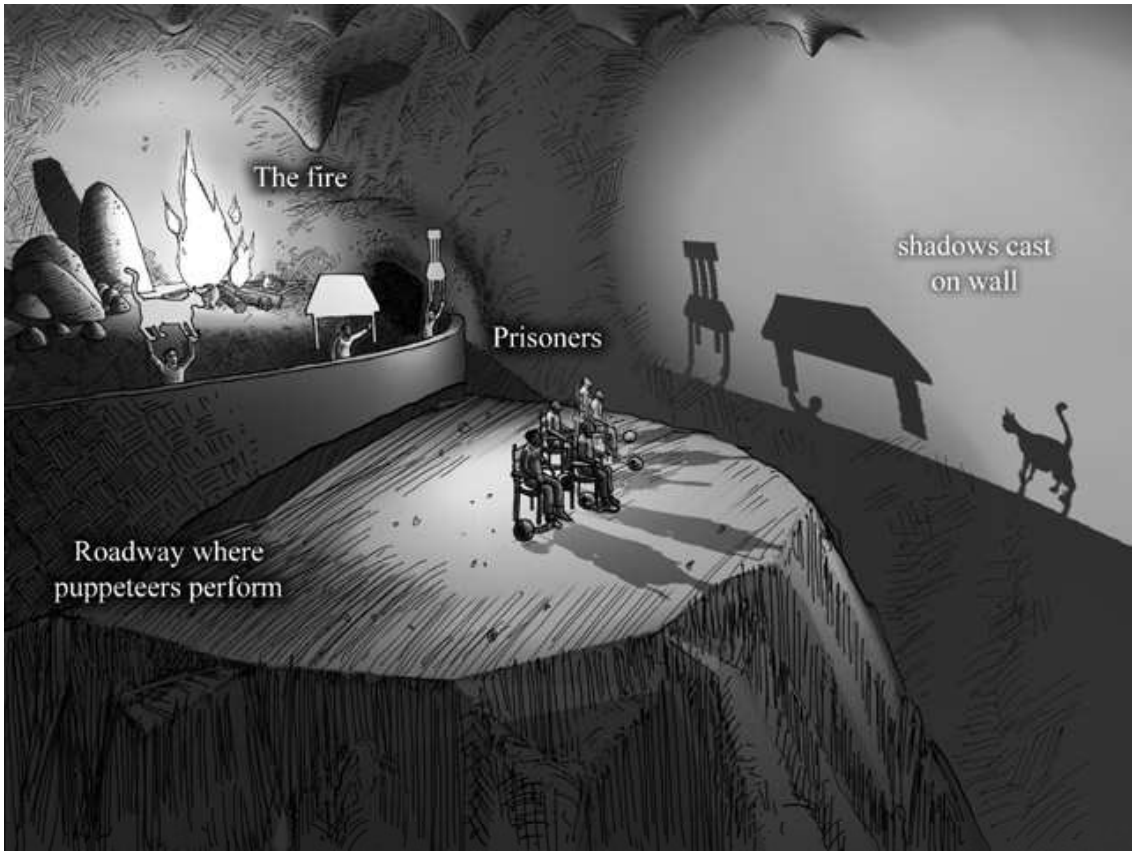
1. Statistická analýza dat

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.

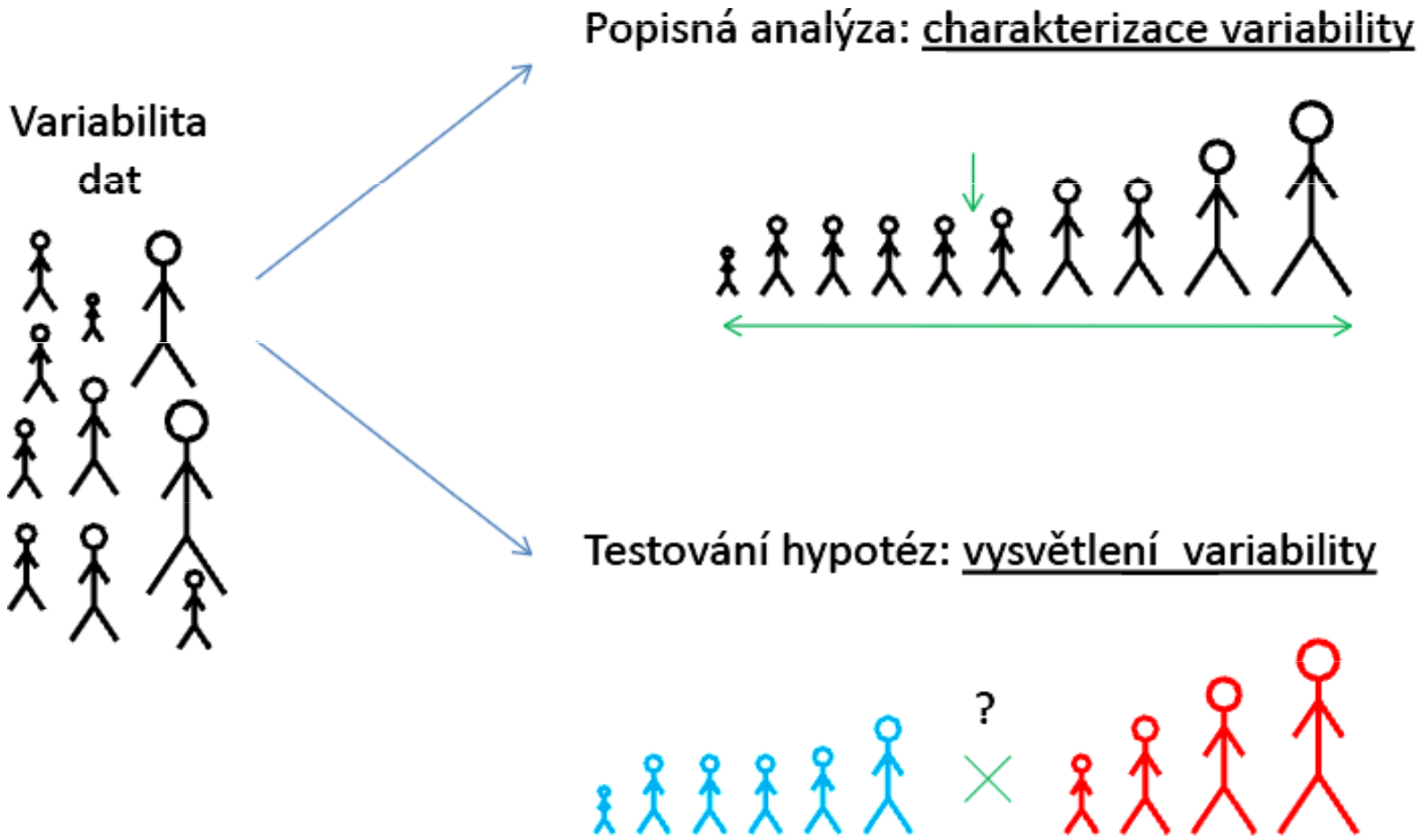


Význam statistické analýzy dat

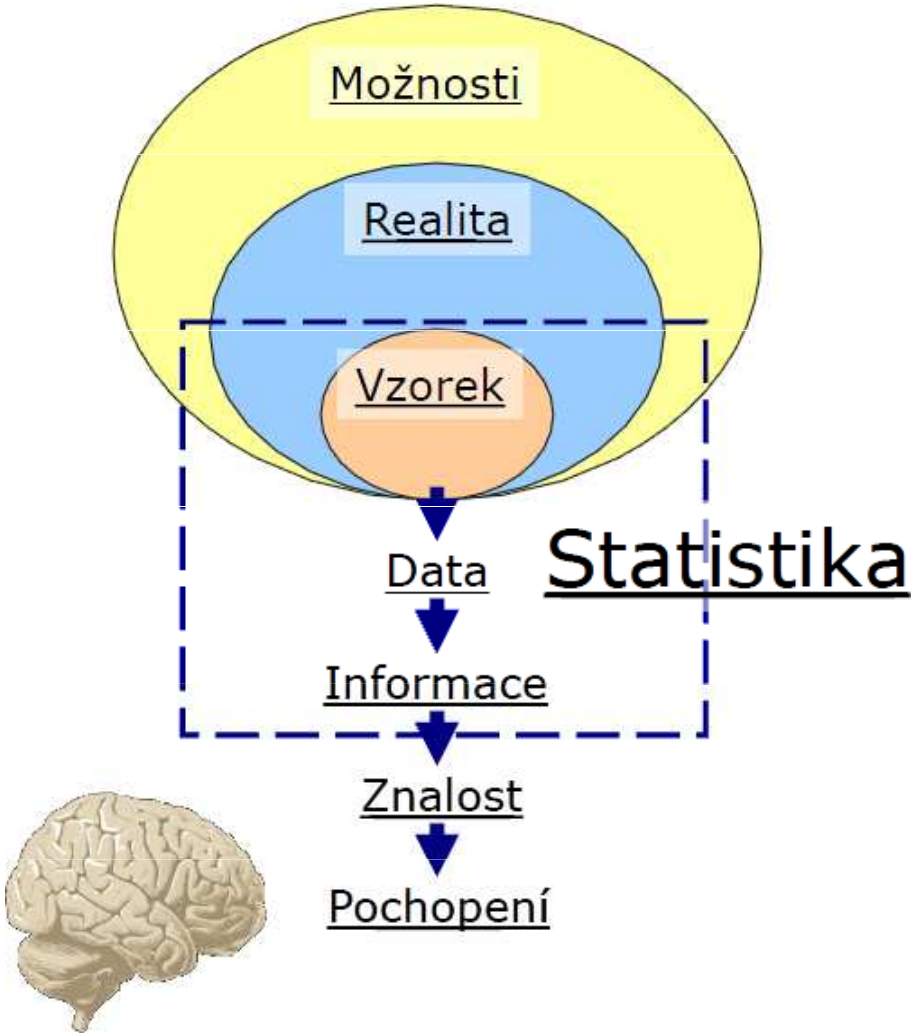
- Realita je variabilní a statistika je věda zabývající se variabilitou
- Korektní analýza variabilita a její pochopení přináší užitečné informace o realitě
- V případě deterministického světa by statistická analýza nebyla potřebná
- V případě zcela chaotického světa by nebyla možná.

Práce s variabilitou v analýze dat

- Dva hlavní přístupy k variabilitě:

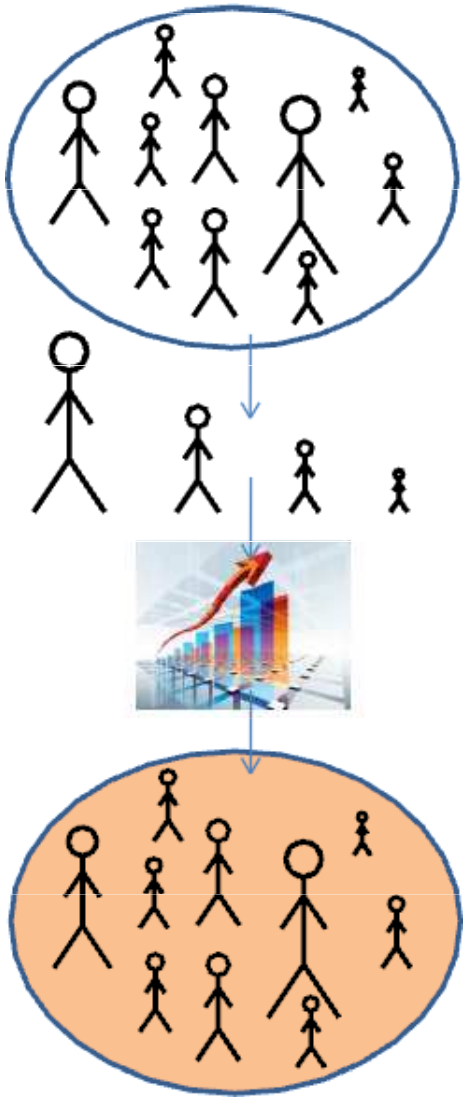


Práce s variabilitou v analýze dat



- 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

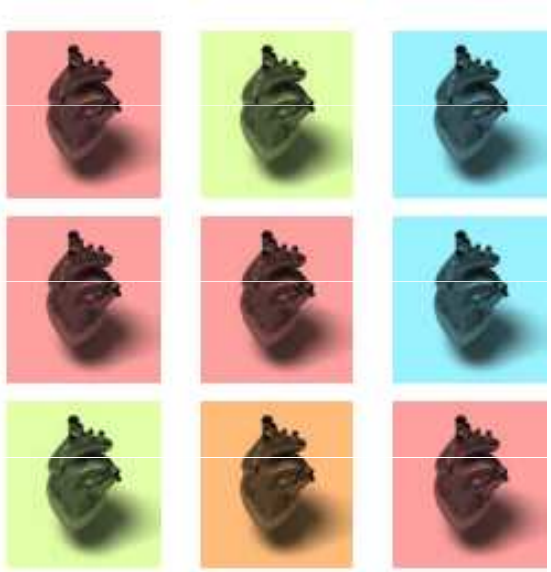
Analýza

Díky zobecnění výsledků známe vlastnosti cílové populace

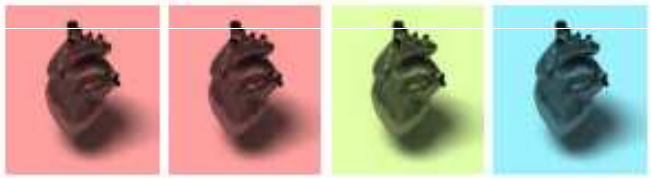
- Cílem analýzy není pouhý popis a analýza vzorku, ale zobecnění výsledků ze vzorku na jeho cílovou populaci.
- Pokud vzorek nereprezentuje cílovou populaci, vede zobecnění k chybným závěrům.

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

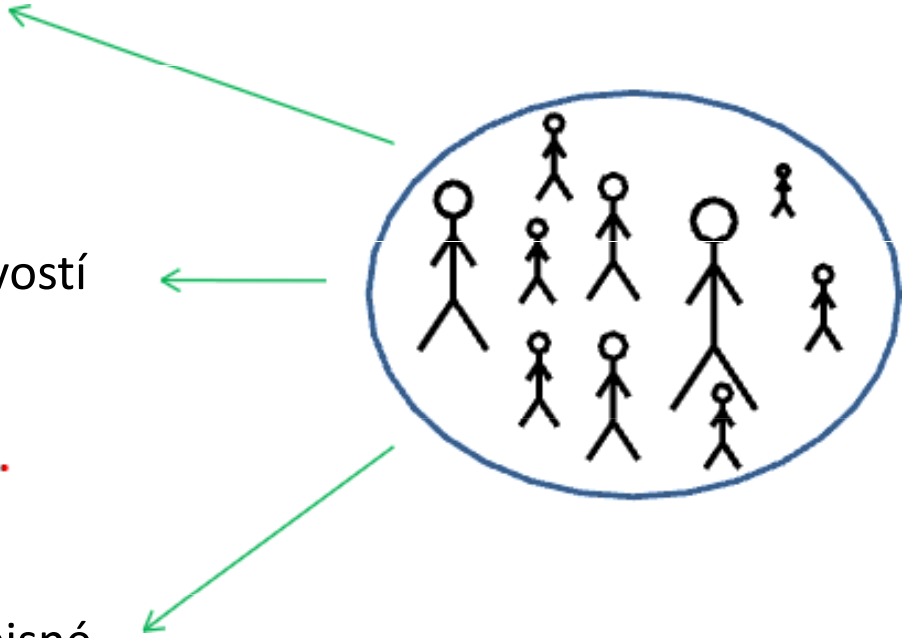
Velikost vzorku a přesnost statistických výstupů

- Existuje skutečné rozložení a skutečný průměr měřené proměnné

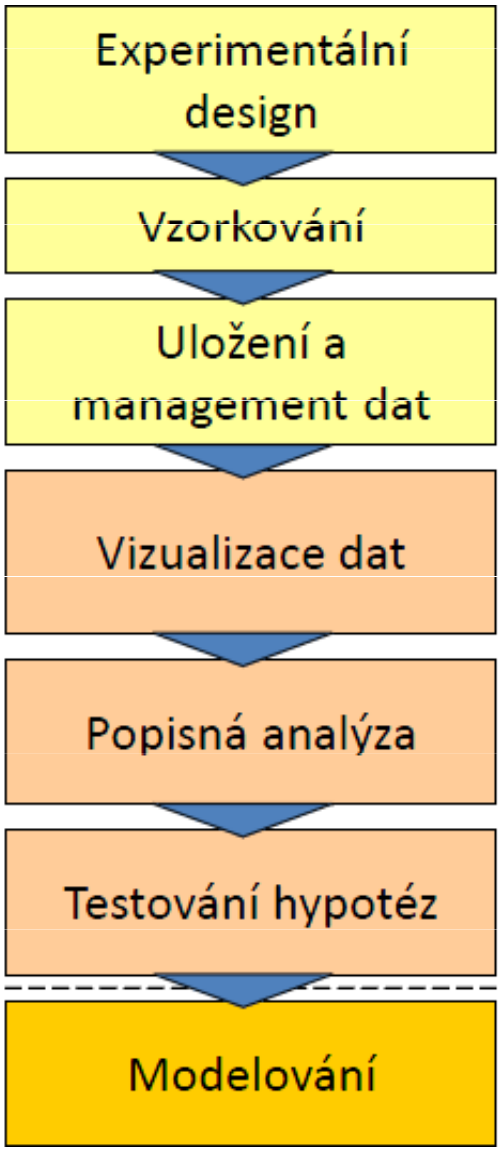
- Z jednoho měření nezjistíme nic

- 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 případech nereálný.



Obecné schéma aplikace statistické analýzy



Jak velký vzorek je nezbytný pro statisticky relevantní výsledky? Klíčová stratifikační kritéria cílové populace.

Vzorkovací plán zabezpečující náhodnost a reprezentativnost vzorku.

Uložení dat ve vhodné formě a jejich vyčištění předcházející vlastní analýze je klíčovým krokem statistické analýzy.

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 umožňuje vyhodnotit srovnáním s existující literaturou realističnost naměřených rozsahů dat.

Testování vazeb mezi různými proměnnými s cílem navzájem vysvětlit jejich variabilitu a tím přispět k pochopení řešeného problému.

Možným vyvrcholením analýzy je využití získaných znalostí a pochopení problému k vytvoření prediktivních modelů.

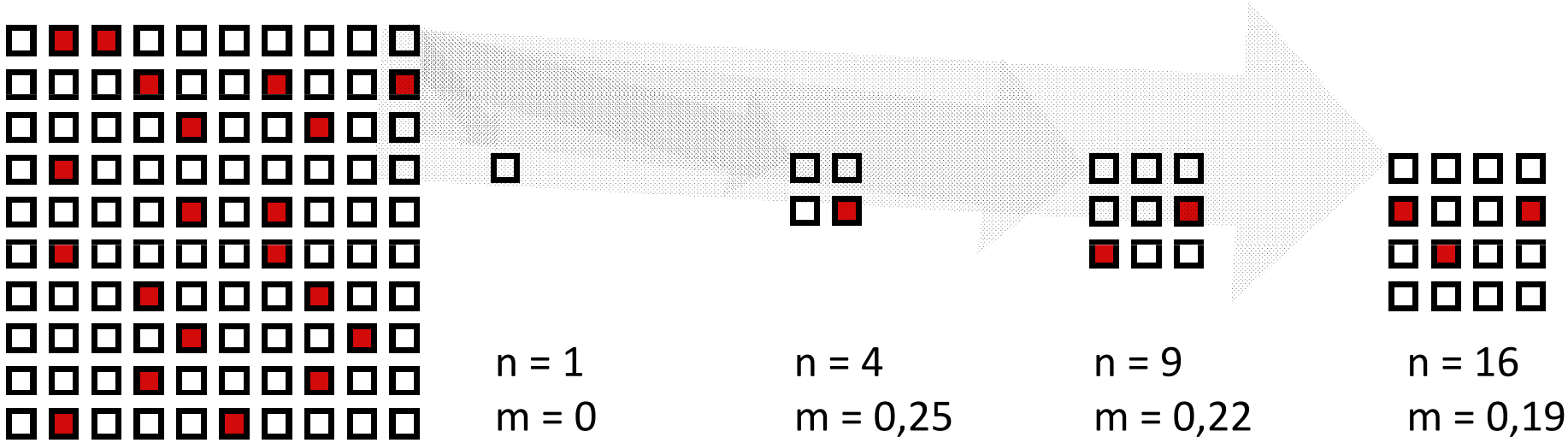
1a. Teoretické pozadí statistické analýzy

Jak vznikají informace
Rozložení dat

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

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 opakování pokusu platilo pro relativní četnost výsledku:

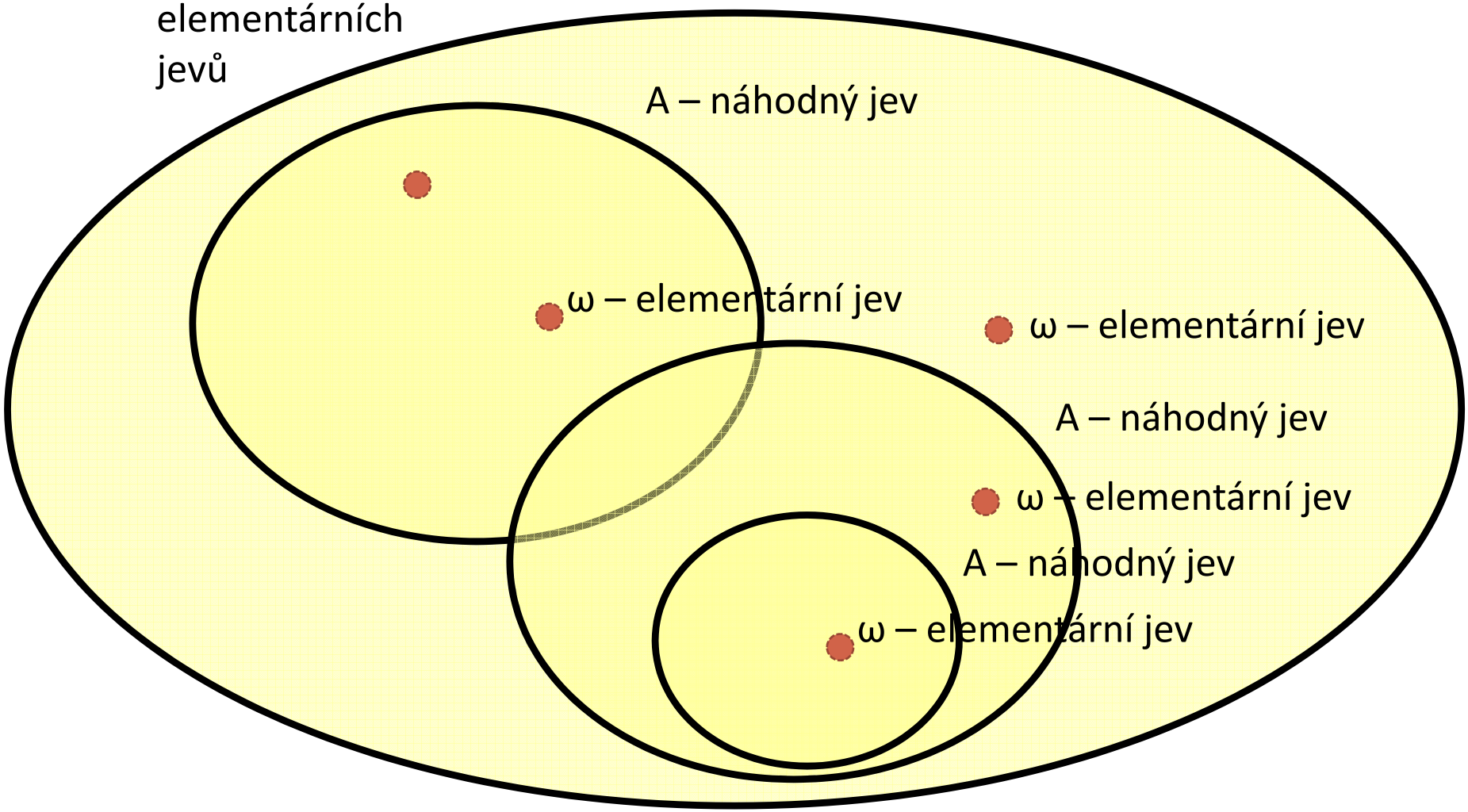
$$\lim_{n\to\infty} f(A) = \lim_{n\to\infty} \frac{n_A}{n} = p(A)$$

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

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

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

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.

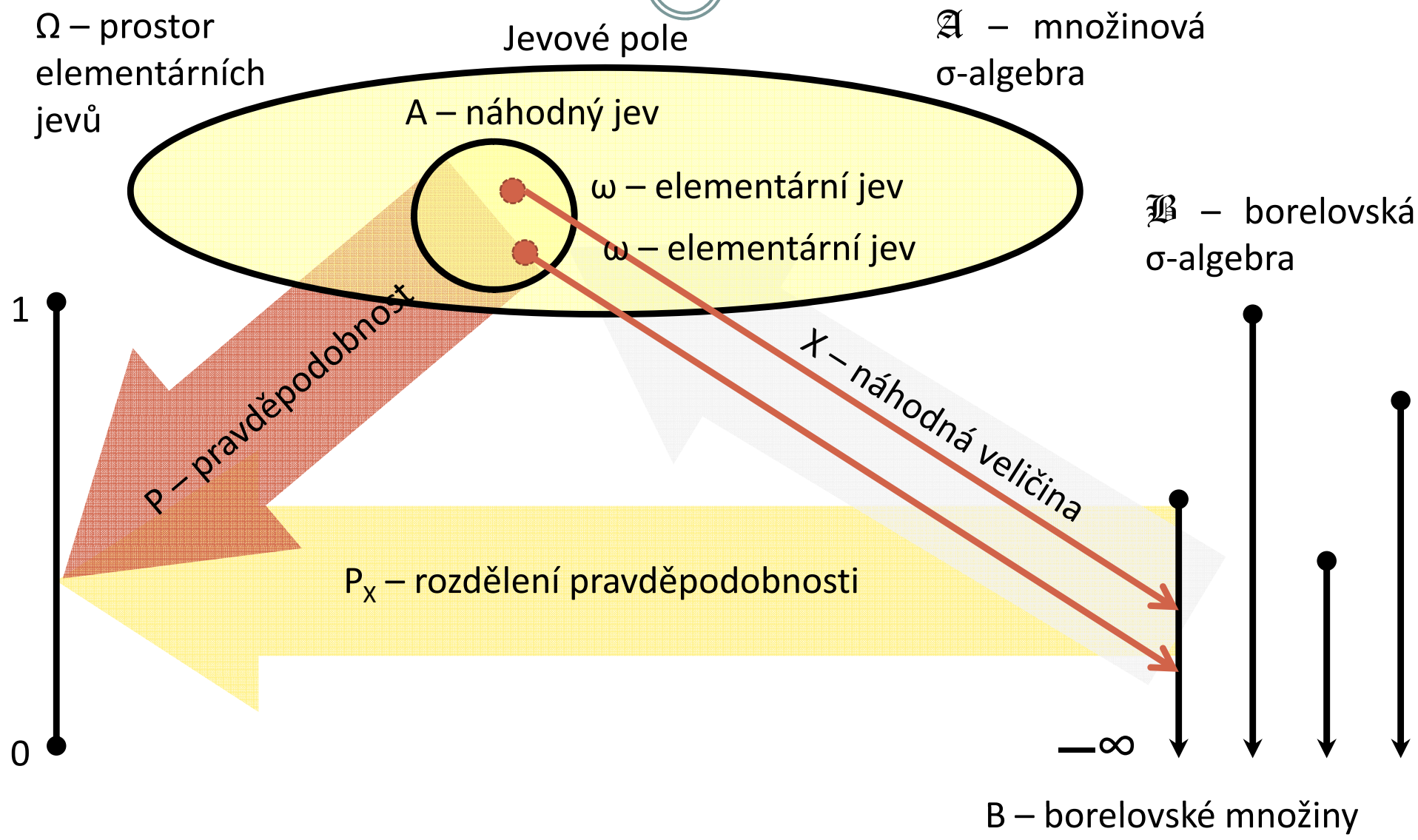
Lze ukázat, že dostatečnou podmínkou pro to, aby X byla náhodná veličina je vztah $\forall x \in \mathbb{R}$: $\{X < x\} \in \mathfrak{A}$.

Rozdělení pravděpodobnosti — množinová funkce, 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
0
P – pravděpodobnost
X – náhodná veličina
P_X – rozdělení pravděpodobnosti
−∞
B – borelovské množiny

JAK vznikají informace ? základní pojmy

Skutečnost

Náhoda
(vybere jednu z možností pokusu)

Jev

podmnožina množiny všech možných výsledků (elementárních jevů) pokusu/děje, o které lze říct, zda nastala nebo ne

Pozorovatel

Rozliší, co nastalo
a) podle možností
b) podle toho, jak potřebuje

Jevové pole

třída všech jevů, které jsme se rozhodli nebo jsme schopni sledovat

Skutečnost + Jevové pole = Měřitelný prostor

Experimentální jednotka - objekt, na kterém se provádí šetření

Populace - soubor experimentálních jednotek Znak - vlastnost sledovaná na objektu

Sledovaná veličina - číselná hodnota vyjadřující výsledek náhodného experimentu

Znak se stává náhodnou veličinou, pokud se jeho hodnota zjišťuje vylosováním objektu ze základního souboru

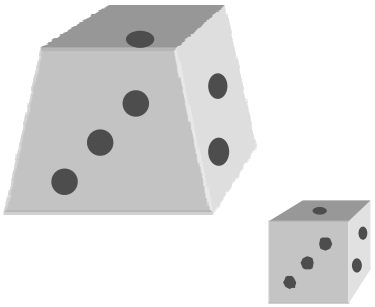
Výběr - výběrová populace - cílová populace

Náhodný výběr Reprezentativnost

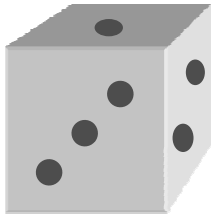
JAK vznikají informace ?

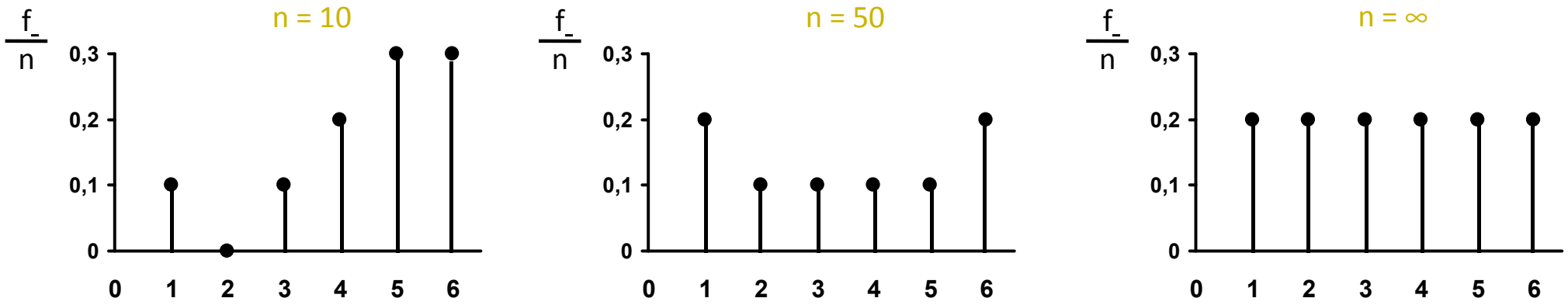
„Empirical approach“

„Classical approach“



Empirický postup





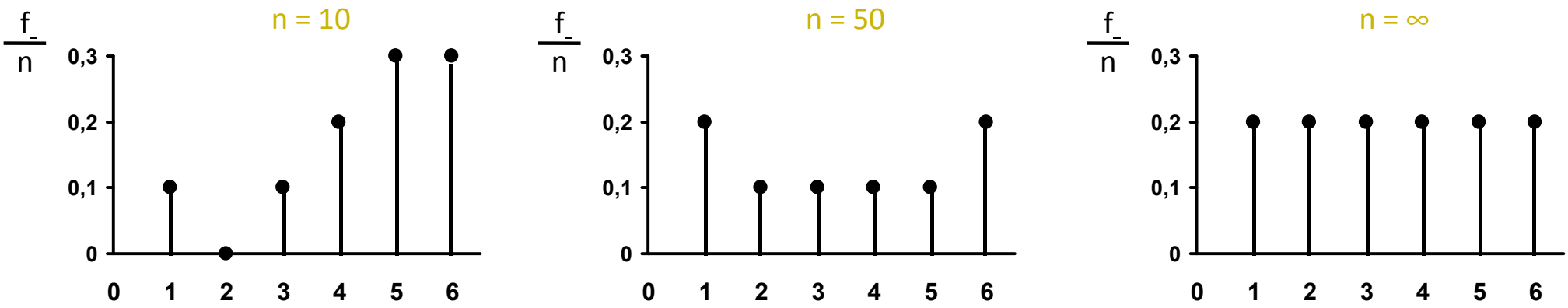
možné jevy: čísla 1 – 6

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 ?

Empirický postup



možné jevy: čísla 1 – 6

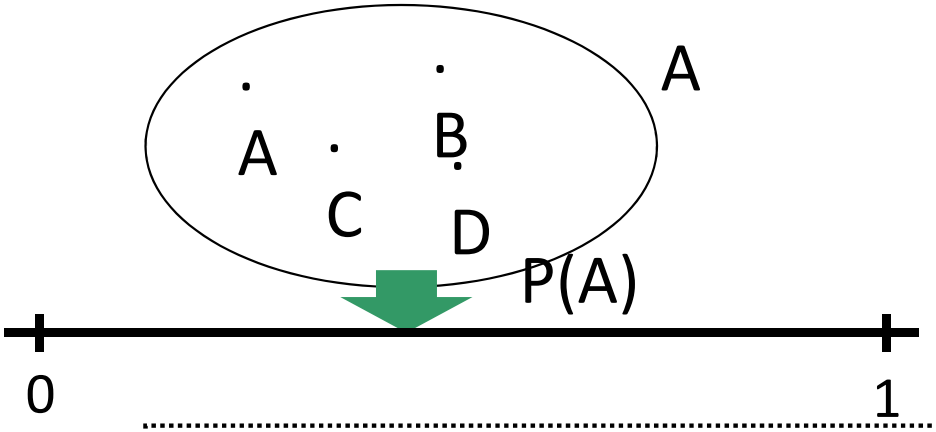
n – počet hodů (opakování)

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

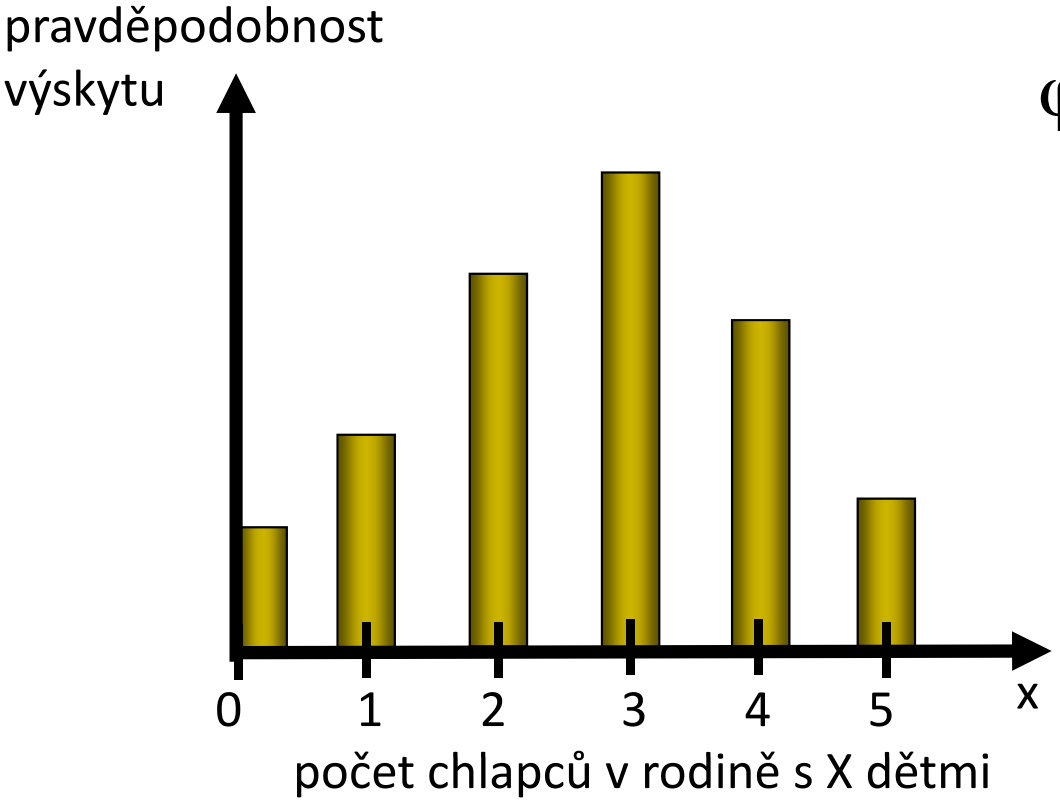
P (A ∩ B) = P (A) . P (B)..... nezávislé jevy

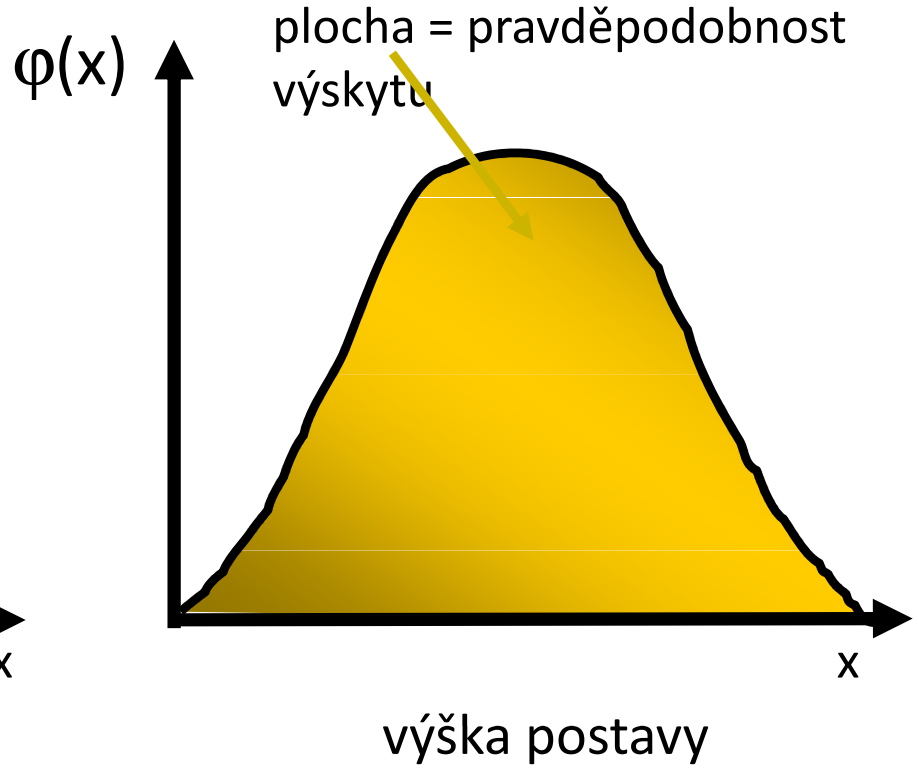
P (A ∩ B) = P (A) . P (B/A) .....závislé jevy

P (A / B) = P (A ∩ 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





2. Základní typy dat

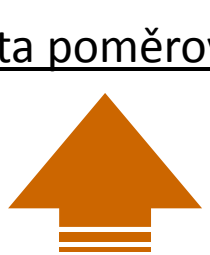
Spojitá a kategoriální data
Základní popisné statistiky
Grafický popis dat

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

Anotace

- Realitu můžeme popisovat různými typy dat, každý z nich se specifickými vlastnostmi, výhodami, nevýhodami a vlastní sadou využitelných statistických metod - od binárních přes kategoriální, ordinální až po spojitá data roste míra informace v nich obsažené.
- Základním přístupem k popisné analýze dat je tvorba frekvenčních tabulek a jejich grafických reprezentací – histogramů.

Jak vznikají informace ?
– různé typy dat znamenají různou informaci



Data poměrová

Kolikrát ?

Spojitá data

Data intervalová

O kolik ?

Data ordinální

Větší, menší ?

Kategoriální otázky

Diskrétní data

Data nominální

Rovná se ?

Otázky „Ano/Ne“

Podíl hodnot větší/menší než specifikovaná hodnota ?

Procenta odvozené hodnoty

Samotná znalost typu dat ale na dosažení informace nestačí .....

Jak vznikají informace ?
– různé typy dat znamenají různou informaci

Statistika středu

Data poměrová

Data intervalová

Data ordinální

Data nominální

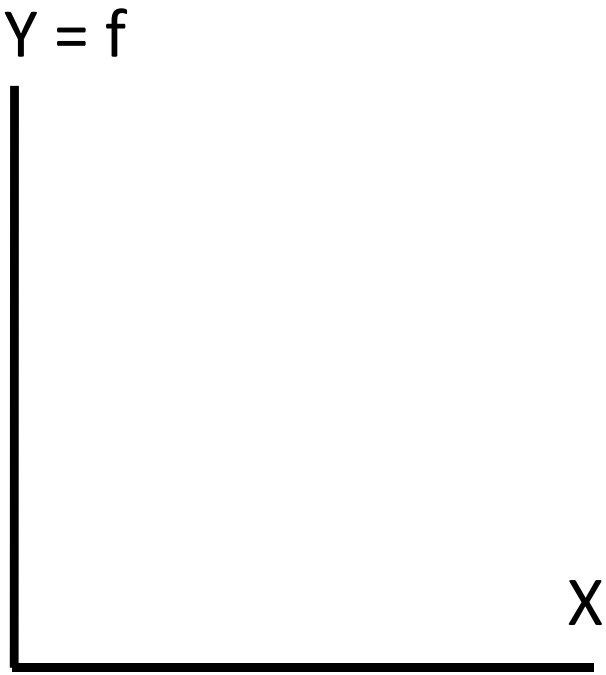
PRŮMĚR

MEDIÁN

MODUS

Spojitá data

Diskrétní data



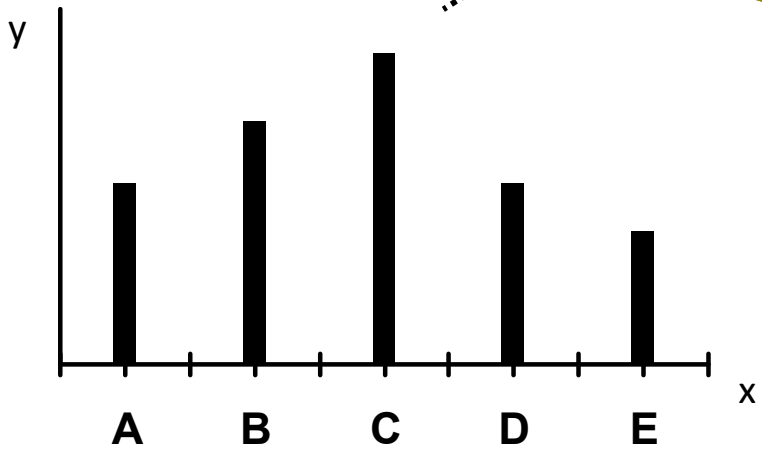
Jak vznikají informace ?
– různé typy dat znamenají různou informaci

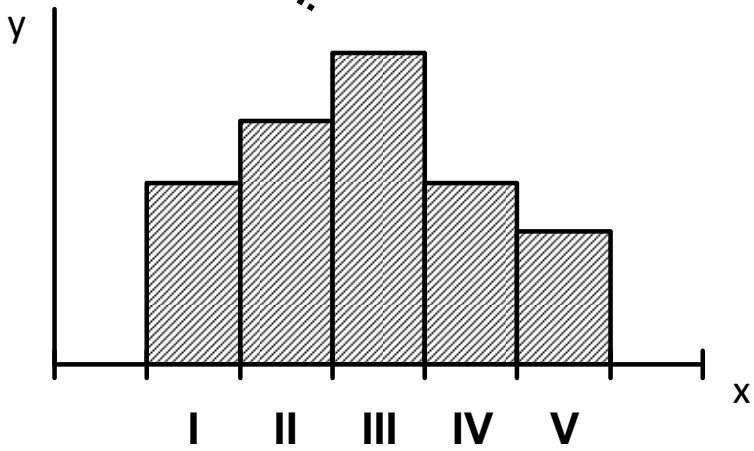
Definice průměru, směrodatné odchylky, mediánu aj.

JAK vznikají informace ?
- opakovaná měření informují rozložením hodnot

Y: frekvence -
absolutní / relativní

KOLIK se naměřilo





CO se naměřilo

X: měřený znak

Diskrétní data

Spojitá data

Odvozená data: Pozor na odvozené indexy

Příklad I:

Znak X: Hmotnost
Znak Y: Plocha

Příklad II:

X: Průměrný počet výrobků v prodejně
Y: Odhad prostoru průměrně nabízeného k vystavení výrobku

průměr : (min - max)

X: 1,2 : (1,15 - 1,24) → + / - 3,8 %

Y: 1,8 : (1,75 - 1,84) → + / - 2,5 %

$$X/Y = 0{,}667 : \left(\frac{1{,}15}{1{,}84} - \frac{1{,}24}{1{,}75}\right)$$ → + / - 6,2 %

Nová veličina má jinou šířku rozpětí než ty, ze kterých je odvozená

Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

DISKRÉTNÍ DATA

Primární data

Počty epizod pro n = 100 hemofiliků

0
0
1
2
1
1
3
1
1
2
.
.
.
.
.
.
n = 100

Frekvenční sumarizace

N: 100 dětí (hemofiliků)
x: znak: počet krvácivých epizod za měsíc

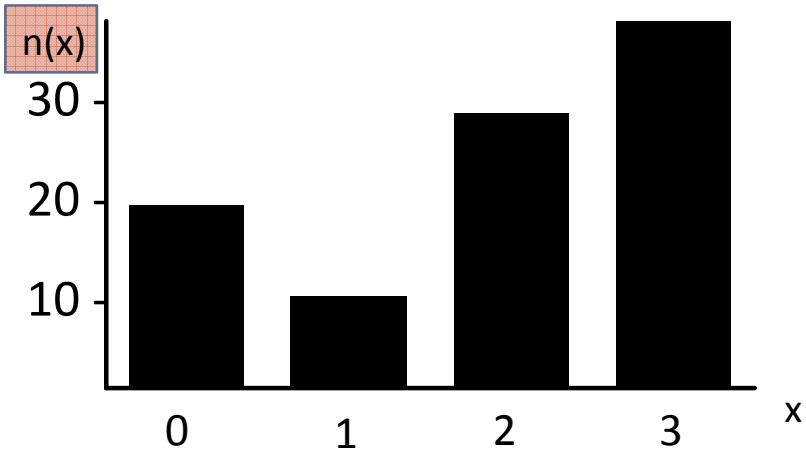
x	n(x)	N(x)	p(x)	F(x)
0	20	20	0,2	0,2
1	10	30	0,1	0,3
2	30	60	0,3	0,6
3	40	100	0,4	1,0

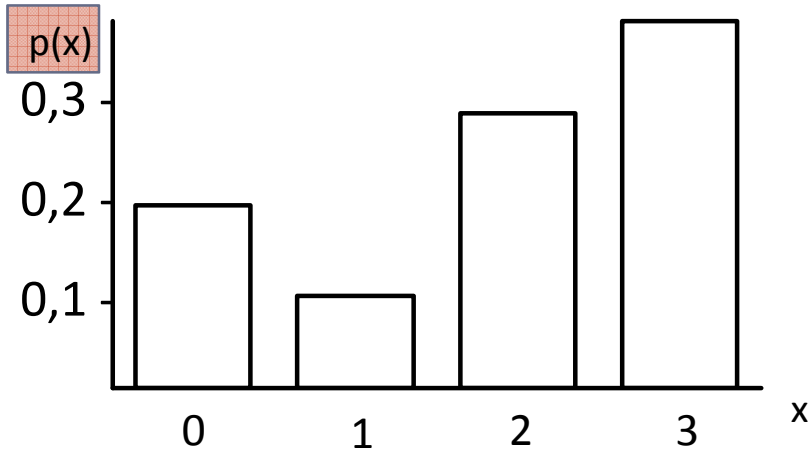
n(x) – absolutní četnost x
N(x) – kumulativní četnost hodnot nepřevyšujících x;

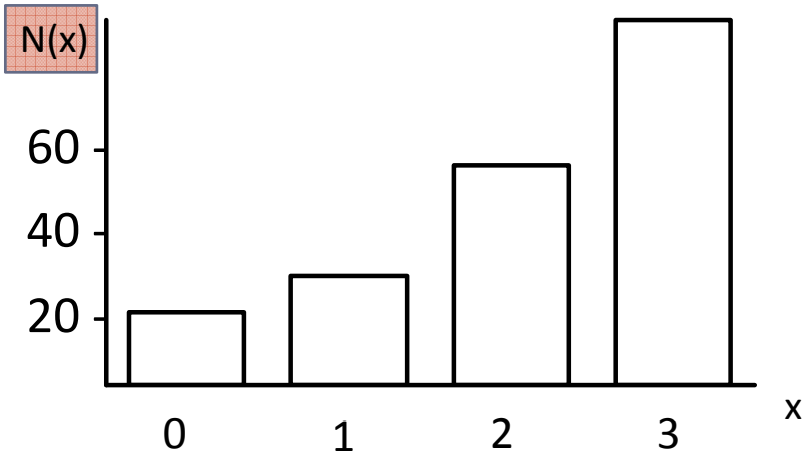
$$N(x) = \sum_{t \le x} n(t)$$

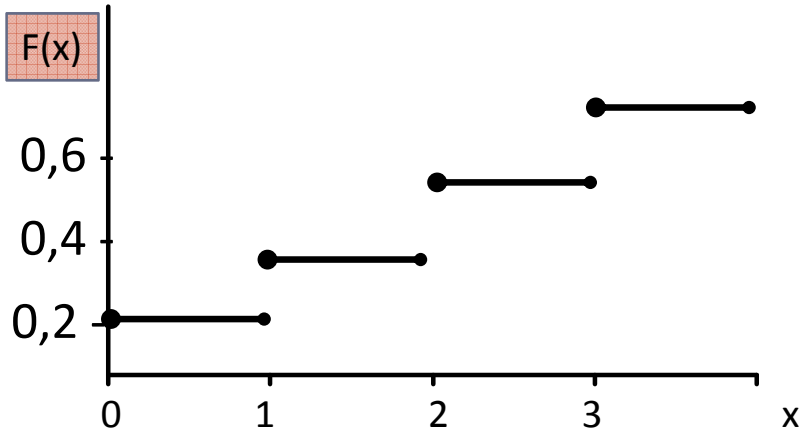
p(x) – relativní četnost; p(x) = n(x) / n
F(x) – kumulativní relativní četnost hodnot nepřevyšujících x; F(x) = N(x) / n

Jak vznikají informace ?
Grafické výstupy z frekvenční tabulky









Jak vznikají informace ?
- frekvenční tabulka jako základní nástroj popisu

SPOJITÁ DATA

Příklad: **x: koncentrace látky v krvi n = 100 pacientů**

Primární data

Hodnoty pro n = 100 osob

1,21
1,48
1,56
0,31
1,21
1,33
0,33
.
.
.
n = 100

Frekvenční sumarizace

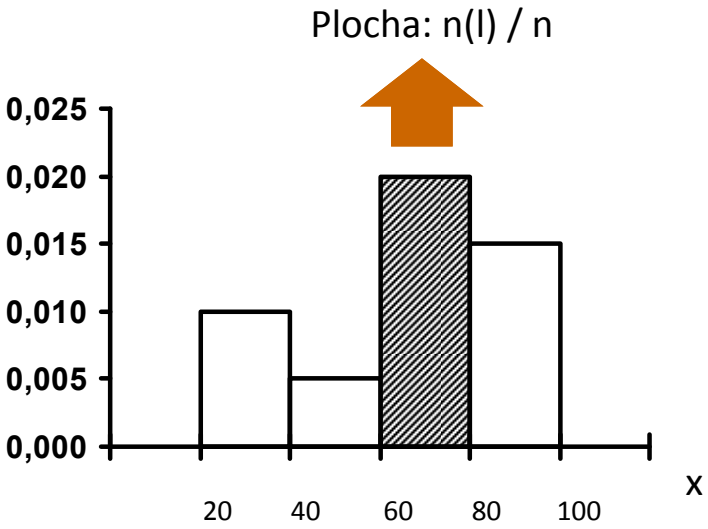
n = 100 opakovaných měření (100 pacientů)
x: koncentrace sledované látky v krvi (20 – 100 jednotek)

interv	d(l)	n(l)	n(l)/n	N(x'')	F(x'')
<20, 40)	20	20	0,2	20	0,2
<40, 60)	20	10	0,1	30	0,3
<60, 80)	20	40	0,4	70	0,7
<80, 100)	20	30	0,3	100	1,0

d(l) – šířka intervalu
n(l) – absolutní četnost
n(l) / n – intervalová relativní četnost
N(x'') – intervalová kumulativní četnost do horní hranice X''
F(x'') – intervalová relativní kumulativní četnost do horní hranice X''

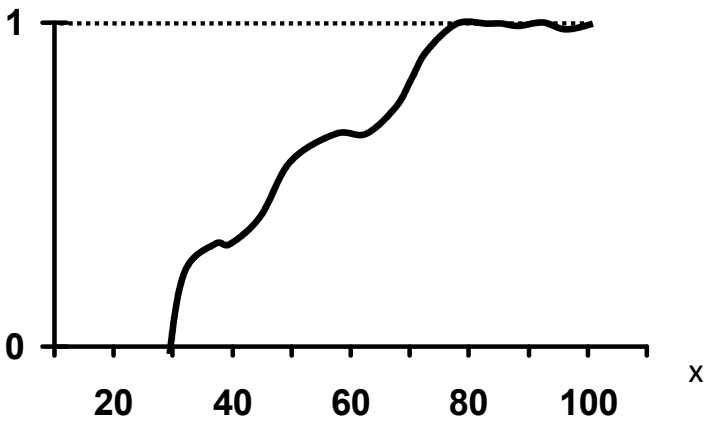
Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Histogram



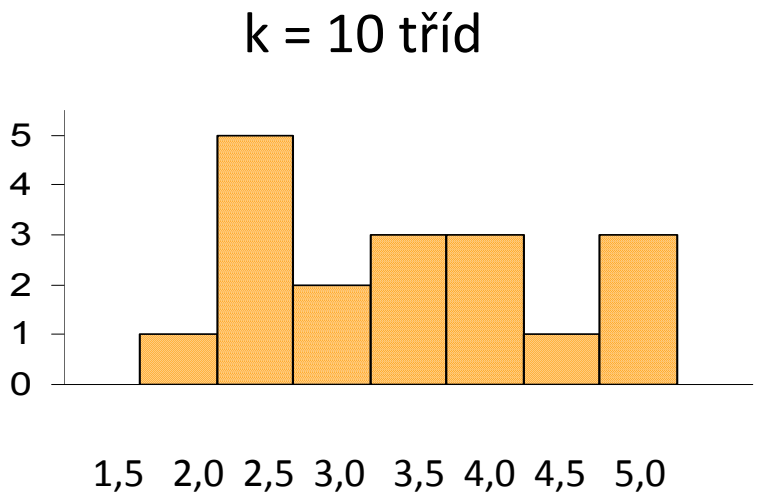
$$f(x) = \frac{n(l) / n}{d(l)} \Rightarrow$$ Intervalová hustota četnosti

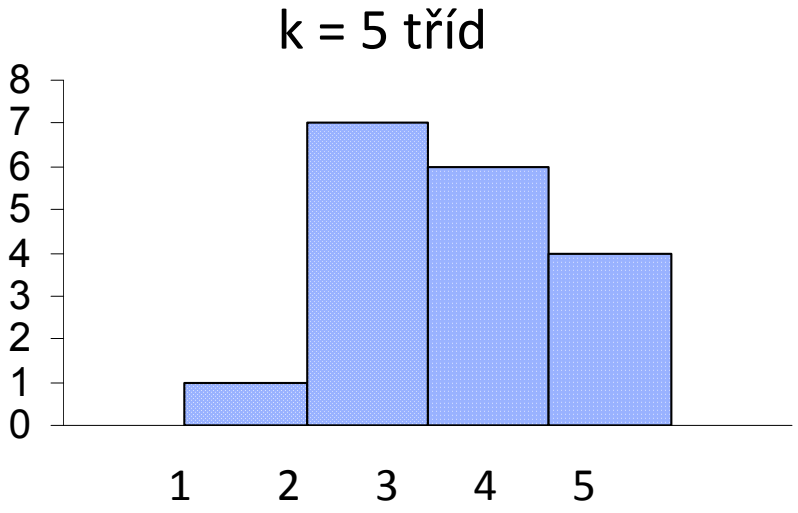
Výběrová distribuční funkce

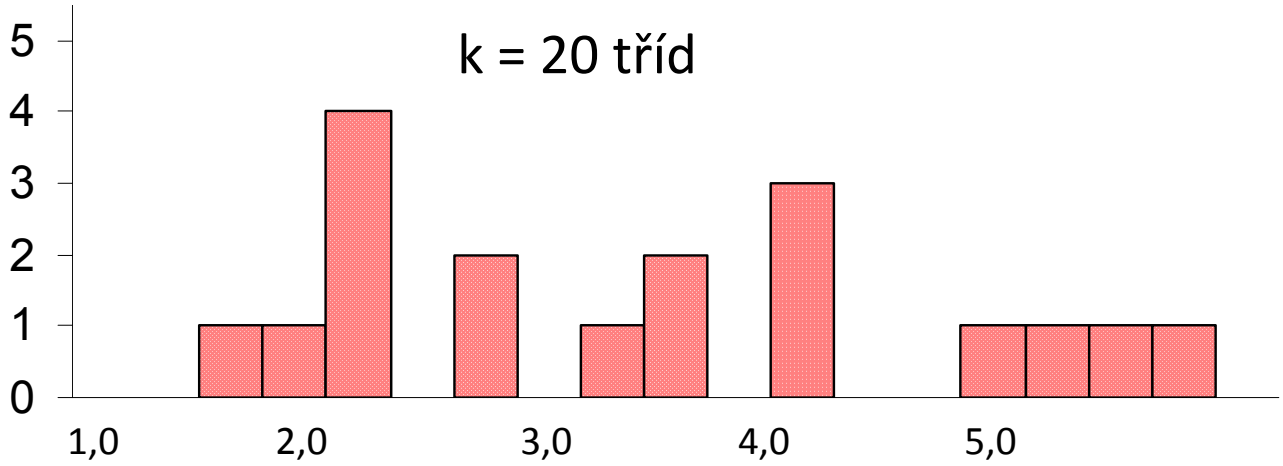


$F(x) \Rightarrow$ Intervalová relativní kumulativní četnost

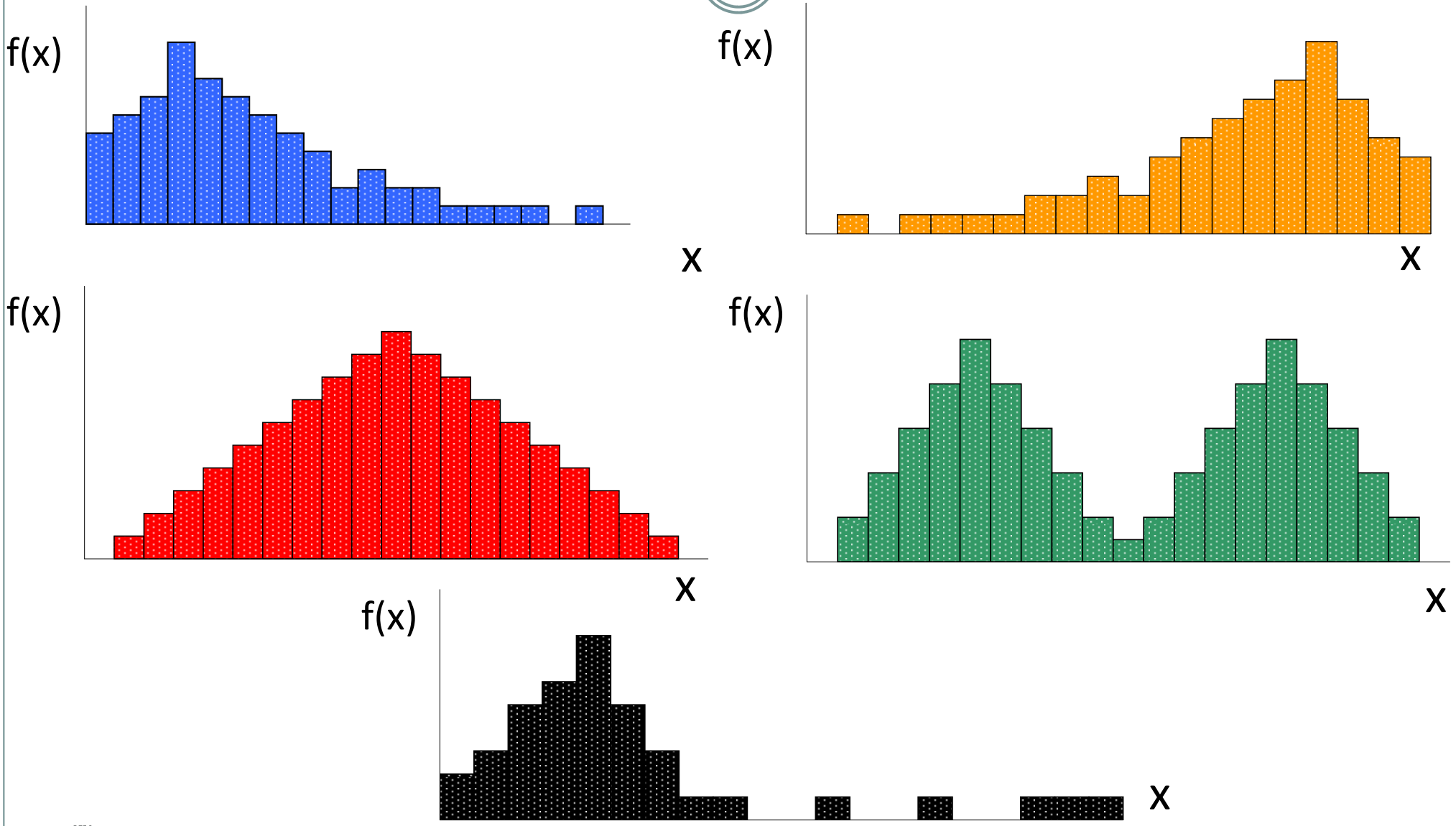
Počet zvolených tříd a velikost souboru určují kvalitu výstupu



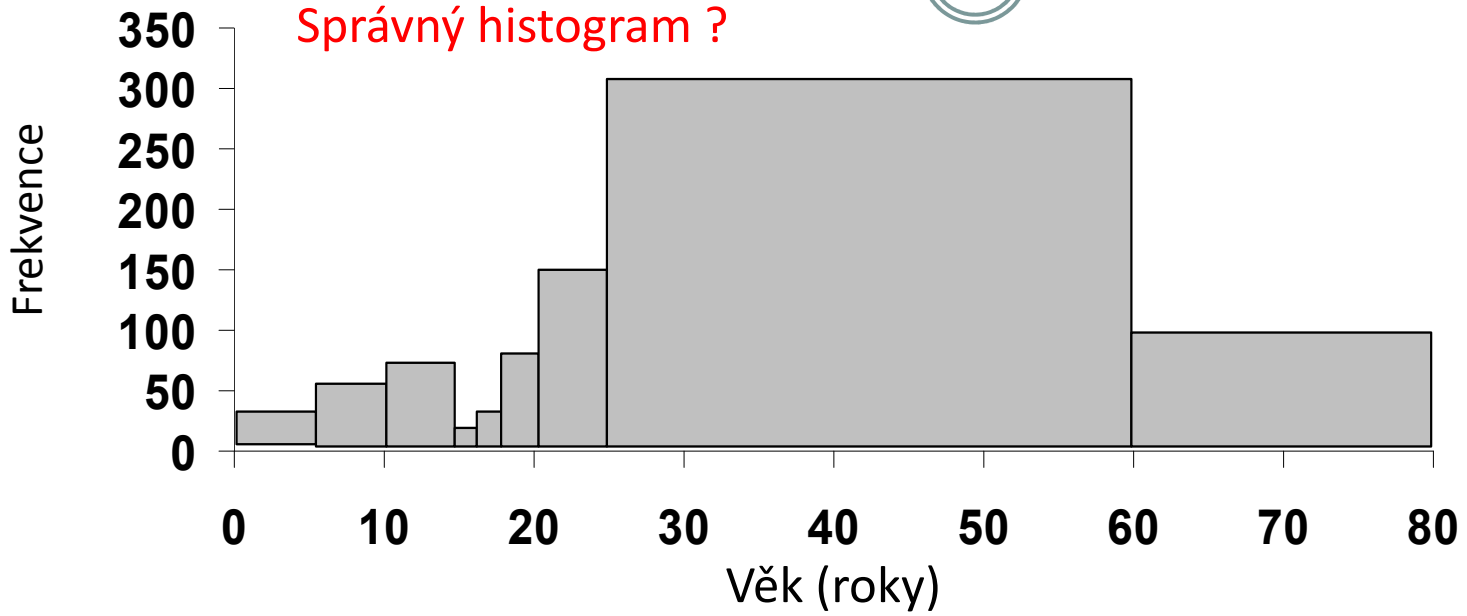


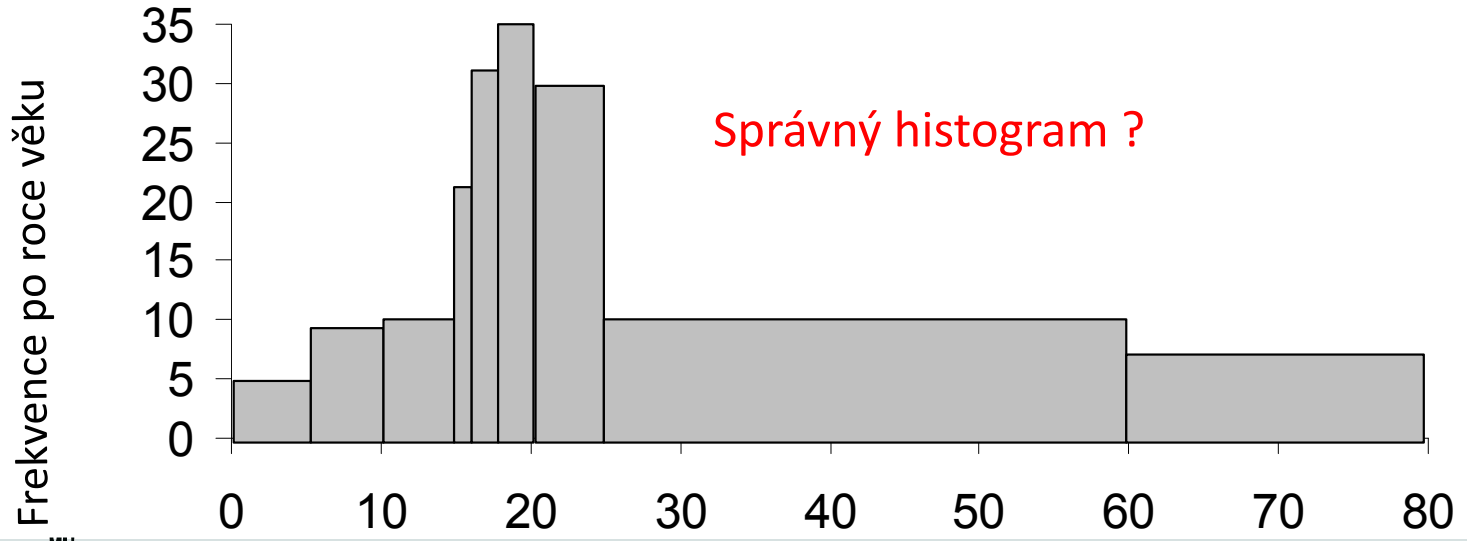


Histogram vyjadřuje tvar výběrového rozložení



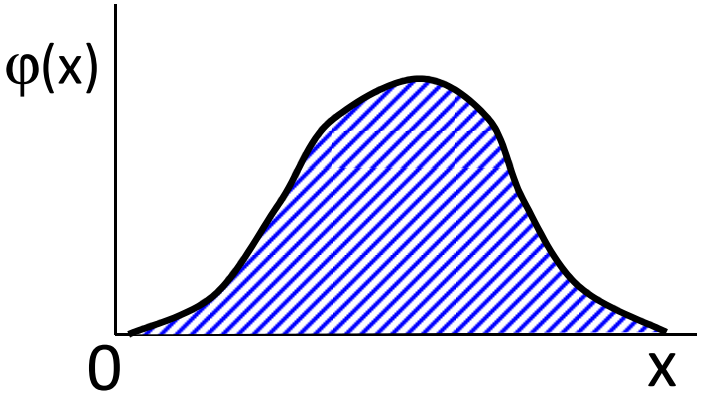
Příklad: věk účastníků vážných dopravních nehod



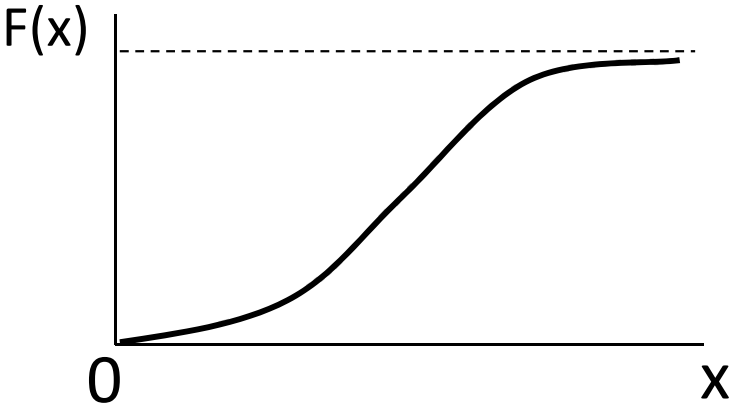


Věk	f
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103

Pojem ROZLOŽENÍ - příklad spojitých dat



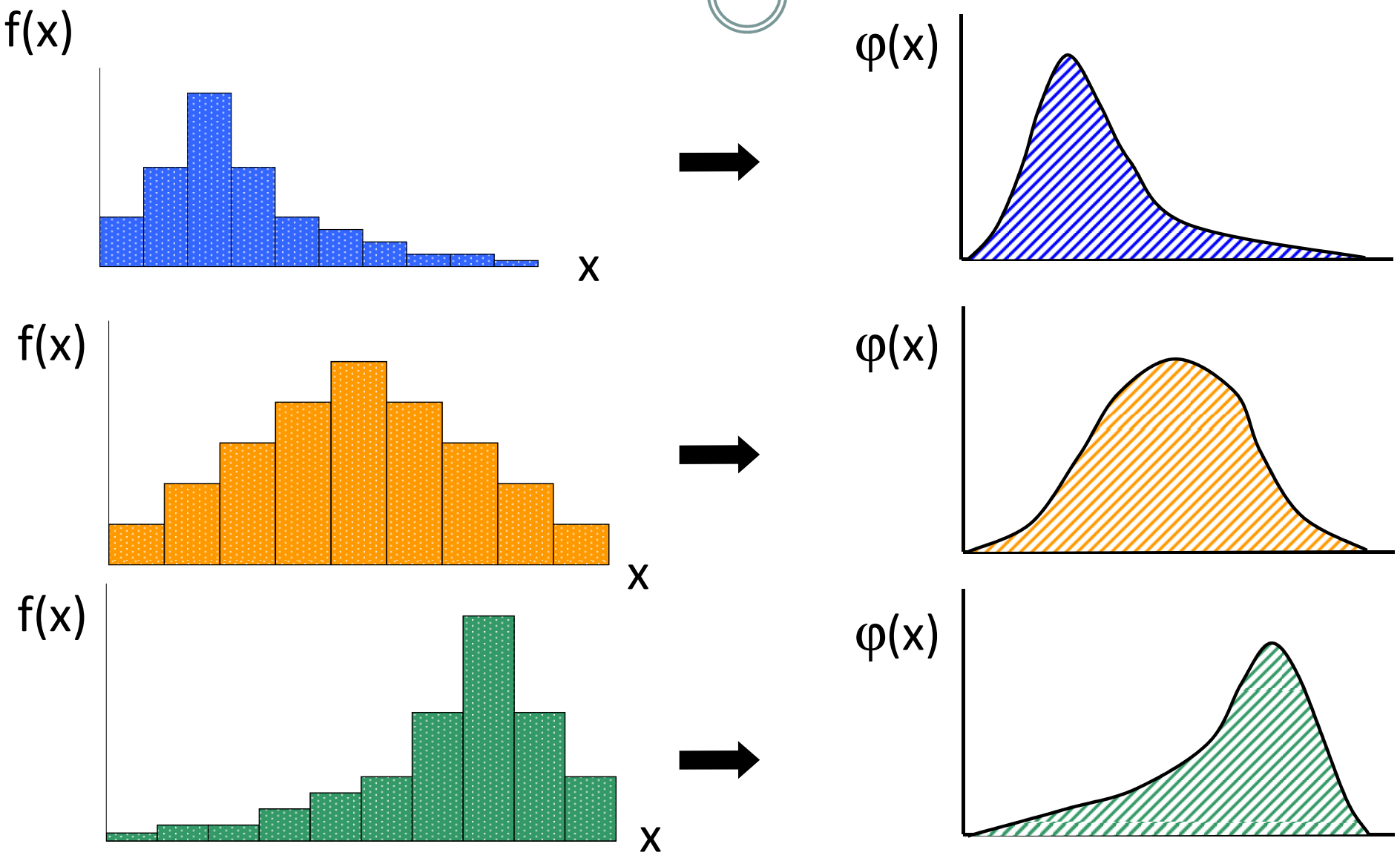
Rozložení



Distribuční funkce

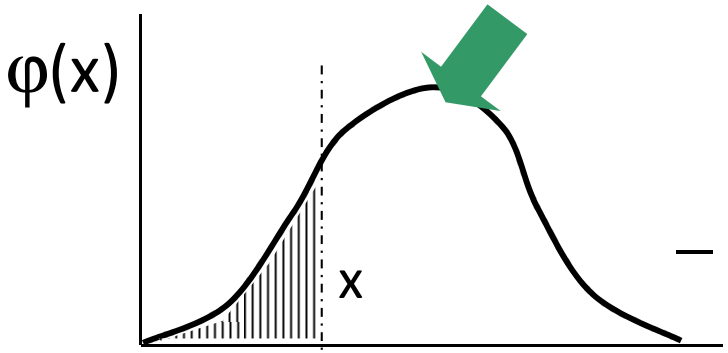
Je - li dána distribuční funkce,
je dáno rozložení

Výběrové rozložení hodnot lze modelově popsat a definovat tak pravděpodobnost výskytu X

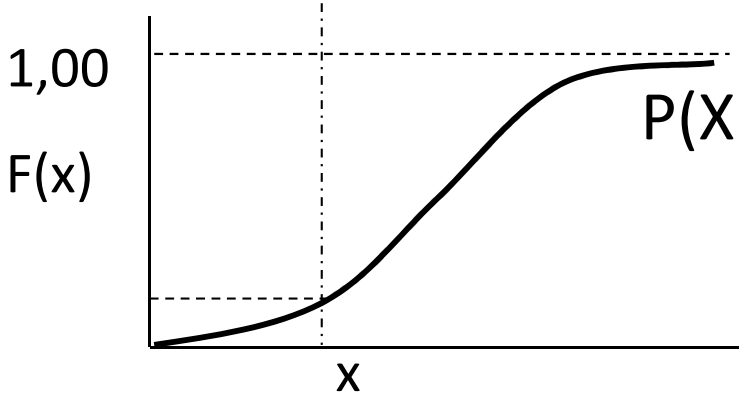


Distribuční funkce jako užitečný nástroj pro práci s rozložením

Plocha = relativní četnost



$$\int_{-\infty}^{\infty} \varphi(x)\ d(x) = 1$$

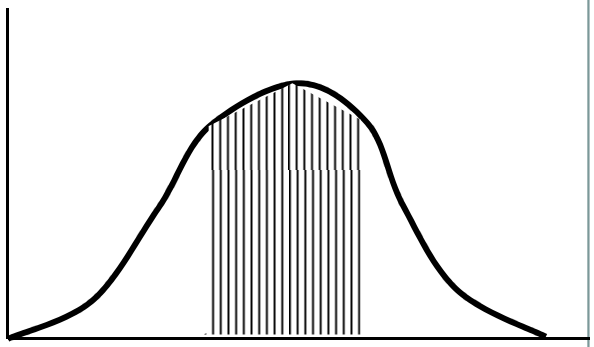


$$P(X \leq x) = \Phi(x) = F(x'')$$

$\Phi(x)$ … distribuční funkce

$$P(X \leq x) = \int_{M} \varphi(x)\ d(x)$$

F(x): Pravděpodobnost, že se X vyskytuje v intervalu M



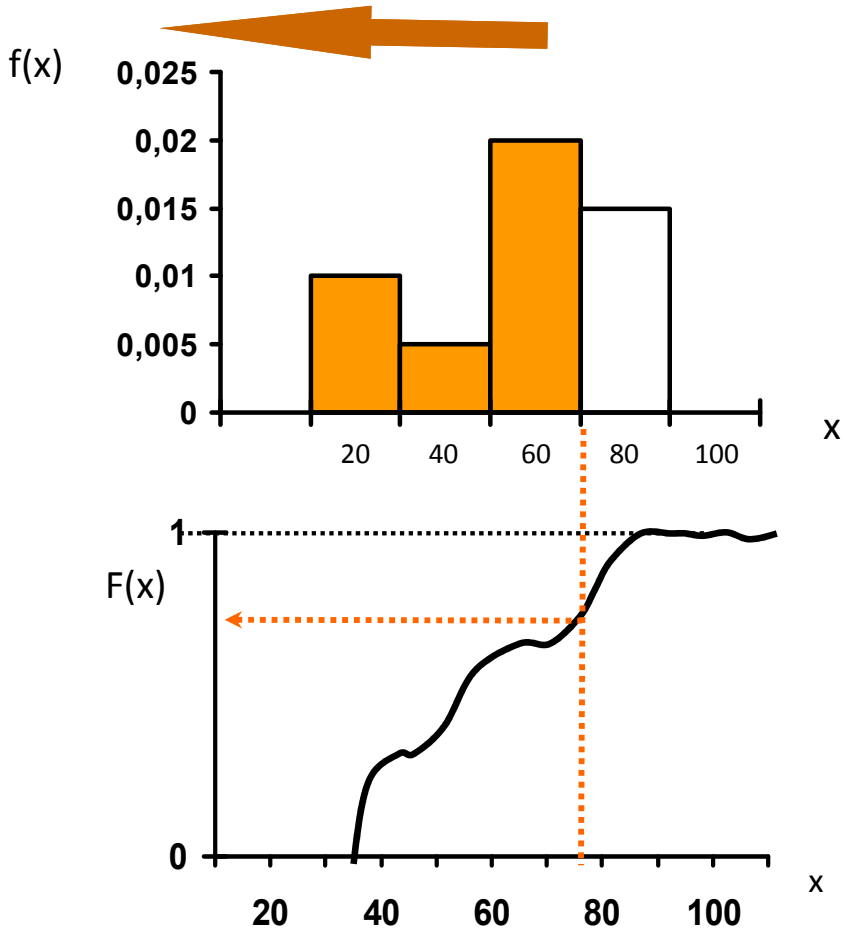
M

Známe-li distribuční funkci, pak známe rozložení sledované veličiny.

Pro jakoukoli množinu hodnot (M) lze určit P, že X do této množiny patří.

Jak vznikají informace ?
- frekvenční sumarizace spojitých dat

Grafické výstupy z frekvenční tabulky – spojitá data



Uspořádání čísel podle velikosti a konstrukce rozložení umožňuje pravděpodobnostní zařazení každé jednotlivé hodnoty

KVANTIL

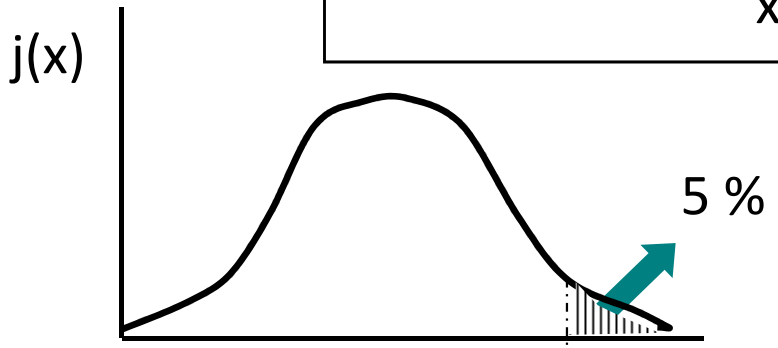
$X_{0.1}; X_{0.9}; X_{0.5}; X_{\theta}$

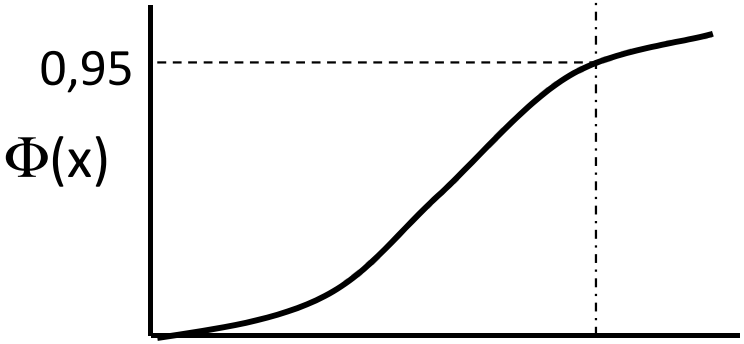
Otázka: Jak velké musí být X, aby 5 % všech hodnot bylo nad ním?

θ = 0,95 … Pravděpodobnost

Hledáme: $P(X \leq x_\theta) = 0,95 = \theta$

$x_\theta = (X_{0,95}) = ?$





$F(x\theta) = \theta$

Kvantil je číslo, jehož hodnota distribuční funkce je rovna P, pro kterou je kvantil definován

Jakékoliv číslo na ose x je kvantilem