme sledovat jednu molekulu vody v moři, bude se její pohyb jevit náhodný
- pokud budeme sledovat významný podíl molekul, jsme schopni pozorovat vlnění, příliv a odliv, mořské proudy...

Zákon velkých čísel

- Platí ve všech oblastech statistiky
 - pokud budeme mít jazykový korpus o např. 100 slovech, jsme příliš ovlivněni náhodností výběru a statistické charakteristiky nemají smysl
 - → korpusy o velikosti miliard slov

Statistické testování hypotéz

- Cíl: statistická průkaznost
 - ověřit, zda příslušná statistická data potvrzují nějakou hypotézu
- Příklad: hádání karet
 - člověk se pokouší uhádnout barvu karty, která je mu ukázána z rubu
 - kolikrát musí uhodnout (např. z 25 pokusů), abychom mohli říct, že „je jasnovidec“?
 - uhádne 5x – nejspíš náhoda
 - uhádne 24x – nejspíš je jasnovidec
 - uhádne 11x – ?
 - jak určit hranici?

Statistické testování hypotéz

- I 25 úspěšných pokusů může být náhoda
 - vyloučit to neumíme
 - umíme ale vyjádřit pravděpodobnost takové události

Statistické testování hypotéz: pojmy

■ Nulová hypotéza H_0

- výchozí názor, který chceme vyvrátit
- „dotyčný není jasnovidec“

■ Alternativní hypotéza H_1

- ta, pro kterou hledáme oporu v datech
- „dotyčný je jasnovidec“

■ Chyba typu I

- potvrdíme alternativní hypotézu, ta přitom neplatí
- prohlásíme dotyčného za jasnovidce, ten přitom jen tipoval
- **chceme minimalizovat pravděpodobnost této chyby**

Statistické testování hypotéz: pravděpodobnost chyby

■ Pravděpodobnost chyby

- 1 tip má pravděpodobnost úspěchu $1/4$
- 25/25 úspěšných pokusů: $(1/4)^{25}$, tj. cca 10^{-15}
- (předpokládáme, že pokusy jsou nezávislé)
- → pokud v případě 25 úspěšných pokusů prohlásíme dotyčného za jasnovidce, spleteme se s pravděpodobností cca 10^{-15}

■ Statistická průkaznost

- pokud pravděpodobnost chyby je menší než 1-5 %
- došlo k **vyvrácení nulové hypotézy**
- = alternativní hypotéza byla statisticky prokázána
- 1 % odpovídá 12/25 úspěšným pokusům v hádání karet

Statistické testování hypotéz: nástrahy

■ Opakování pokusů

- pokud zopakujeme pokus vícekrát, pravděpodobnost chyby se zvětšuje
- musíme uvažovat všechny provedené pokusy

■ Pokud se nepodaří vyvrátit nulovou hypotézu

- neznamená to, že nulová hypotéza platí
- „nepodařilo se prokázat souvislost“ \neq „podařilo se prokázat, že souvislost neexistuje“
- podobně neprokázání viny u soudu neprokazuje nevinu obžalovaného