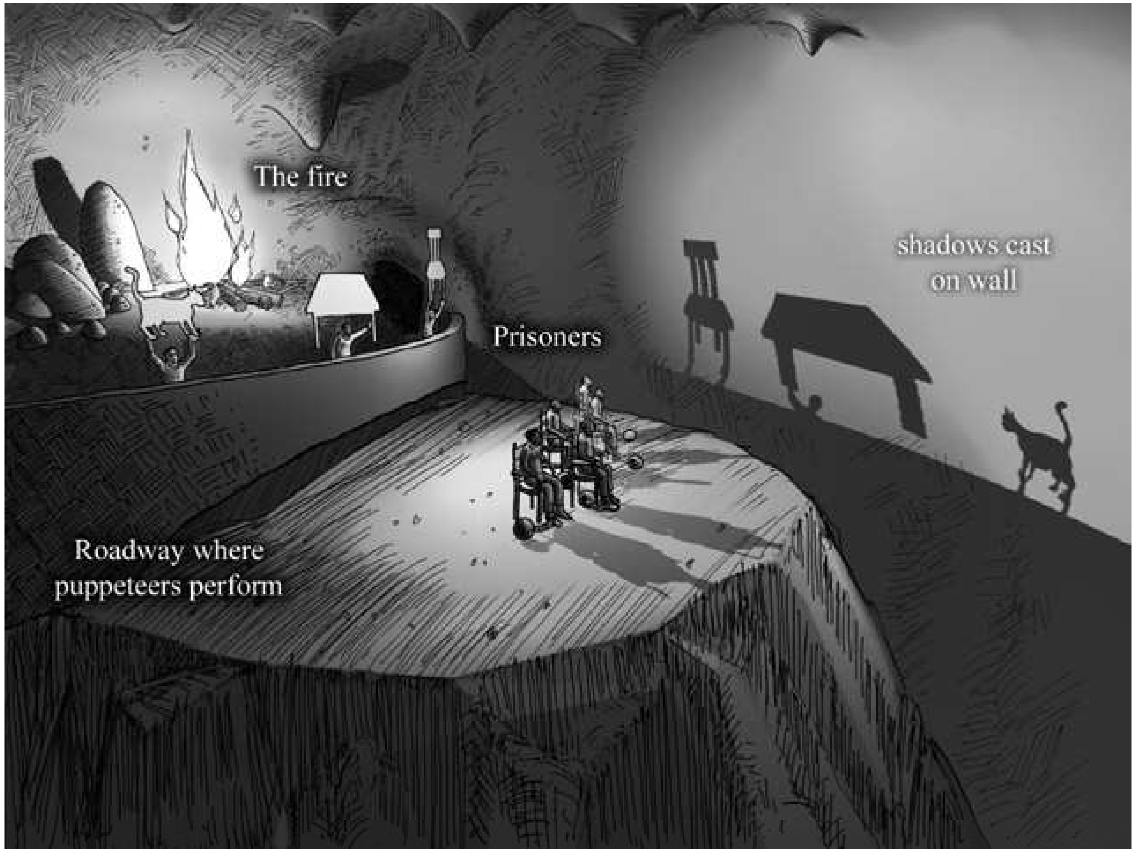
1. Statistická analýza dat

Jak vznikají informace
Rozložení dat

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, S. Littnerová, J. Kalina

Význam statistické analýzy dat

- Sběr a vyhodnocování dat je způsobem k uchopení a pochopení reality.
- Chápání reality je vždy nedokonalé a nepřesné.
- Statistika umožňuje vnést do pochopení reality určitou spolehlivost a ukázat, jak je velká.

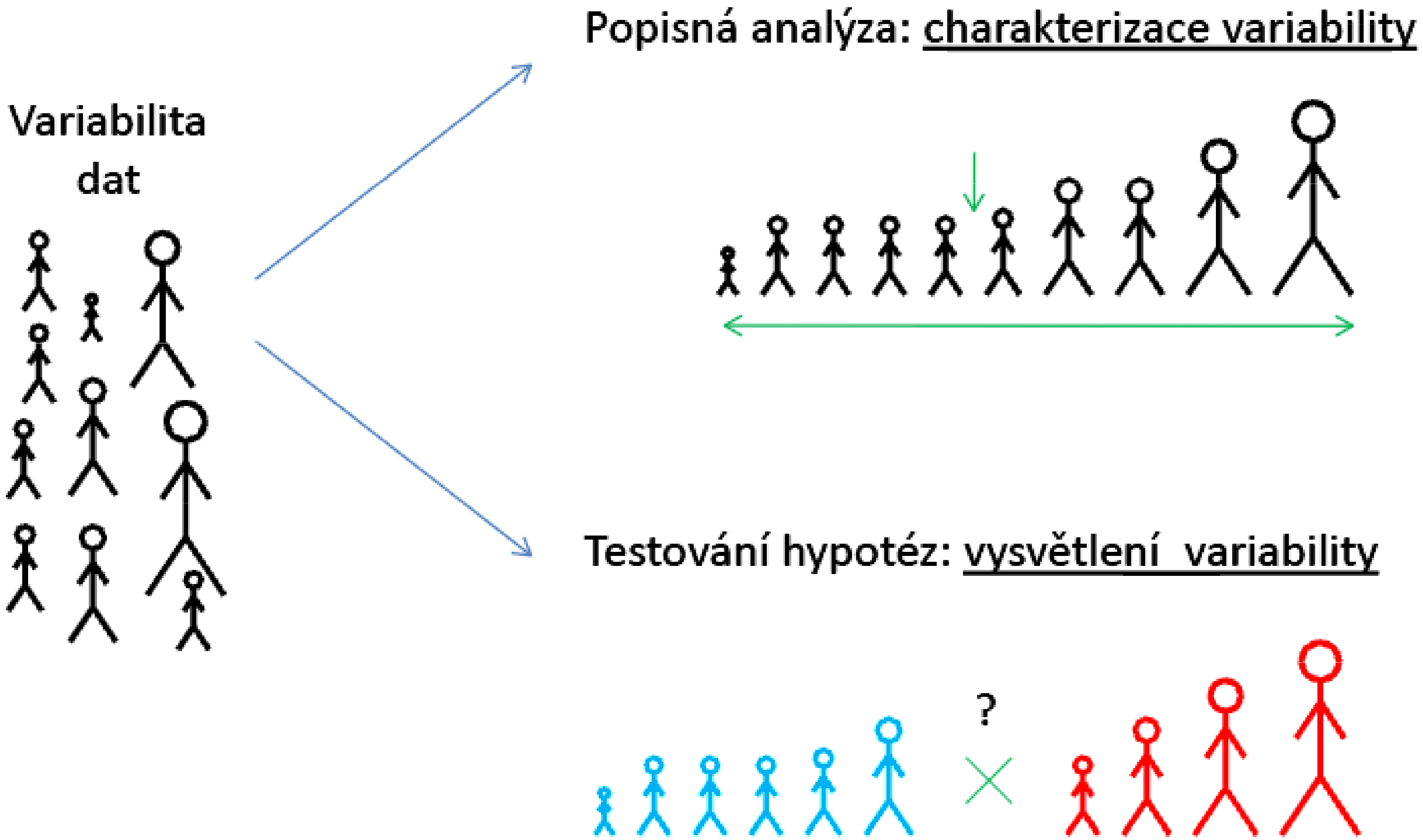


Význam statistické analýzy dat

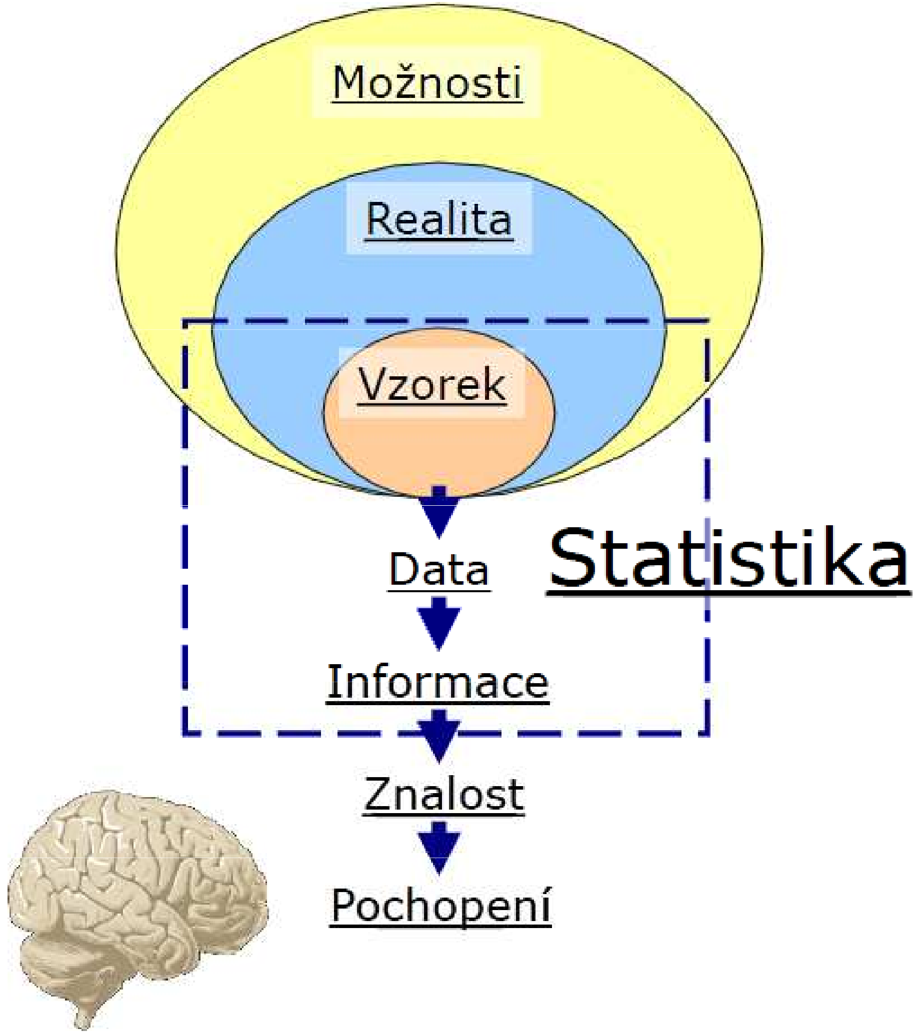
- Realita je variabilní a statistika je věda zabývající se variabilitou.
- Korektní analýza variability a její pochopení přináší užitečné informace o realitě.
- V případě deterministického světa by statistická analýza nebyla potřebná.
- V případě zcela chaotického světa by nebyla možná.

Práce s variabilitou v analýze dat

- Dva hlavní přístupy k variabilitě:

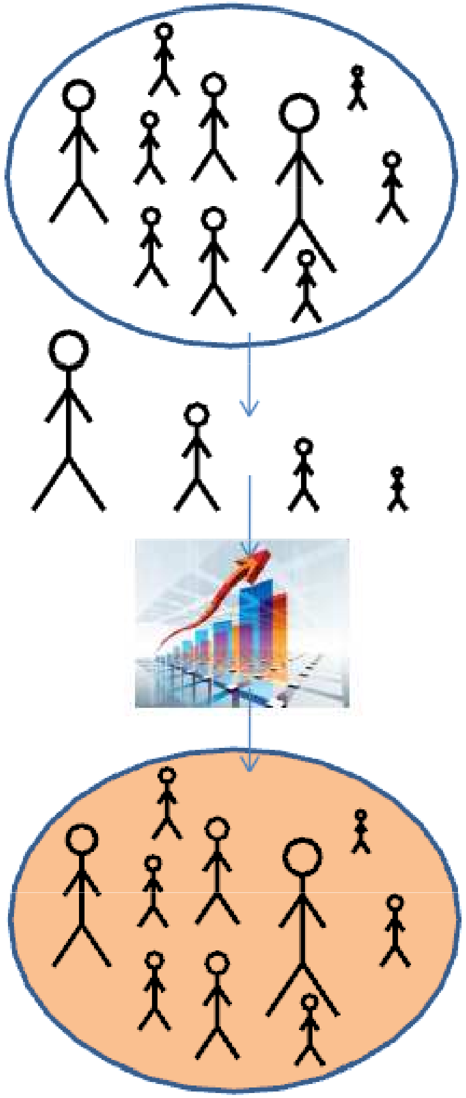


Práce s variabilitou v analýze dat



- Statistika není schopna činit závěry o jevech neobsažených ve zkoumaném vzorku.
- Statistika je nasazena v procesu získání informací ze vzorkovaných dat a je podporou v získání znalosti a pochopení problému.
- Statistika není náhradou naší inteligence!

Práce s variabilitou v analýze dat



Neznámá cílová populace

Vzorek

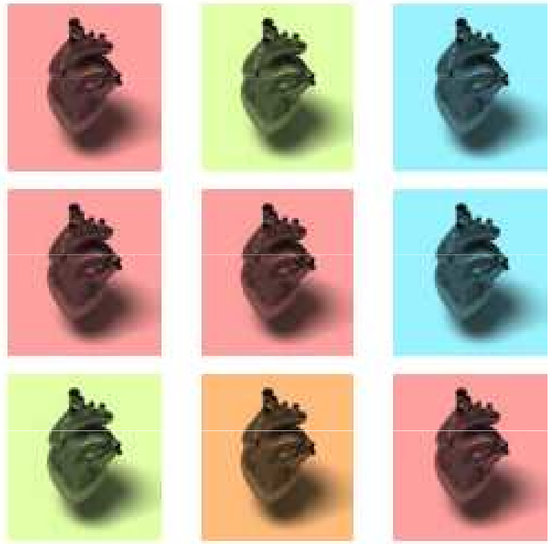
Analýza

Díky zobecnění výsledků známe vlastnosti cílové populace *(s určitou pravděpodobností chyby)*

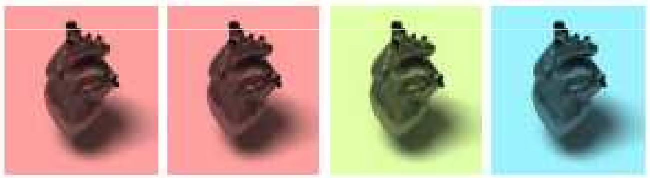
- Cílem analýzy není pouhý popis a analýza vzorku, ale zobecnění výsledků ze vzorku na jeho cílovou populaci.
- Pokud vzorek nereprezentuje cílovou populaci, vede zobecnění k chybným závěrům.

Význam vzorkování ve statistice

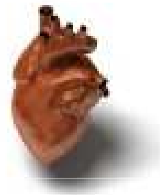
- Statistika hovoří o realitě prostřednictvím vzorku!!!
- Statistické předpoklady korektního vzorkování je nutné dodržet

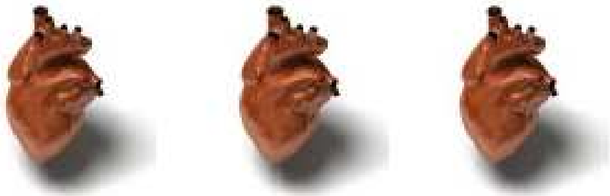


- Náhodný výběr z cílové populace
- Representativnost: struktura vzorku musí maximálně reflektovat realitu



- Nezávislost: několikanásobné vzorkování téhož objektu nepřináší ze statistického hlediska žádnou novou informaci

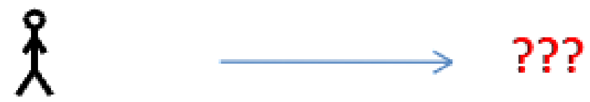




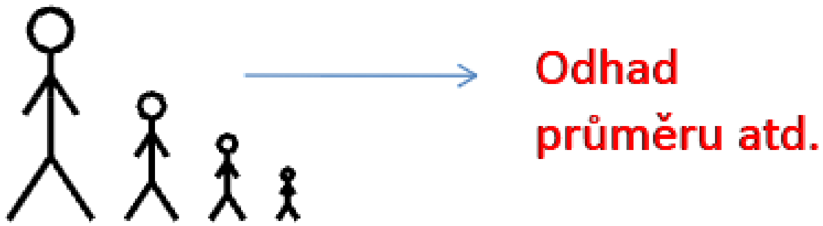
Velikost vzorku a přesnost statistických výstupů

- Existuje skutečné rozložení a skutečný průměr měřené proměnné

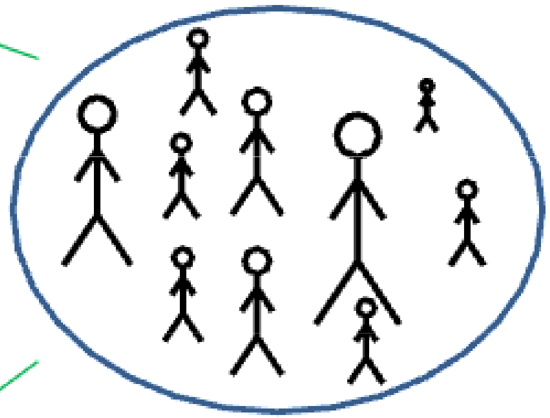
- Z jednoho měření nezjistíme nic



- Vzorek určité velikosti poskytuje odhad reálné hodnoty s definovanou spolehlivostí



- Vzorkování všech existujících objektů poskytne skutečnou hodnotu dané popisné statistiky, nicméně tento přístup je ve většině případů nereálný.



Obecné schéma aplikace statistické analýzy

Experimentální design	Jak velký vzorek je nezbytný pro statisticky relevantní výsledky? Klíčová stratifikační kritéria cílové populace.
Vzorkování	Vzorkovací plán zabezpečující náhodnost a reprezentativnost vzorku.
Uložení a management dat	Uložení dat ve vhodné formě a jejich vyčištění předcházející vlastní analýze je klíčovým krokem statistické analýzy.
Vizualizace dat	Grafická inspekce dat je nezbytným krokem analýzy vzhledem ke schopnosti lidského mozku primárně akceptovat obrazová data. Poskytne vhled do dat, představu o jejich rozložení, vazbách proměnných apod.
Popisná analýza	Popisná analýza umožňuje vyhodnotit srovnáním s existující literaturou realističnost naměřených rozsahů dat.
Testování hypotéz	Testování vazeb mezi různými proměnnými s cílem navzájem vysvětlit jejich variabilitu a tím přispět k pochopení řešeného problému.
Modelování	Možným vyvrcholením analýzy je využití získaných znalostí a pochopení problému k vytvoření prediktivních modelů.

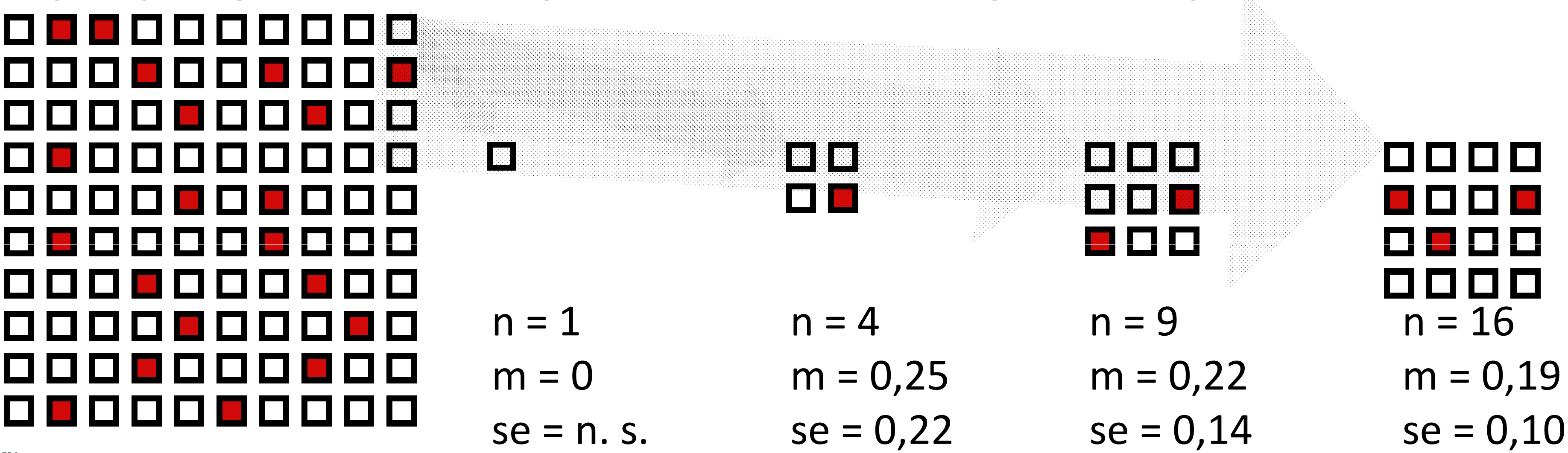
1a. Teoretické pozadí statistické analýzy

Jak vznikají informace

Rozložení dat

Anotace

- Základním principem statistiky je pravděpodobnost výskytu nějaké události. Prostřednictvím vzorkování se snažíme odhadnout skutečnou pravděpodobnost událostí.
- Klíčovou otázkou je velikost vzorku, čím větší vzorek, tím větší šance na projevení se skutečné pravděpodobnosti výskytu jevu (a tím je také nákladnější analýza).



Definice

Náhodný jev: značíme velkým latinským písmenem, např. A. Jde o jev, pro který požadujeme tzv. statistickou stabilitu, tj. aby při n opakováních pokusu platilo pro relativní četnost výsledku:

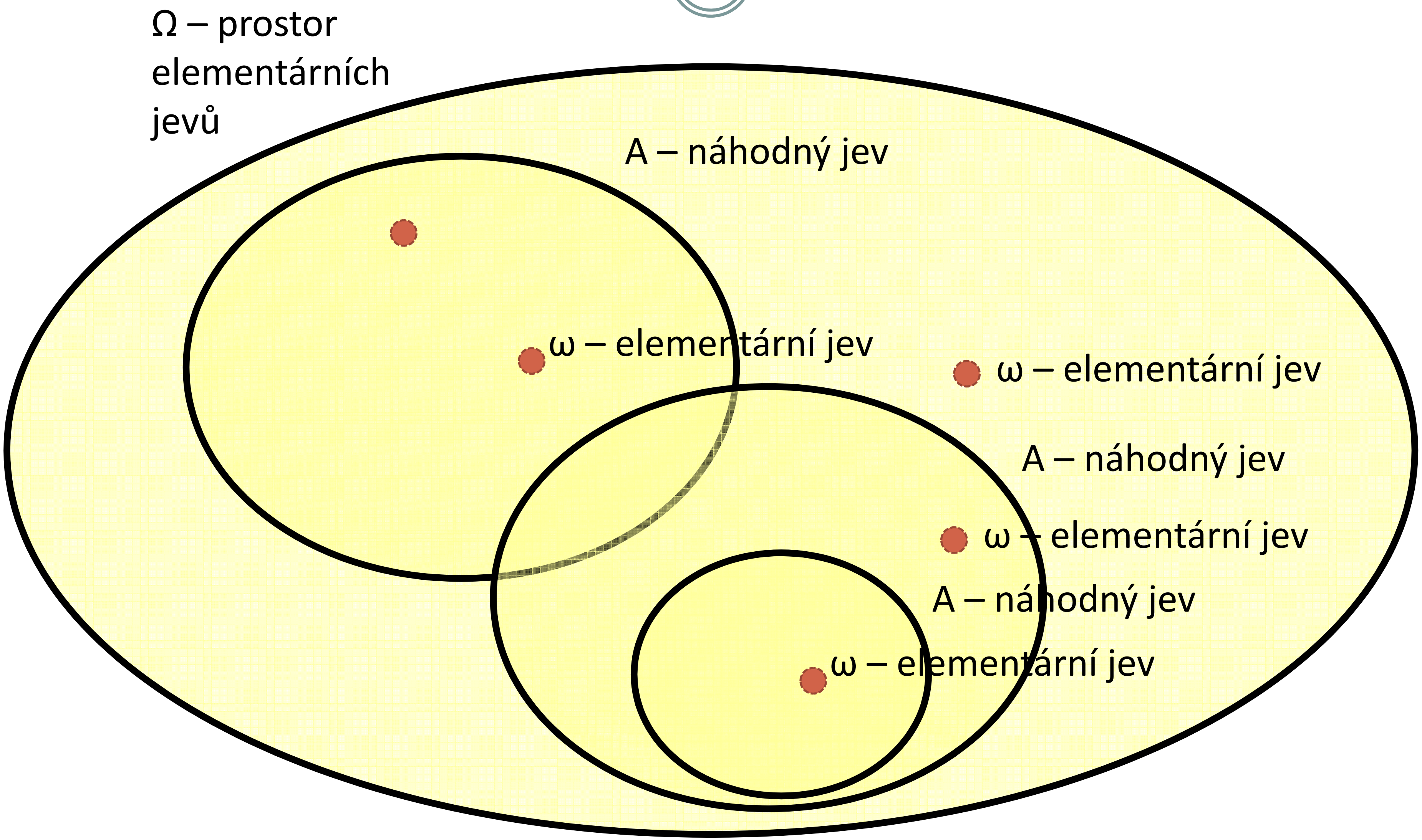
$$\lim_{n \to \infty} f(A) = \lim_{n \to \infty} \frac{n_A}{n} = p(A)$$

Elementární jev: nejjemnější možný náhodný jev, tj. náhodný jev, který nelze vyjádřit jako sjednocení dvou jiných neprázdných náhodných jevů. Značí se obvykle ω.

Prostor elementárních jevů: značíme obvykle Ω, jde o libovolnou neprázdnou množinu (její prvky nazýváme elementárními jevy).

Platí tedy, že elementární jevy jsou prvky prostoru elementárních jevů, rovněž jsou prvky náhodných jevů a náhodné jevy jsou podmnožiny prostoru elementárních jevů.

Definice


Ω – prostor elementárních jevů
A – náhodný jev
ω – elementární jev
ω – elementární jev
A – náhodný jev
ω – elementární jev
A – náhodný jev
ω – elementární jev

Definice

σ-algebra systém (množina) podmnožin prostoru elementárních jevů A (označujeme $\mathfrak{A}$) splňující následující podmínky:

1. $\mathfrak{A}$ je neprázdná množina,
2. $A \in \mathfrak{A} \;\Rightarrow\; \mathfrak{A} \setminus A \in \mathfrak{A}$
3. sjednocení libovolného počtu $A_i \in \mathfrak{A}$.

Jevové pole uspořádaná dvojice prostoru elementárních jevů a na něm definované σ-algebry $(\Omega, \mathfrak{A})$. Jevové pole se také někdy nazývá měřitelný prostor.

Pravděpodobnost reálná množinová funkce P definovaná na množině A σ-algebry $(\Omega, \mathfrak{A})$ tak, že jsou dodrženy následující podmínky:

(podle Kolmogorova)

1. $P(\Omega) = 1$
2. $\forall A \in \mathfrak{A}: \; P(A) \geq 0$
3. pravděpodobnost součtu neslučitelných jevů je rovna součtu pravděpodobnosti těchto neslučitelných jevů.

Definice

Pravděpodobnostní prostor uspořádaná trojice prostoru elementárních jevů, na něm definované σ-algebry a jim příslušné pravděpodobnostní funkce $(\Omega, \mathfrak{A}, P)$.

Borelovská σ-algebra je σ-algebra $\mathfrak{B}$ generovaná systémem borelovských množin S, tj. množin splňujících podmínku:

1. $S = (-\infty, x\rangle$, kde $x \in \mathbb{R}$.

Náhodná veličina reálná množinová funkce X definovaná na prostoru elementárních jevů Ω nějakého pravděpodobnostního prostoru $(\Omega, \mathfrak{A}, P)$, splňující pro nějakou borelovskou σ-algebru $\mathfrak{B}$ předpoklad:

1. $B \in \mathfrak{B} \;\Rightarrow\; \{\omega \in \Omega:\; X(\omega) \in B\} \in \mathfrak{A}$.

Pravděpodobnostní prostor je měřitelný prostor s přidanou funkcí pravděpodobnosti.

Definice

Náhodná veličina se někdy také nazývá náhodná proměnná nebo měřitelná funkce, borelovské množiny se někdy též nazývají měřitelné množiny.

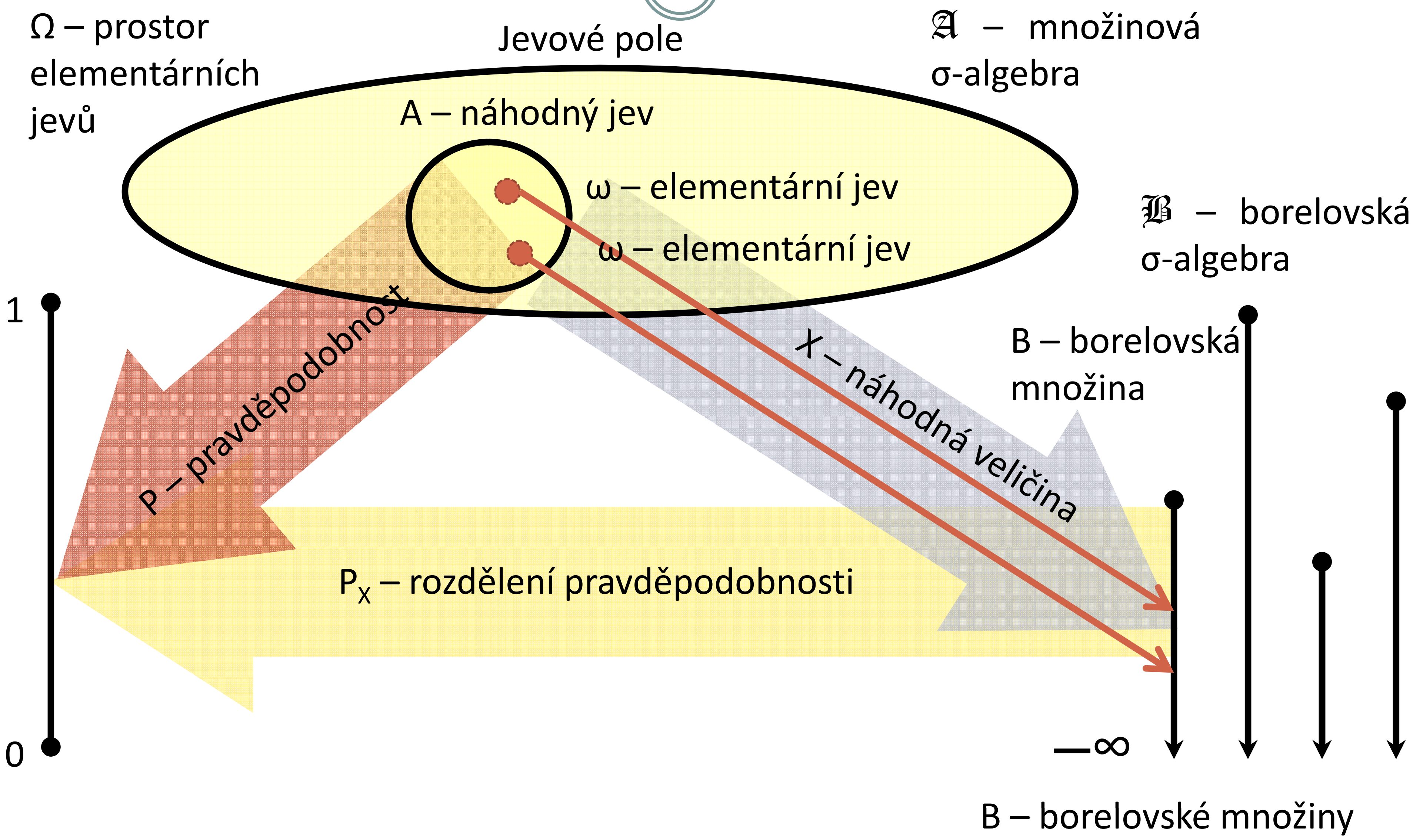
Lze ukázat, že dostatečnou podmínkou pro to, aby X byla náhodná veličina je vztah $\forall x \in \mathbb{R}$: $\{X < x\} \in \mathfrak{A}$.

Rozdělení pravděpodobnosti — množinová funkce P_X, která každé borelovské množině B přiřadí pravděpodobnost tak, že je dodržena následující podmínka:

1. $P_X(B) = P(\{\omega \in \Omega: X(\omega) \in B\})$ pro $B \in \mathfrak{B}$.

Náhodná veličina přiřazuje náhodným jevům měřitelné hodnoty (reálná čísla), rozdělení pravděpodobnosti pak každé takové hodnotě (reprezentované nějakou borelovskou množinou B) přiřazuje pravděpodobnost, tj. hodnotu mezi 0 a 1 takovou, že jsou dodrženy předpoklady po definici pravděpodobnosti uvedené dříve.

Definice

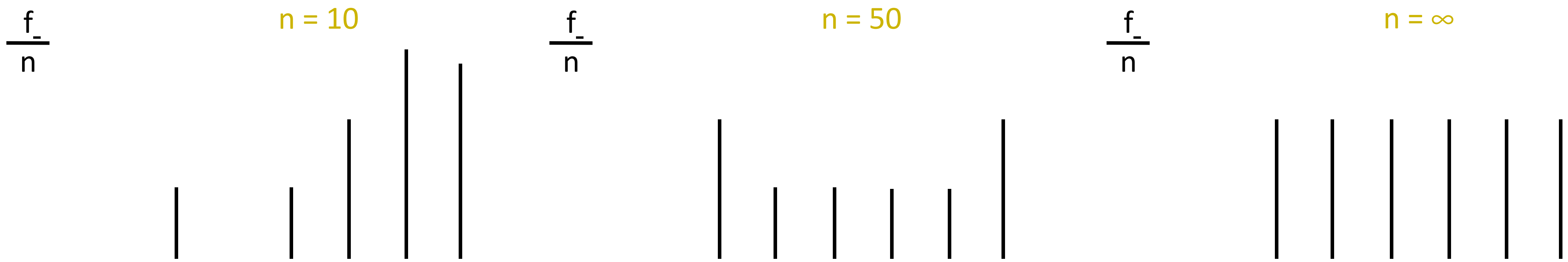


Ω – prostor elementárních jevů
Jevové pole
𝔄 – množinová σ-algebra
A – náhodný jev
ω – elementární jev
ω – elementární jev
𝔅 – borelovská σ-algebra
1
P – pravděpodobnost
X – náhodná veličina
B – borelovská množina
P_X – rozdělení pravděpodobnosti
0
−∞
B – borelovské množiny

JAK vznikají informace ?

„Empirical approach“

„Classical approach“

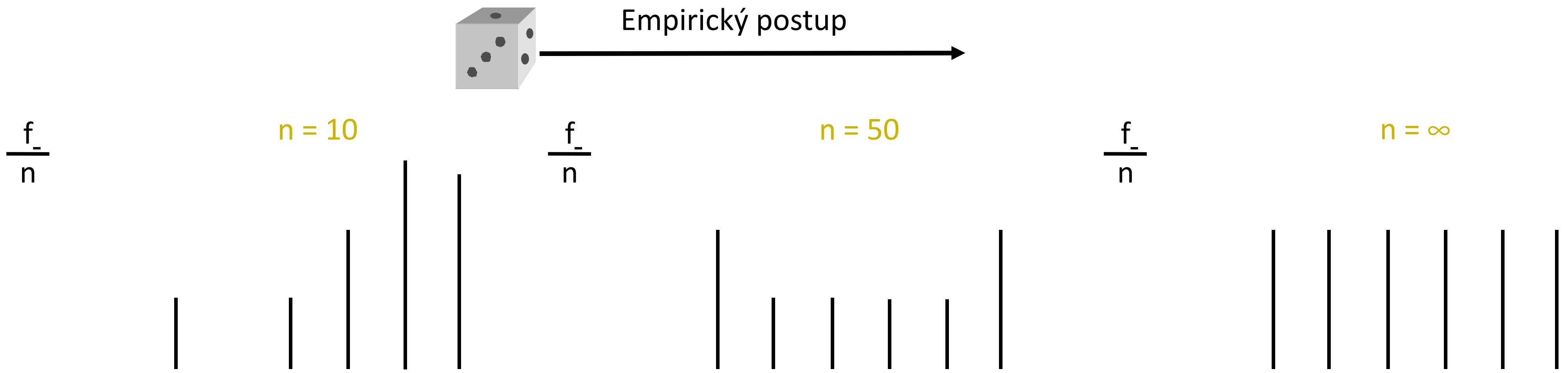


možné jevy: čísla 1 – 6

n – počet hodů (opakování)

U složitých stochastických systémů se pravdě blížíme až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit

JAK vznikají informace ?



možné jevy: čísla 1 – 6

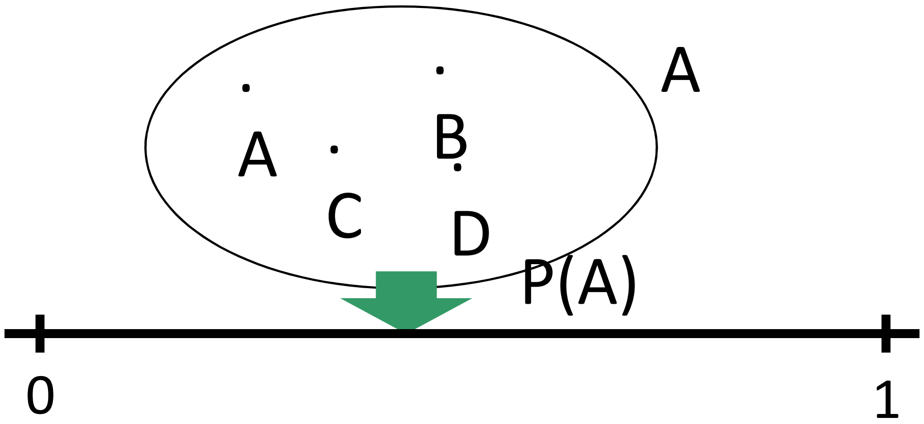
n – počet hodů (opakování)

Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější) …diskutabilní je ale ovšem míra zobecnění konkrétního experimentu

Empirický zákon velkých čísel

Při opětovné nezávislé realizaci téhož náhodného experimentu se podíl výskytů sledovaného jevu mezi všemi dosud provedenými realizacemi zpravidla ustaluje kolem konstanty.

Pravděpodobnost je libovolná reálná funkce definovaná na jevovém poli A, která každému jevu A přiřadí nezáporné reálné číslo P(A) z intervalu 0 - 1.



Z praktického hlediska je pravděpodobnost idealizovaná relativní četnost

P (A) = 1 .... jev jistý

P (A) = 0 .... jev nemožný

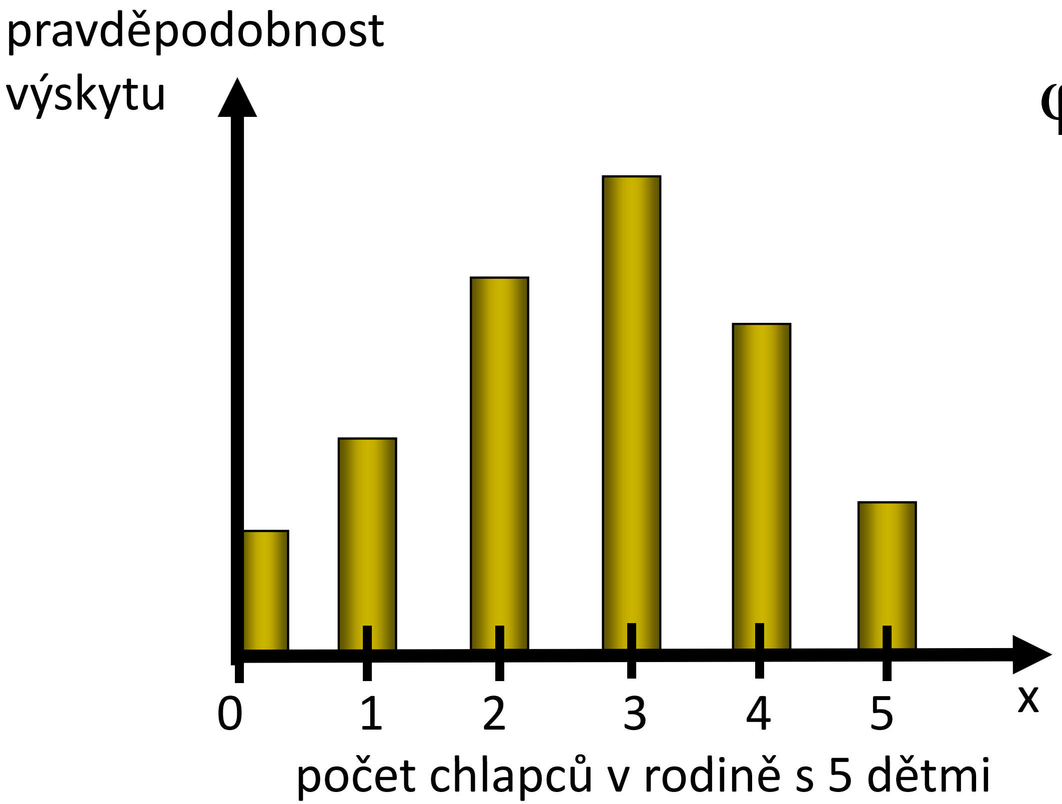
$P(A \cap B) = P(A) \cdot P(B)$............ nezávislé jevy

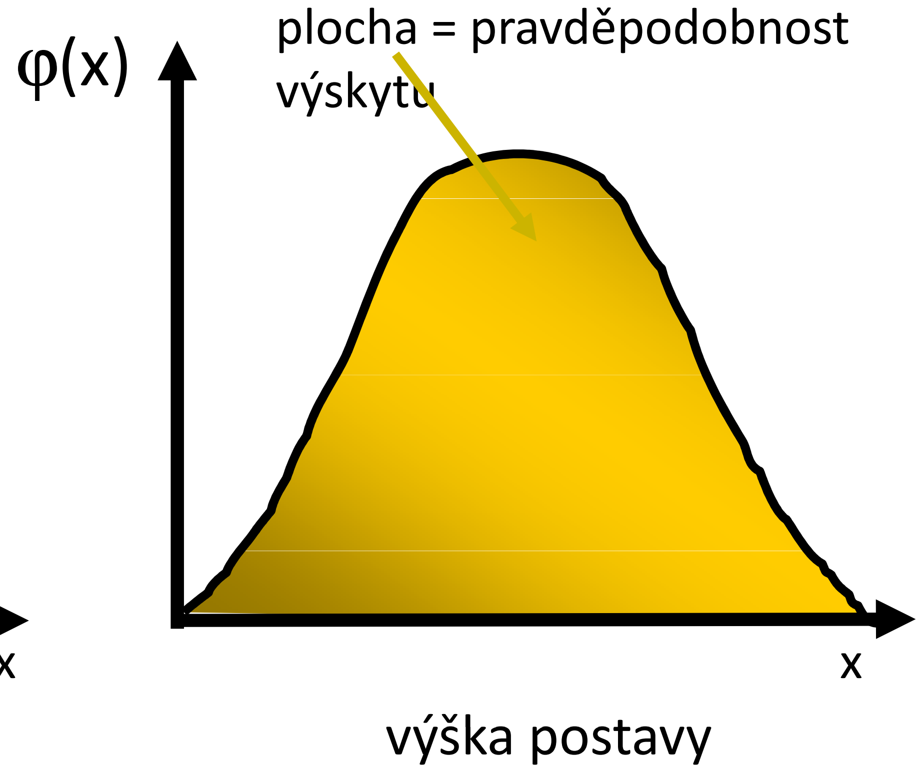
$P(A \cap B) = P(A) \cdot P(B/A)$ ....závislé jevy

$P(A / B) = P(A \cap B) / P(B)$ ....podmíněná pravděpodobnost

Pravděpodobnost výskytu jevu – rozložení dat

- existuje pravděpodobnost výskytu jevů (nedeterministické závěry)
- „vše je možné“: pouze jev s pravděpodobností 0 nikdy nenastane
- pravděpodobnost lze zkoumat retrospektivně i prospektivně





Centrální limitní věta

Pokud lze náhodnou veličinu X vyjádřit jako součet náhodných veličin X_1, X_2,…, X_n, které mají shodné rozdělení, konečnou střední hodnotu a konečný rozptyl, platí, že rozdělení veličiny X se vzrůstajícím n konverguje (poměrně rychle) k normálnímu rozdělení.