

Pravděpodobnost, náhoda, kostky

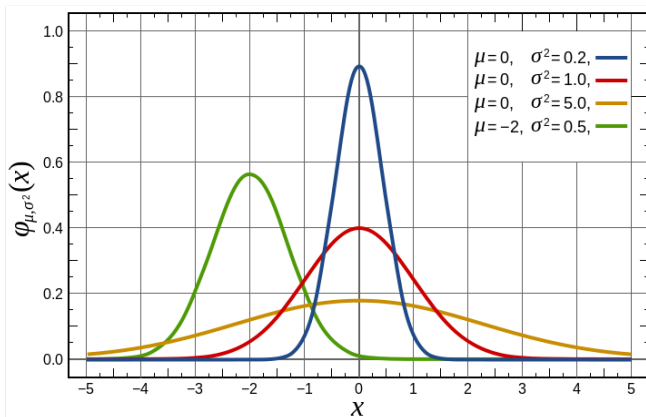
Radek Pelánek

IV122, jaro 2013

- pravděpodobnost
- náhodná čísla
- lineární regrese
- detekce shluků

- pravděpodobnost, podmíněná pravděpodobnost, nezávislost
- střední hodnota, rozptyl, směrodatná odchylka
- distribuční funkce
- normální distribuce

Normální distribuce



Wikipedia

Normální distribuce

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ – průměr
- σ – standardní odchylka

Monty Hall Problem

- troje dveře, za jedněmi z nich je poklad, cílem je najít poklad
- vyberete jedny dveře
- já otevřu jedny z nevybraných dveří, za kterými není poklad
- vy nyní můžete zůstat u své volby nebo změnit své rozhodnutí
- co je rozumné udělat?
 - zůstat u své volby
 - změnit rozhodnutí
 - je to úplně jedno (můžeme se rozhodnout náhodně)

Monty Hall Problem: řešení

- je výhodnější změnit rozhodnutí:
 - zůstat u své volby: 33 %
 - změnit rozhodnutí: 66 %
 - rozhodnout se náhodně: 50 %
- problém známý tím, že i mnoho matematiků se v něm snadno splete
- pro vybudování intuice (přesvědčení skeptiků) se hodí simulace...

Monty Hall: experimentálně

- implementujte simulátor hry
- vyzkoušejte strategie „zůstat při původním rozhodnutí“, „změnit rozhodnutí“, „náhodně měnit rozhodnutí“
- experimentálně vyhodnoťte úspěšnost strategií v dlouhém běhu

Náhodná čísla

- aplikace:
 - počítačové hry, loterie
 - kryptografie
 - vědecké výpočty, simulace
- zdroje:
 - „pseudonáhodná čísla“ – běžné `random()`, „deterministické s chaotickým chováním“
 - „opravdová náhoda“ – např. atmosférický tlak, www.random.org

Co to jsou náhodná čísla?

„Házení kostkou“ – čísla 1-6

Která z následující posloupností je více pravděpodobná?

- 1 1 2 2 3 3 4 4 5 5 6 6
- 1 5 2 3 4 6 2 3 3 1 2 4

Co to jsou náhodná čísla?

„Házení kostkou“ – čísla 1-6

Která z následující posloupností je více pravděpodobná?

- 1 1 2 2 3 3 4 4 5 5 6 6
- 1 5 2 3 4 6 2 3 3 1 2 4

Obě mají stejnou pravděpodobnost $(\frac{1}{6})^{12}$

Úkol: (ne)náhodné posloupnosti

- máte k dispozici několik posloupností čísel
„hody kostkou“ \sim celá čísla 1 až 6
- určete, které z nich jsou „nenáhodné“ a proč
- co to znamená, že posloupnost je „náhodná“?

Testování náhodnosti

DILBERT By SCOTT ADAMS



Testování náhodnosti

- nenáhodná posloupnost:
 - predikovatelná – dokážete předpovědět další číslo (lépe než náhodným tipem)?
 - zdroje nenáhodnosti např. zkreslení, korelace, vzory, periodičita
- existují rozsáhlé sady testů náhodnosti
- vztah statistické testy

Testování náhodnosti: frekvence

Frekvence čísel ve 300 hodech

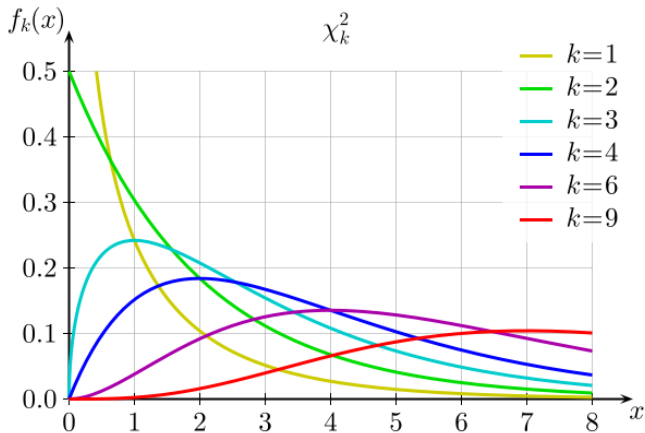
	1	2	3	4	5	6
očekávané	50	50	50	50	50	50
série 1	42	57	61	44	41	55
série 2	58	47	51	46	62	36
série 3	52	65	66	34	36	48

Odpovídá to náhodnému generování?

Testování náhodnosti: Chí kvadrát test

- O_i – očekávaný počet
- P_i – pozorovaný počet
- $S = \sum_{i=1}^6 \frac{(P_i - O_i)^2}{P_i}$
- S – pro velké n má přibližně χ^2 -rozložení o 5 stupních volnosti
- $\chi^2(k) = \sum_{i=1}^k Z_i^2$
kde Z_i má standardní normální rozdělení
- test: určíme p-hodnotu $\chi^2(5)$ pro S , pokud příliš malá – zamítnout

Chí kvadrát



Wikipedia

Generování náhodných čísel

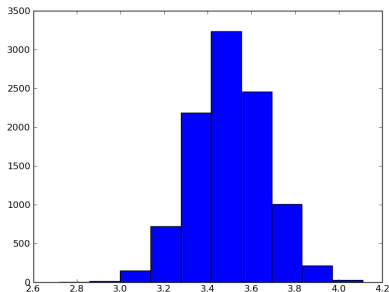
- uniformní distribuce: základ – rekurentní vztahy, např.
$$x_n = (ax_{n-1} + c) \bmod m$$
- neuniformní distribuce:
 - např. normálně rozložená čísla, náhodné body v kruhu, podle zadané distribuce, ...
 - chytré transformace, rejection sampling, Monte Carlo Markov Chain, ...

Centrální limitní věta

- nezávislé a identicky rozložené proměnné
- vzorky velikosti n
- pro velké n je průměr vzorku přibližně normálně rozložen

Centrální limitní věta: příklad

hody (férovou) kostkou
vzorky velikosti 100
počet vzorků 10000



Centrální limitní věta: poznámky

- umožňuje modelovat mnoho „neznámých vlivů“ pomocí normální distribuce
- typický příklad – šum v datech (chyba měření):
 - předpokládáme, že šum je výsledkem mnoha dílčích vlivů
 - modelujeme pomocí normální distribuce
- pozor na:
 - předpoklad „nezávislé a identicky rozložené“
 - platí pro aritmetický průměr („aditivní“ veličiny)
 - rychlost konvergence závisí na výchozí distribuci

Centrální limitní věta: příklady kostky

K_a = zatížená kostka, která preferuje vyšší čísla
(pravděpodobnost úměrná počtu teček)

K_b = inverzně zatížená kostka

Jak to dopadne (rozmyslete „teoreticky“, vyzkoušejte prakticky):

- hody kostkou K_a
- pro každý hod náhodně vybereme jednu z kostek K_a , K_b
- náhodně vybereme jednu z kostek K_a , K_b a tou házíme všechna čísla ve vzorku

Věnujte pozornost tvaru výsledné distribuce, průměru i směrodatné odchylce.

Bayesova věta

- $P(A|B)$ – podmíněná pravděpodobnost
- Bayesova věta

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesova věta

- D – pozorovaná data
- H_i – hypotézy o vzniku dat
- $P(H_i)$ – „prior“, odhad pravděpodobnosti H_i předtím, než jsme viděli data
- $P(D|H_i)$ – pravděpodobnost dat při dané hypotéze
- $P(H_i|D)$ – „posterior“, odhad pravděpodobnosti H_i korigovaný daty
- Bayesova věta

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{P(D)}$$

- $P(D) = \sum_i P(D|H_i)P(H_i)$ – pravděpodobnost dat

Bayesova věta – klasický příklad

- předpokládejme
 - výskyt AIDS: 6 z 1000
 - spolehlivý test na AIDS:
 - správný výsledek 99.9 % pro ty, co mají AIDS
 - 99 % pro ty, co nemají AIDS
- výsledek testu osoby X je pozitivní
- jaká je pravděpodobnost, že X má AIDS?

Bayesova věta – klasický příklad

hypotézy: A = AIDS, N = nemá AIDS

data: V = pozitivní výsledek

$$\begin{aligned}P(A|V) &= \frac{P(V|A)P(A)}{P(V|A)P(A)+P(V|N)P(N)} \\&= \frac{0.006 \cdot 0.999}{0.006 \cdot 0.999 + 0.994 \cdot 0.01} \sim 0.38\end{aligned}$$

Bayesova věta – příklad kostky

- 1, 3, 4, 5, 1, 4, 6, 5, 1, 5, 4, 5
- posloupnost byla vygenerována jednou z následujících kostek:
 - ① normální kostka
 - ② normální kostka, na které je dvojka přepsaná na pětku
 - ③ zatížená kostka, na které padá 6 s dvakrát větší pravděpodobností než ostatní čísla
- jaký je posterior (jak bychom měli věřit pravděpodobnosti jednotlivých kostek) pro:
 - uniformní prior (dopředu jsme považovali všechny možnosti za stejně pravděpodobné)?
 - prior preferující 3. kostku (věříme: 5 % normální, 5 % přepsaná, 90 % zatížená)?
- napište program:
vstup: prior, posloupnost, výstup: posterior