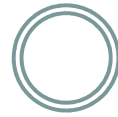


# 5. Teoretické pozadí statistické analýzy



**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.

# 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:

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

**Prostor elementárních jevů** značíme obvykle  $\Omega$ , 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  $\omega$ .

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

# Definice



$\Omega$  – prostor  
elementárních  
jevů

$A$  – náhodný jev

$\omega$  – elementární jev

$\omega$  – elementární jev

$A$  – náhodný jev

$\omega$  – elementární jev

$A$  – náhodný jev

$\omega$  – elementární jev

# Definice



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

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

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

**Pravděpodobnost (podle Kolmogorova)** reálná množinová funkce  $P$  definovaná na množině  $\mathcal{A}$   $\sigma$ -algebry  $(\Omega, \mathcal{A})$  tak, že jsou dodrženy následující podmínky:

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

# Definice



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

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

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

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

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

**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  $\forall x \in \mathbb{R}: \{X < x\} \in \mathcal{A}$ .

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

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

**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



$\Omega$  – prostor elementárních jevů

$\mathcal{A}$  – množinová  $\sigma$ -algebra

$A$  – náhodný jev

$\omega$  – elementární jev

$\omega$  – elementární jev

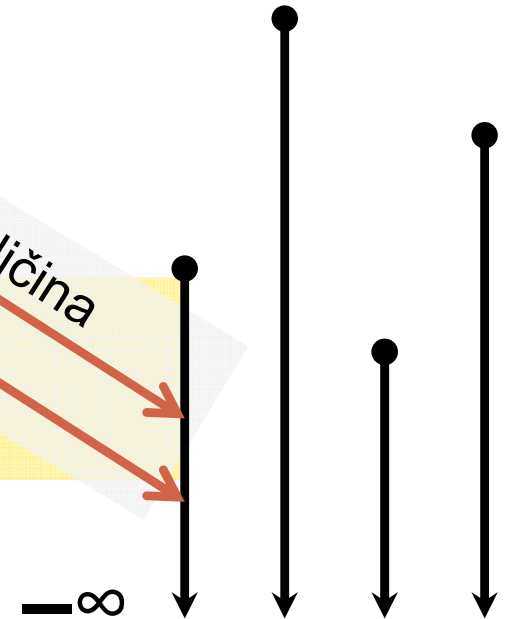
$\mathcal{B}$  – borelovská  $\sigma$ -algebra



$P$  – pravděpodobnost

$P_X$  – rozdělení pravděpodobnosti

$X$  – náhodná veličina



$B$  – borelovské množiny



# JAK vznikají informace ? základní pojmy

Skutečnost

Pozorovatel

**Náhoda**

(vybere jednu z možností pokusu)

**Jev**

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

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

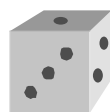
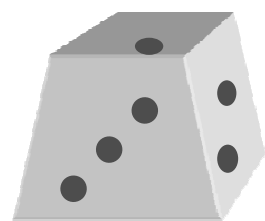
**Náhodný výběr**

**Reprezentativnost**

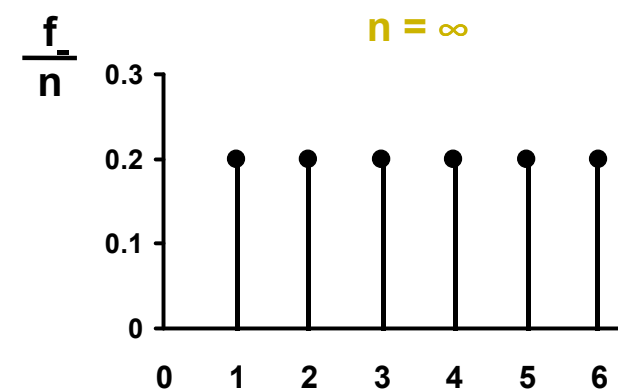
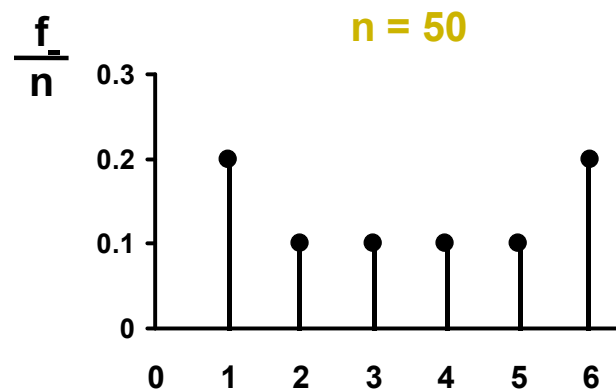
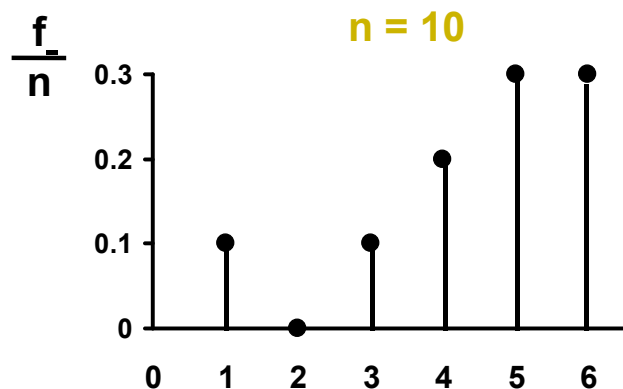
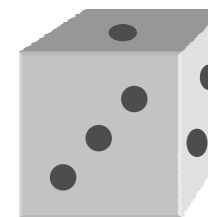
# JAK vznikají informace ?

„Empirical approach“

„Classical approach“



Empirický postup

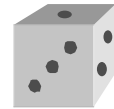
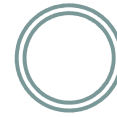


možné jevy: čísla 1 – 6

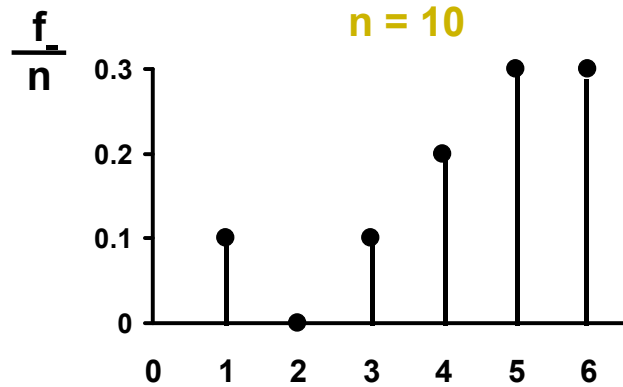
$n$  – počet hodů (opakování)

**U složitých stochastických systémů se pravda získá až po odvedení značného množství experimentální práce: musíme dát systému šanci se projevit**

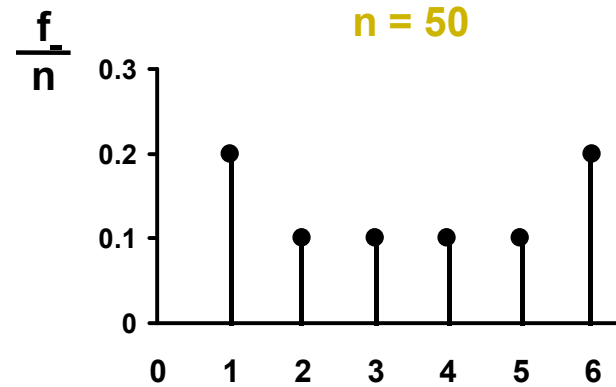
# JAK vznikají informace ?



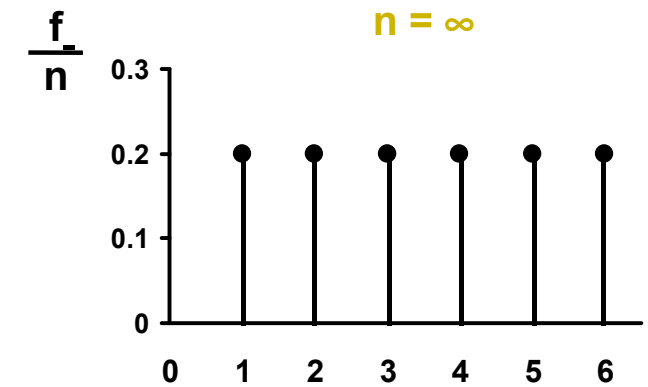
Empirický postup



možné jevy: čísla 1 – 6



$n$  – počet hodů (opakování)



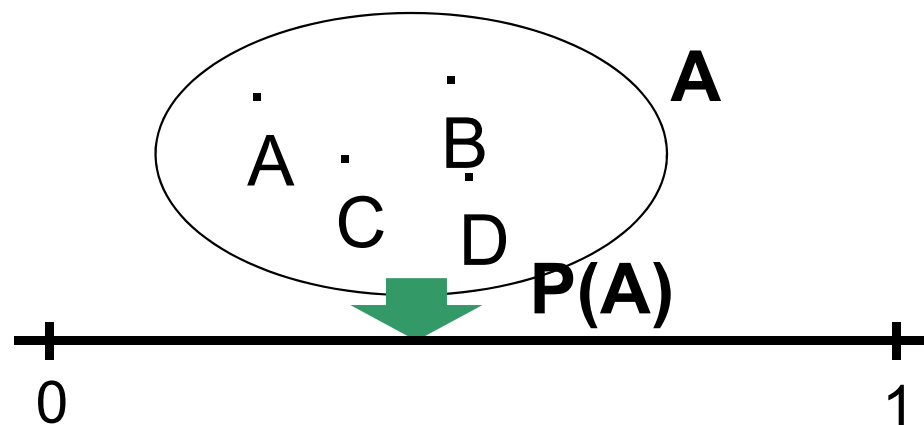
Při realizaci náhodného experimentu roste se zvyšujícím se počtem opakování pravdivá znalost systému (výsledky se stávají stabilnější) ...diskutabilní je ale ovšem míra zobecnění konkrétního experimentu

# 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**

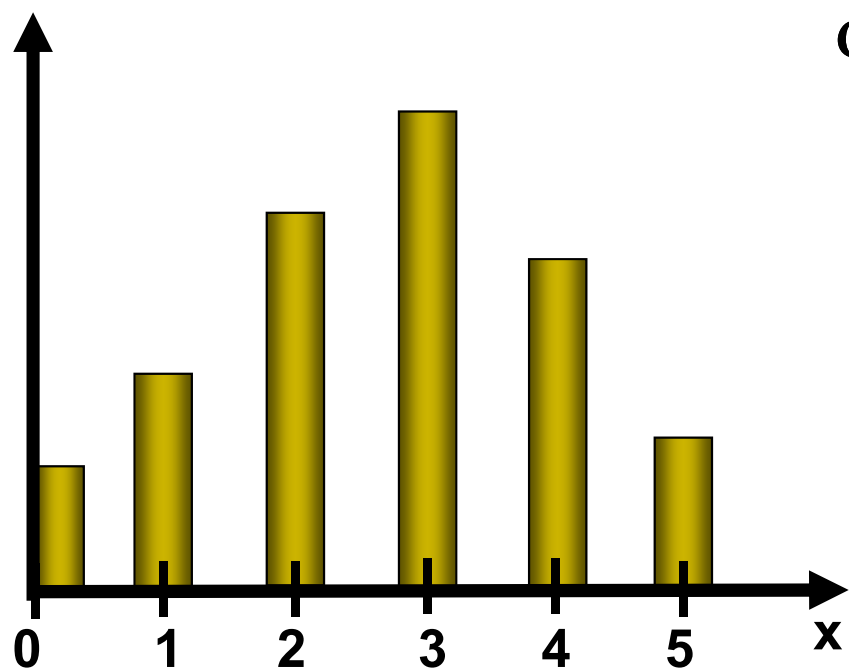
- $P(A) = 1$  ..... jev jistý
- $P(A) = 0$  ..... jev nemožný
- $P(A \cap B) = P(A) \cdot P(B)$  ..... nezávislé jevy
- $P(A \cap B) = P(A) \cdot P(B/A)$  ..... závislé jevy
- $P(A/B) = P(A \cap B) / P(B)$  ..... podmíněná pravděpodobnost

# 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

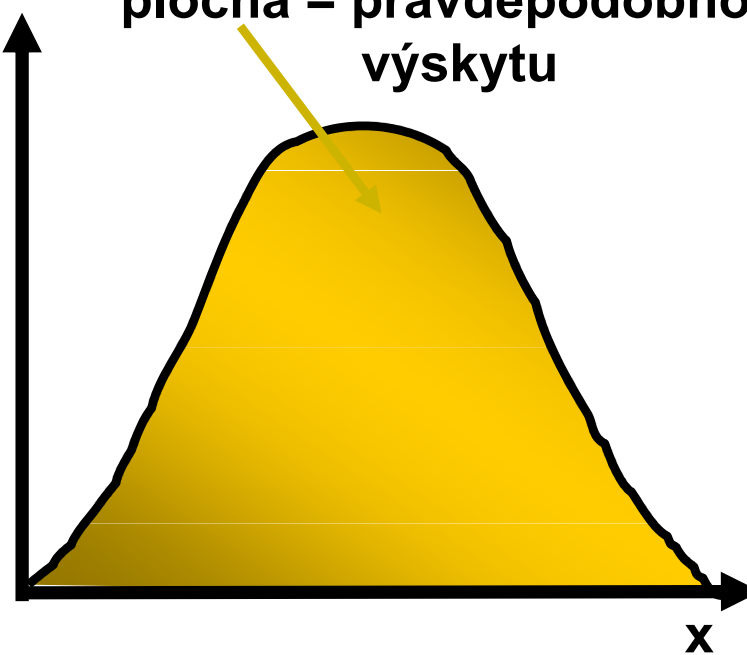
pravděpodobnost  
výskytu



počet chlapců v rodině s X dětmi

$\varphi(x)$

plocha = pravděpodobnost  
výskytu



výška postavy

# 5a. Základní typy dat



**Spojitá a kategoriální data**  
**Základní popisné statistiky**  
**Grafický popis dat**

# 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 ?



Data intervalová

O kolik ?



Data ordinální

Větší, menší ?



Data nominální

Rovná se ?

Spojité data

Diskrétní data

Kategoriální otázky

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



**PRŮMĚR**

**Spojitá  
data**

Data intervalová



**MEDIÁN**

Data ordinální

**Diskrétní  
data**

Data nominální

**MODUS**

$Y = f$

X

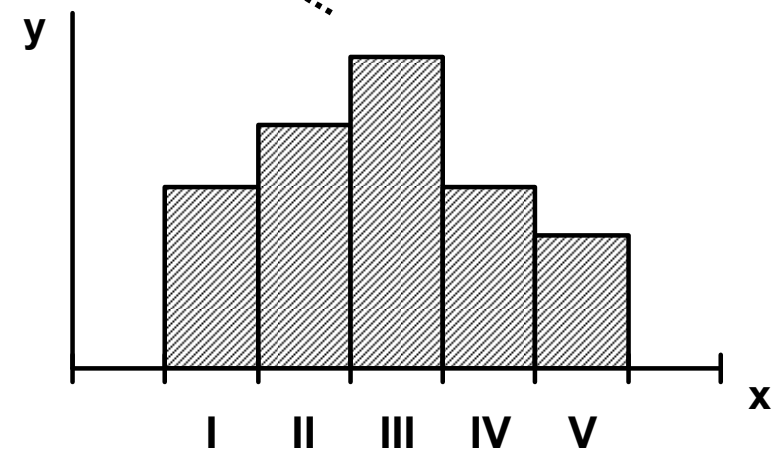
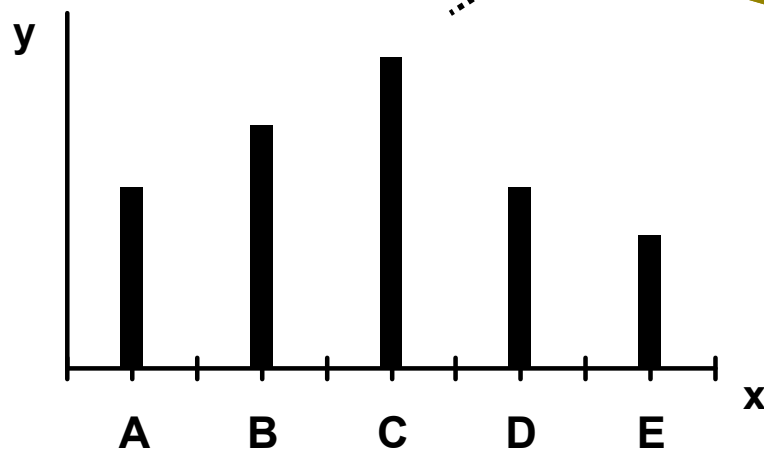
# JAK vznikají informace ?

- opakovaná měření informují rozložením hodnot

**Y: frekvence**

- absolutní / relativní

**KOLIK se naměřilo**



**CO se naměřilo**

**X: měřený znak**

Diskrétní data

Spojité 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  
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;

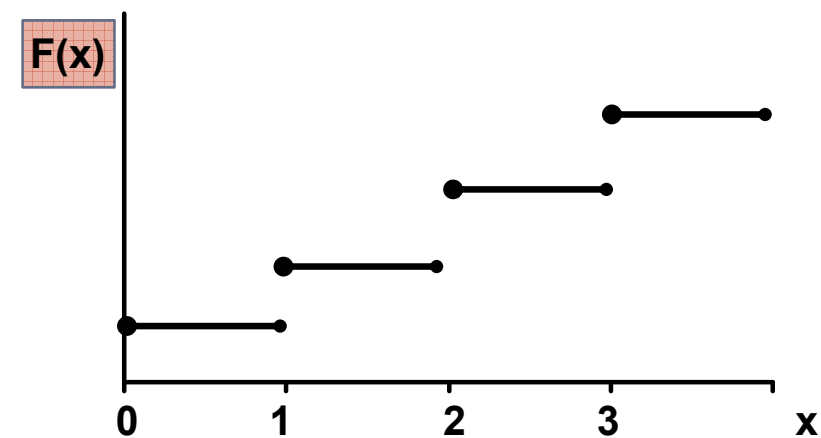
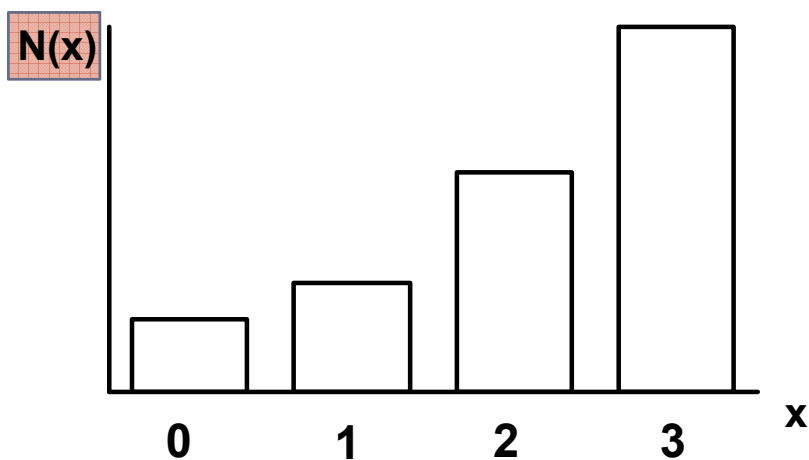
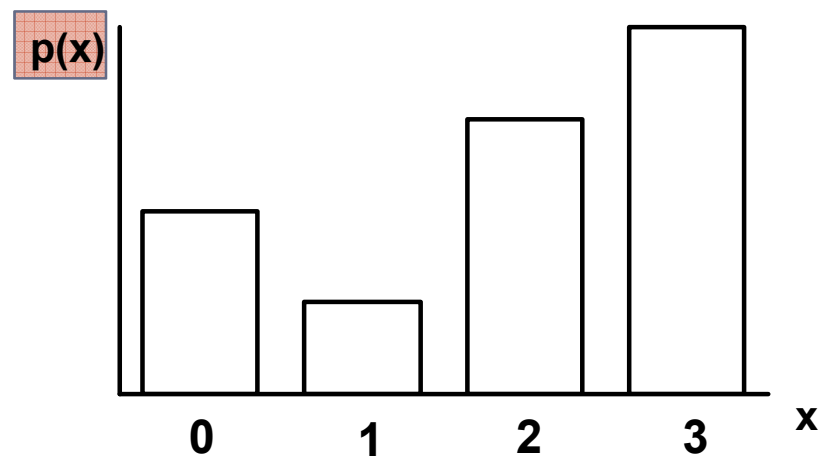
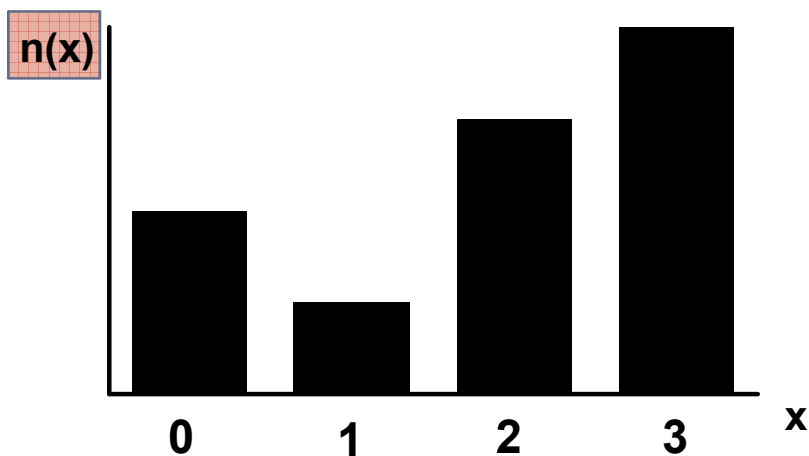
$$N(x) = \sum_{t \leq x} n(t)$$

**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

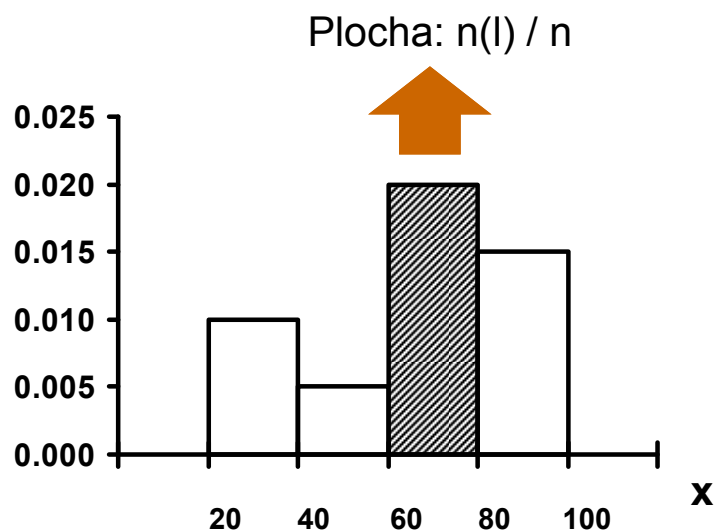
$N(x'')$  – intervalová kumulativní četnost do horní hranice  $X''$

$F(x'')$  – intervalová relativní kumulativní četnost do horní hranice  $X''$

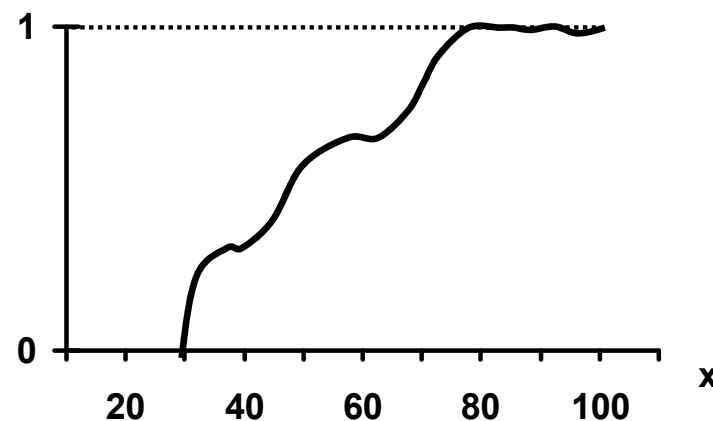
# Jak vznikají informace ?

## - frekvenční sumarizace spojitých dat

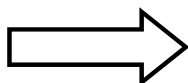
### Histogram



### Výběrová distribuční funkce

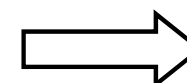


$$f(x) = \frac{n(l) / n}{d(l)}$$



Intervalová  
hustota  
četnosti

$F(x)$

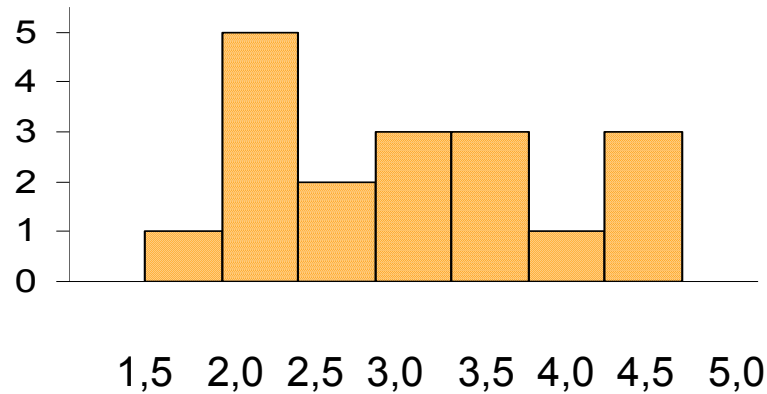


Intervalová  
relativní  
kumulativní  
četnost

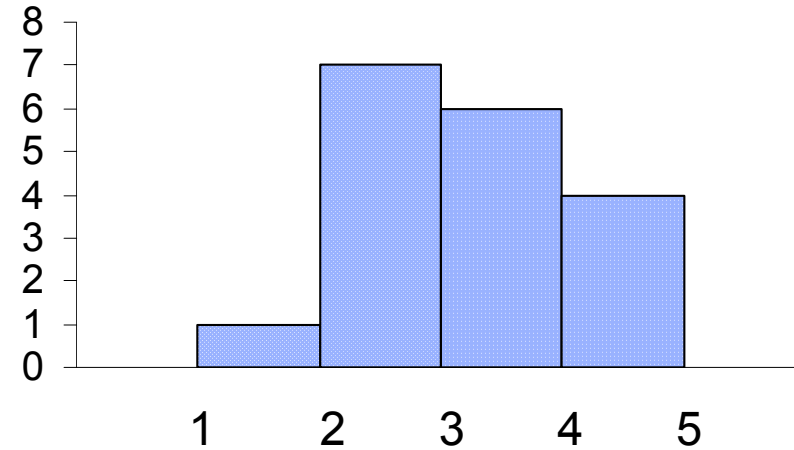
# Počet zvolených tříd a velikost souboru určují kvalitu výstupů



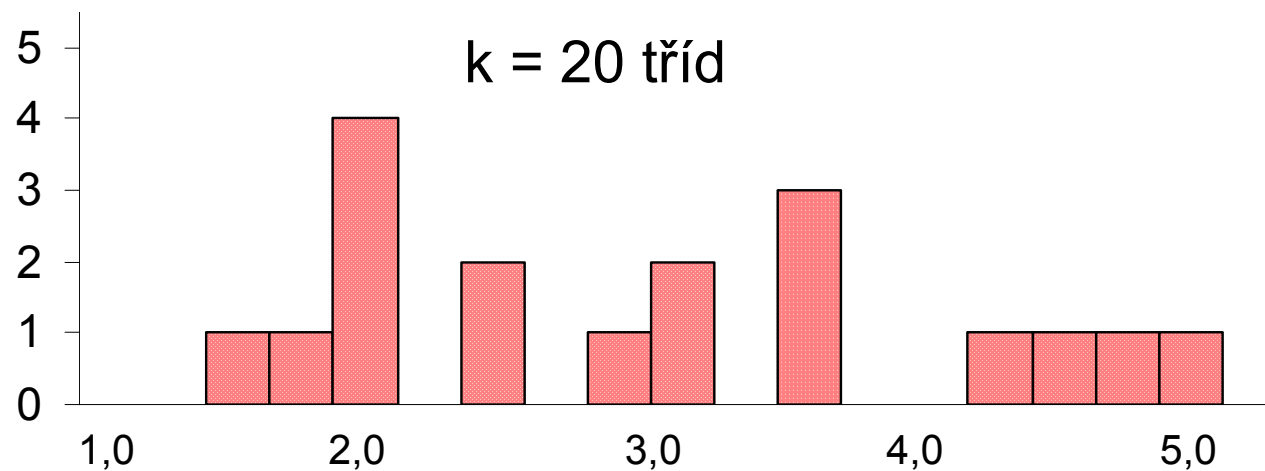
k = 10 tříd



k = 5 tříd

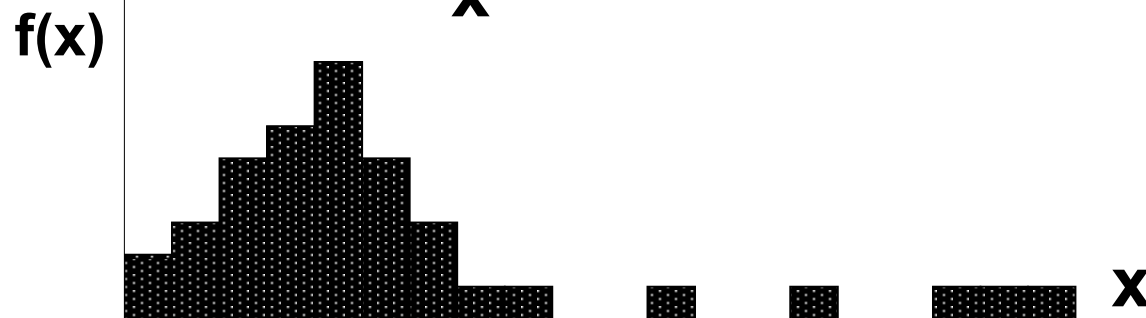
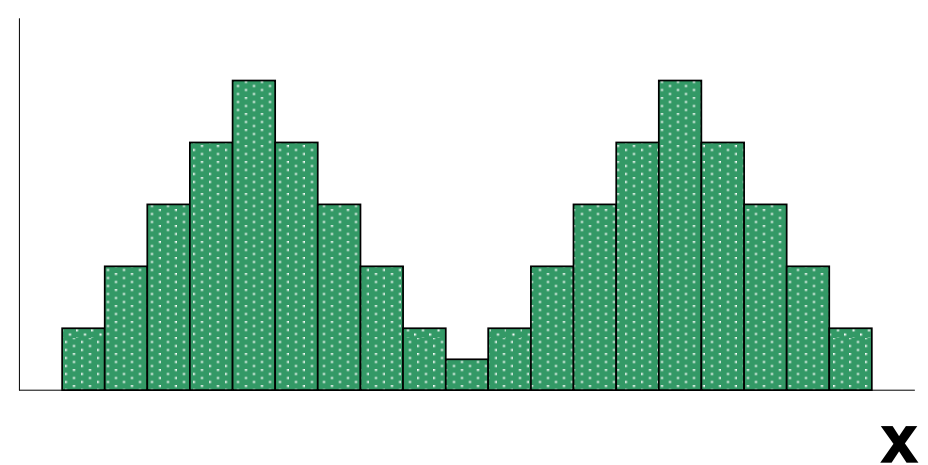
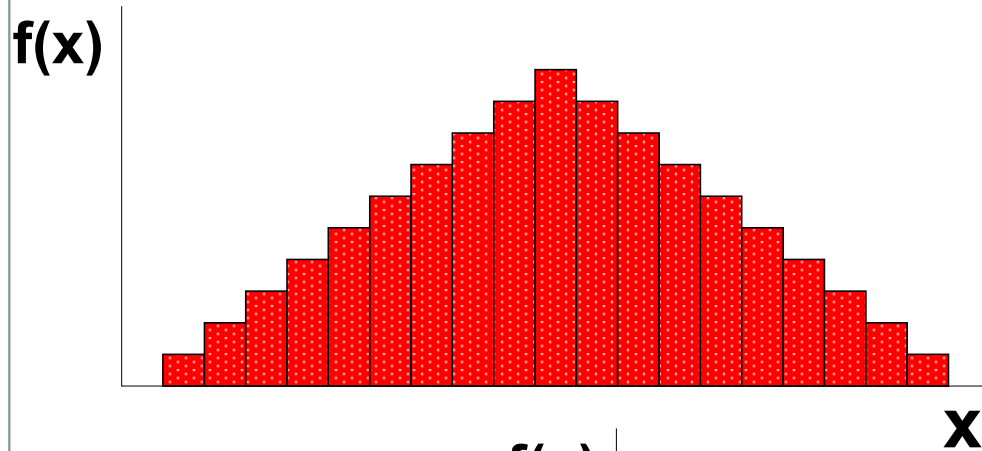
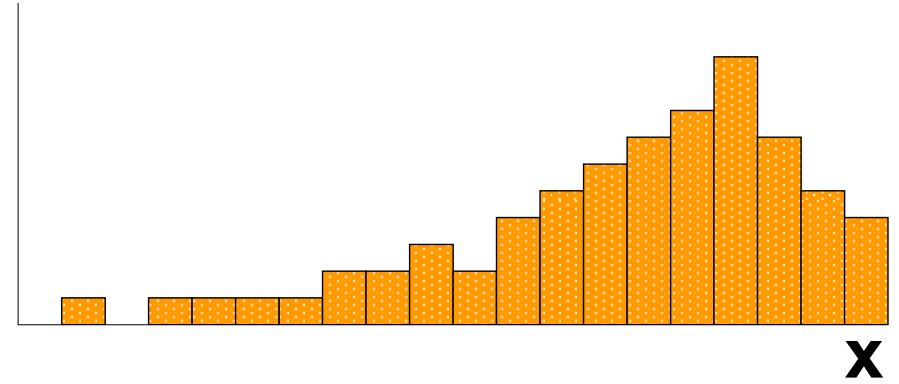
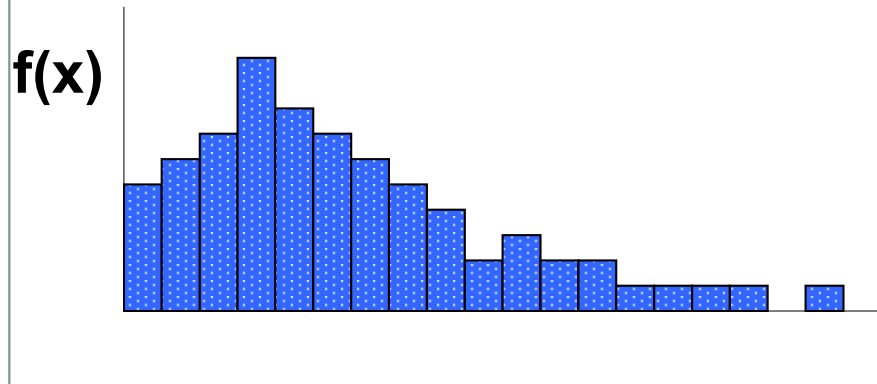


k = 20 tříd

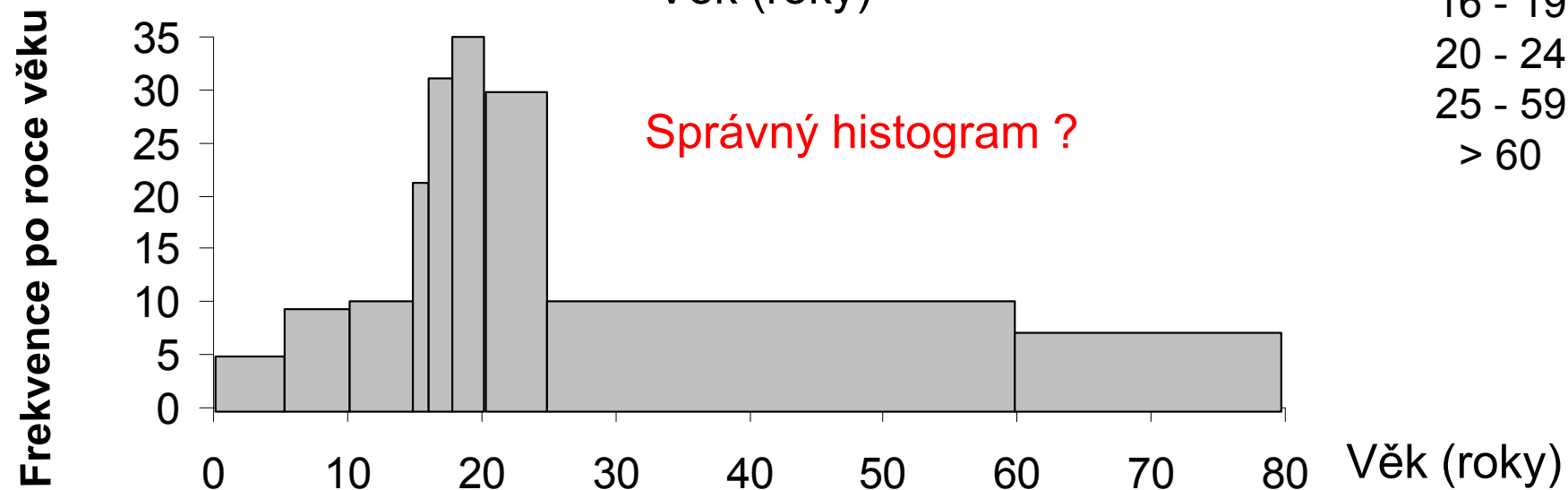
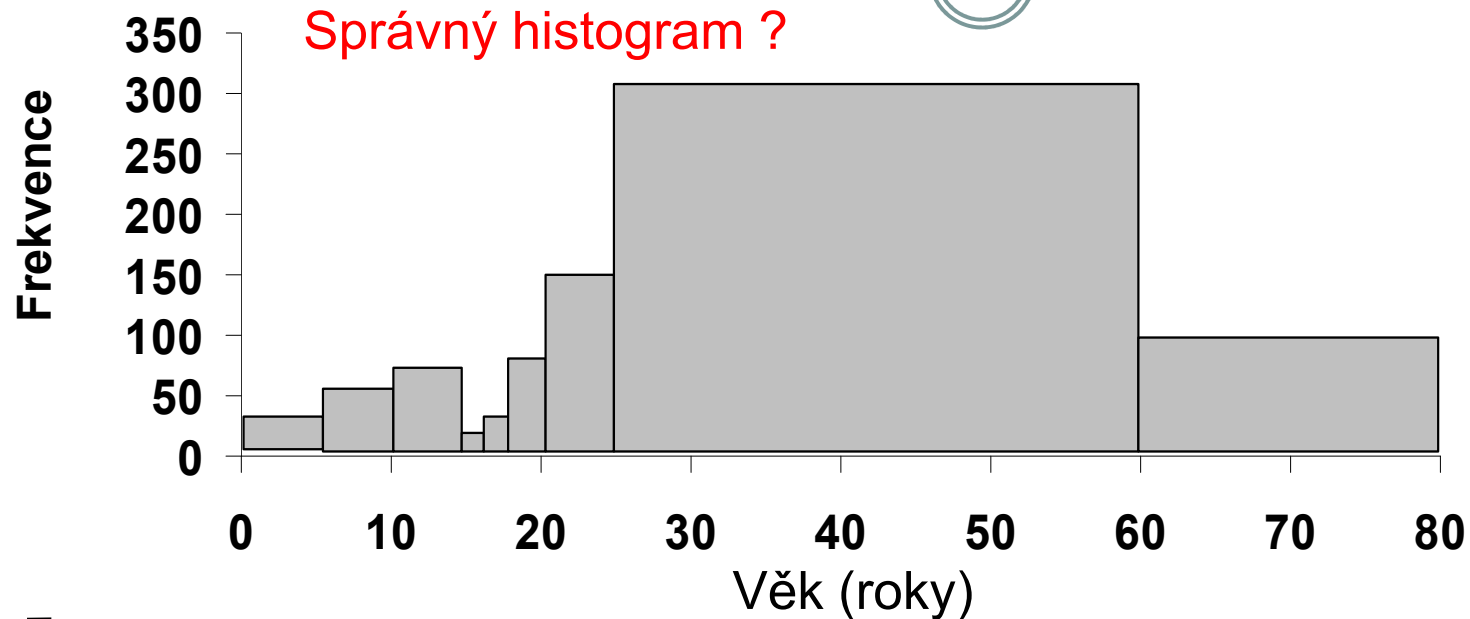




# Histogram vyjadřuje tvar výběrového rozložení

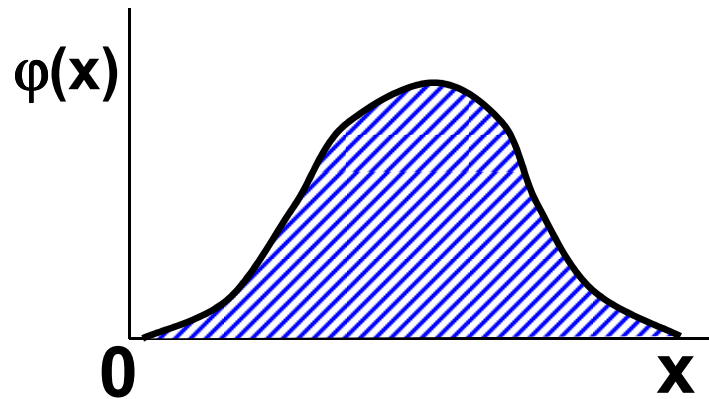


# Příklad: věk účastníků vážných dopravních nehod

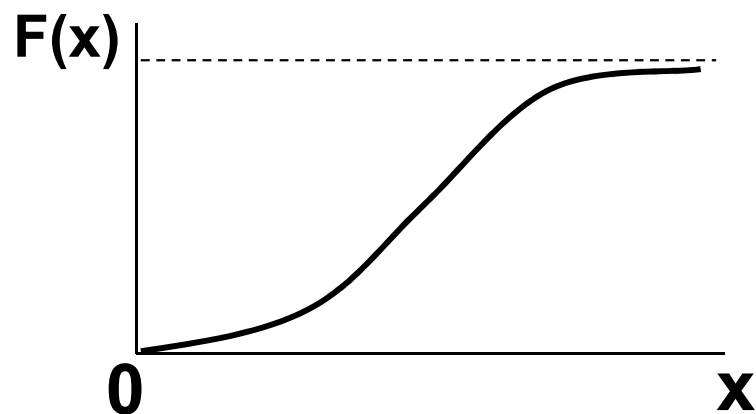


<u>Věk</u>	<u>f</u>
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103

# Pojem ROZLOŽENÍ - příklad spojitých dat



Rozložení

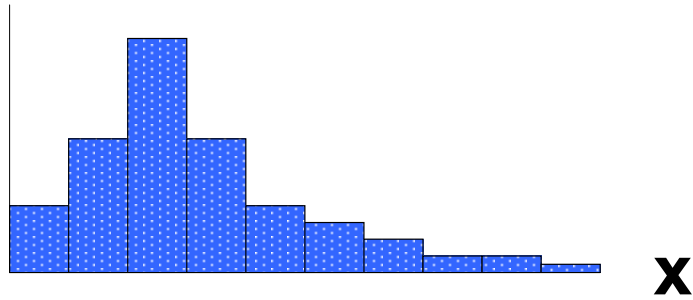


Distribuční funkce

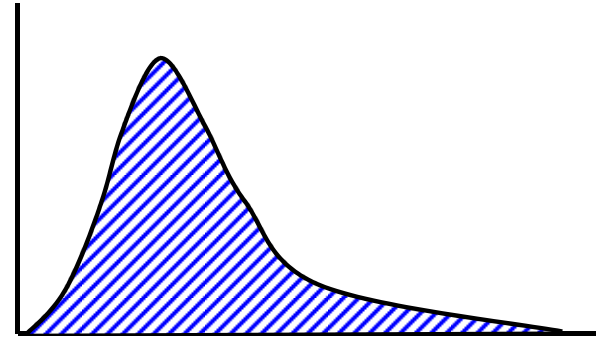
**Je - li dána  
distribuční  
funkce,  
je dáno  
rozložení**

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

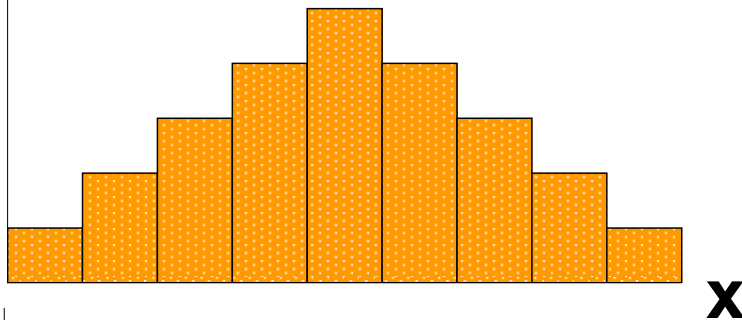
$f(x)$



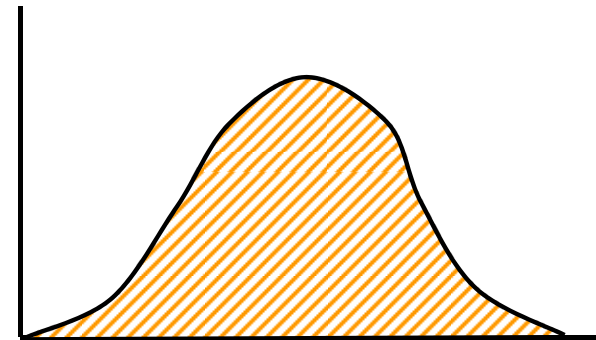
$\varphi(x)$



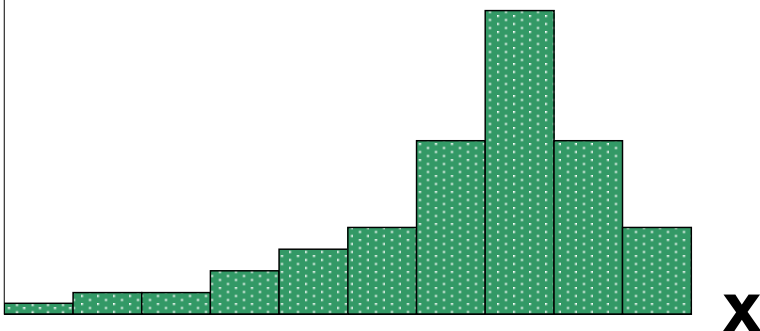
$f(x)$



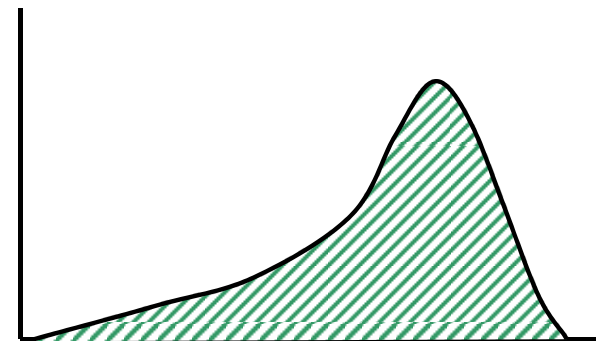
$\varphi(x)$



$f(x)$



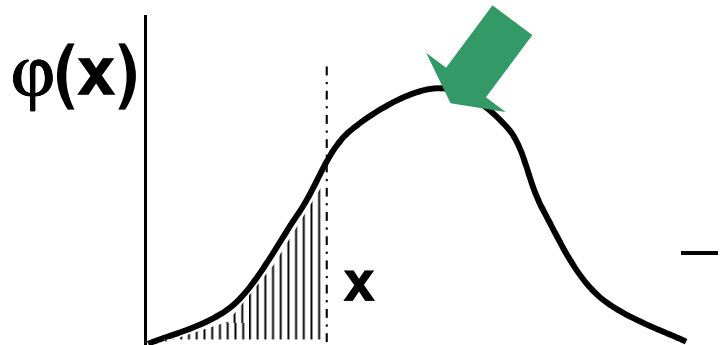
$\varphi(x)$



# Distribuční funkce jako užitečný nástroj pro práci s rozložením

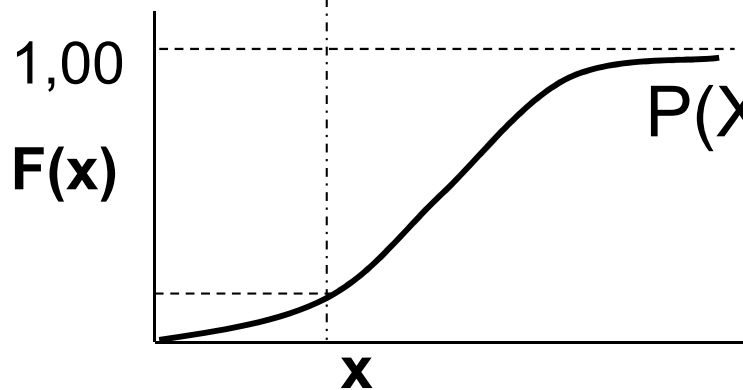
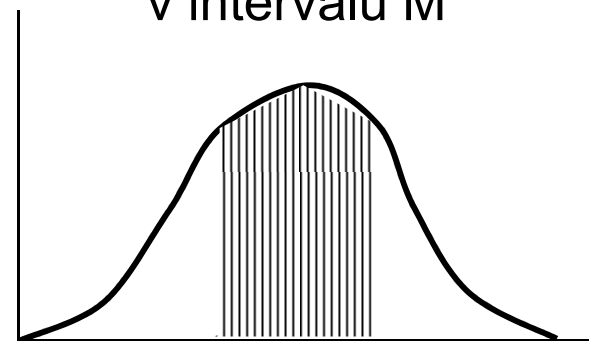


Plocha = relativní četnost



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

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



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

$\Phi(x)$  ... distribuční funkce

$$P(X \leq x) = \int_M \varphi(x) d(x)$$

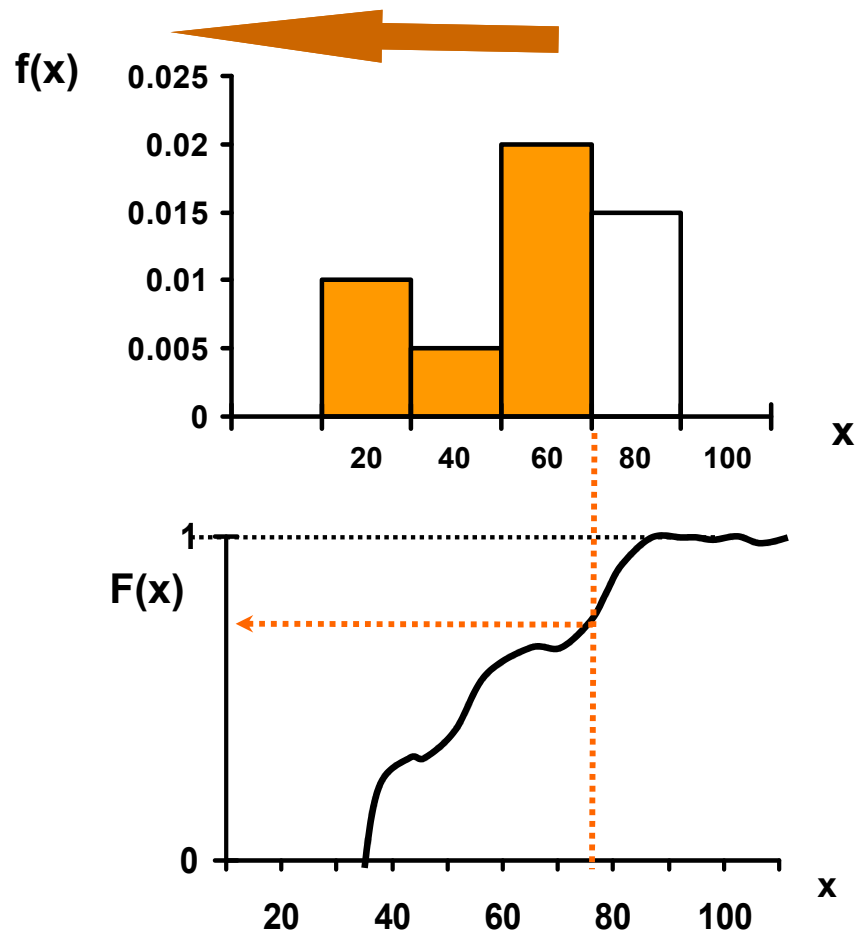
Známe-li distribuční funkci, pak známe rozložení sledované veličiny.

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**

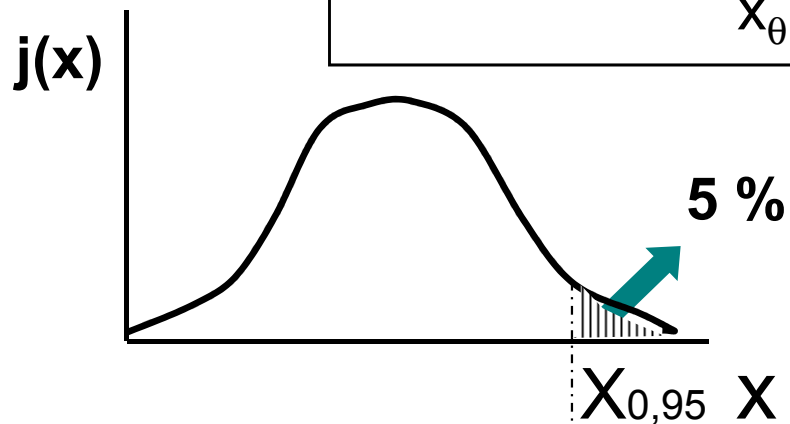
$X_{0.1}; X_{0.9}; X_{0.5}; X_{\theta}$

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



$\theta = 0,95$  ... Pravděpodobnost

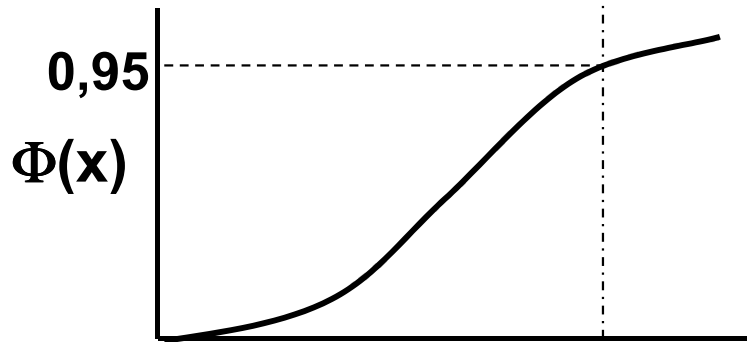
**Hledáme:**  $P(X \leq x_\theta) = 0,95 = \theta$   
 $x_\theta = (X_{0,95}) = ?$



$$F(x_\theta) = \theta$$



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



**Jakékoliv číslo na ose  $x$  je kvantilem**

# 5b. Základní statistické testy



**Jednovýběrový t-test**  
**Párový t-test**  
**Mann-Whitneův test**



# t-Test



Tři varianty parametrického t-testu:

jednovýběrový  
párový  
dvouvýběrový

Předpoklad:

**Měřená náhodná veličina má normální rozdělení.**



**Výběrový průměr má normální rozdělení se stejnou střední hodnotou, skutečný rozptyl ovšem neznáme.**

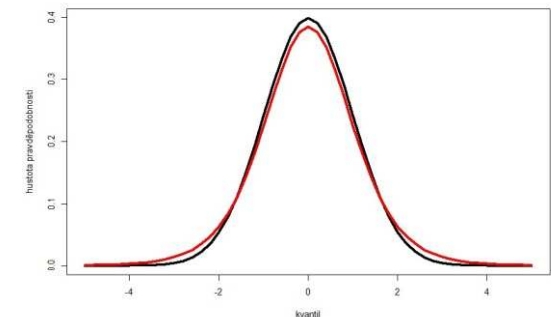


**Rozdíl výběrového průměru od skutečné střední hodnoty má také normální rozdělení.**



**Při využití výběrového rozptylu má rozdíl t-rozdělení.**

Kvantil	-3	-2	-1	0	1	2	3
Norm(0,1)	0,00	0,05	0,24	0,40	0,24	0,05	0,00
$t_7(0,1)$	0,01	0,06	0,23	0,38	0,23	0,06	0,01



# t-Test



## Princip:

podle určené hladiny pravděpodobnosti se stanoví maximální přípustná velikost rozdílu výběrového průměru a skutečné střední hodnoty. Testuje se velikost rozdílu.

## Postup:

Výpočet normalizovaného rozdílu a jeho porovnání s tabelovanou hodnotou (jednostranná a dvoustranná varianta):

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

$H_0$	$H_A$	Testová statistika	Interval spolehlivosti
$\bar{x} \leq \mu$	$\bar{x} > \mu$	$t$	$t > t$
$\bar{x} \geq \mu$	$\bar{x} < \mu$	$t$	$t < t$
$\bar{x} = \mu$	$\bar{x} \neq \mu$	$t$	$ t  > t$

# t-Test



## *Koncentrace antibiotika v cílovém orgánu*

**Při 1000 měřeních antibiotika byla zjištěna v cílovém orgánu průměrná koncentrace 202,5 jednotek a směrodatná odchylka 44 jednotek.**

**Požadovaná koncentrace antibiotika je 200 jednotek.**

- 1) Je daný rozdíl 2,5 významný vzhledem k variabilitě znaku na hladině významnosti 5 %?**
- 2) Jaká je skutečná hladina významnosti?**

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} = \frac{2,5}{44} \sqrt{1000} = 1,797$$

# Mann-Whitneyův U test



**Neparametrická** varianta t-testu se skoro stejnou silou v případě normálně rozdělených dat. Vždy pro dvě skupiny naměřených hodnot.

**Předpoklad:** **Pravděpodobnost že  $X > Y$  = pravděpodobnosti, že  $Y > X$ .**



**Vypočtená U statistika má přibližně normální rozdělení (pro malé počty jsou hodnoty tabelovány zvlášť).**

**Postup:**

Hodnoty z obou sad měření se seřadí podle velikosti.

Počítá se U statistika pro první nebo druhou sadu (obvykle pro tu s nižšími hodnotami)

$U_1$  je součet počtů hodnot ze sady 2 nižších než jednotlivé prvky sady 1 (postupně se sčítá pro všechny prvky ze sady 1).

Alternativní výpočet:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

$R_1$  je součet pořadí skupiny 1.

# Mann-Whitneyův U test



Provede se normalizace:

$$z = \frac{U - m_U}{\sigma_U} \quad m_U = \frac{n_1 n_2}{2} \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

$z$  je normalizovaná statistika

$m_U$  je průměr statistiky  $U$

$\sigma_U$  je směrodatná odchylka statistiky  $U$

Vypočtená statistika  $z$  se porovná s tabelovanými hodnotami normálního rozdělení resp. pro nižší počty s tabelovanými hodnotami pro Mann-Whitneyův  $U$  test.