2. Základní typy dat

Spojitá a kategoriální data
Základní popisné statistiky
Frekvenční tabulky
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

Binární data

Diskrétní data

Data nominální

Rovná se ?

Otázky „Ano/Ne“

Podíl hodnot větší/menší než specifikovaná hodnota ?

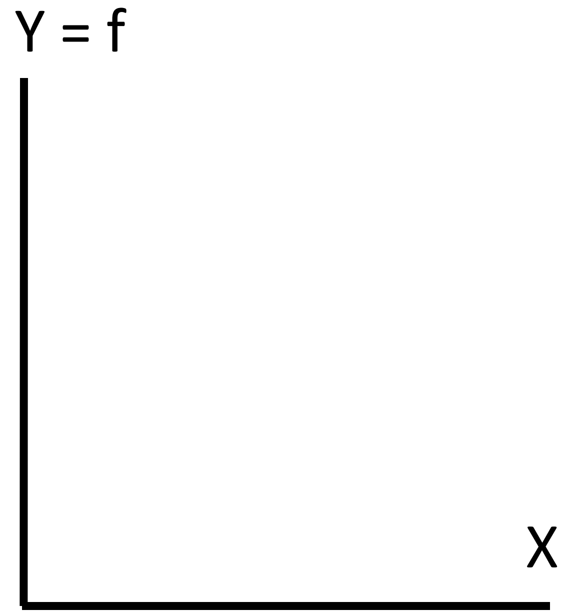
Procenta odvozené hodnoty

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

Jak vznikají informace ?
– různé typy dat znamenají různou informaci

Statistika středu

Data poměrová	PRŮMĚR	Spojitá data
Data intervalová		
Data ordinální	MEDIÁN	
		Diskrétní data
Data nominální	MODUS	



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

Jak vznikají informace ? – různé typy dat znamenají různou informaci

Data:

$$\{x_i\}_{i=1}^{n}$$

Průměr:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Rozptyl (výběrový):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Směrodatná odchylka (výběrová):

$$s = \sqrt{s^2}$$

p-tý kvantil

$$q_p = x_j : \left|\{x_k : x_k \le x_j\}\right| = p \cdot n$$

Medián:

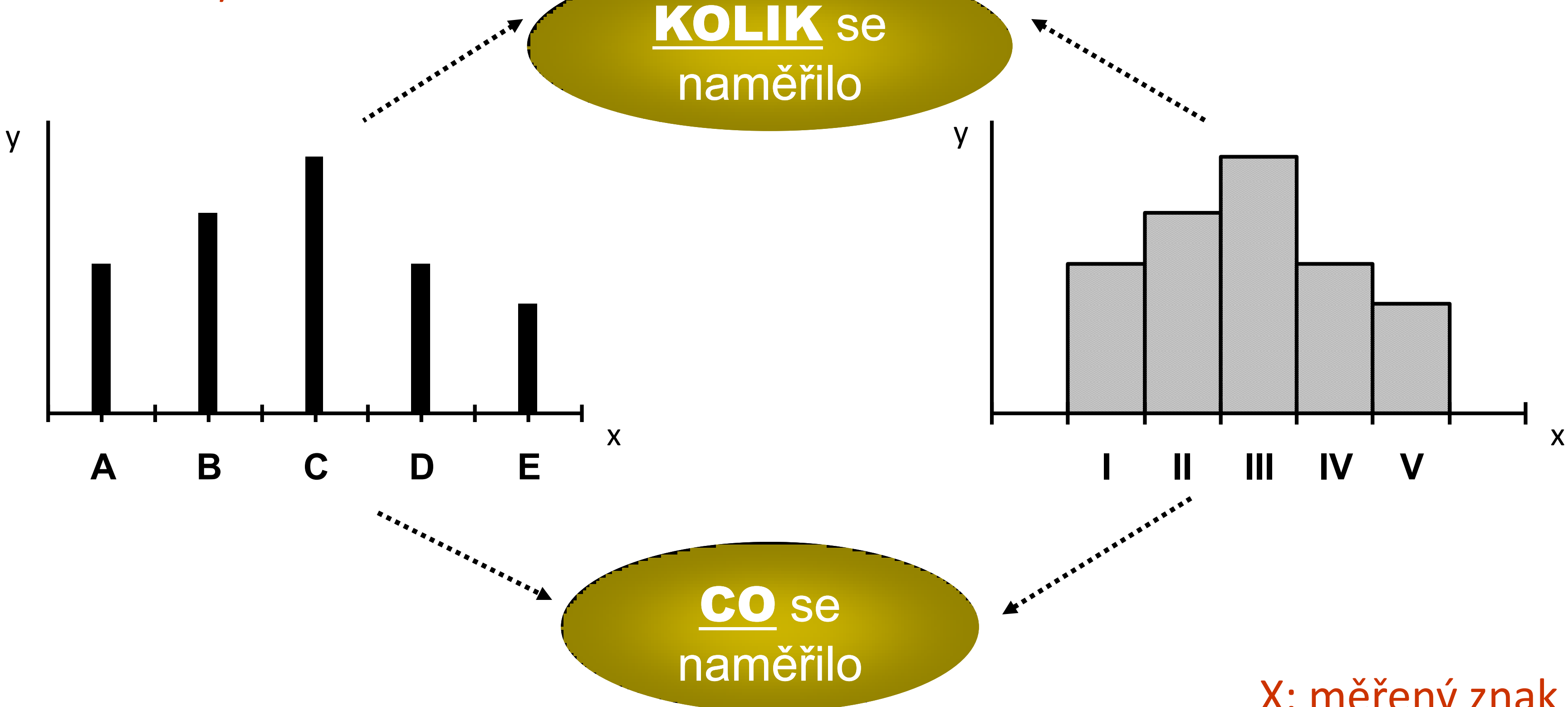
$$\tilde{x} = x_j : \left|\{x_k : x_k \le x_j\}\right| = \frac{n}{2}$$

Modus:

$$\hat{x} = max_j \left|\{x_k : x_k = x_j\}\right|$$

JAK vznikají informace ?
- opakovaná měření informují rozložením hodnot

Y: frekvence -
absolutní / relativní



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

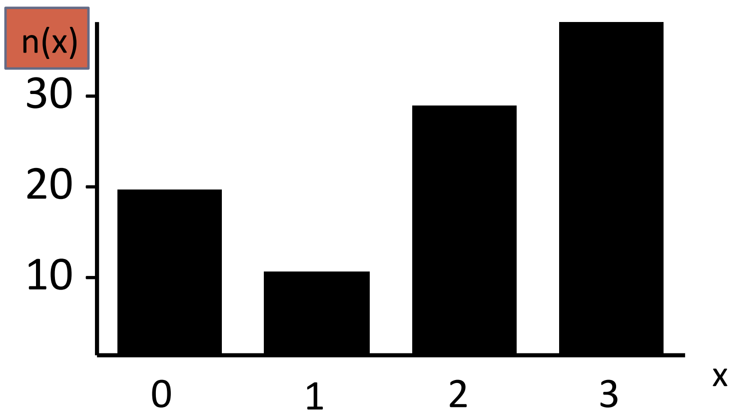
x	n(x)	N(x)	p(x)	F(x)
0	20	20	0,2	0,2
1	10	30	0,1	0,3
2	30	60	0,3	0,6
3	40	100	0,4	1,0

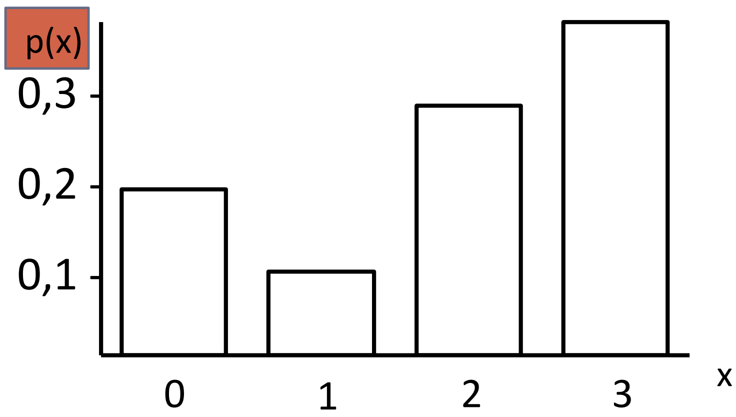
n(x) – absolutní četnost x
N(x) – kumulativní četnost hodnot nepřevyšujících x;

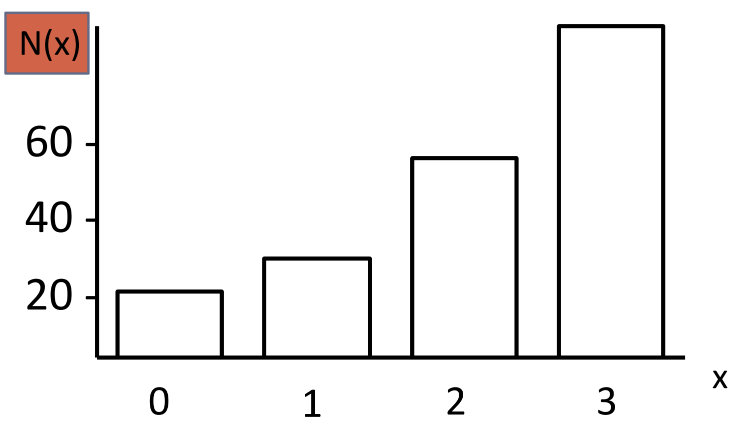
$$N(x) = \sum_{t \leq x} n(t)$$

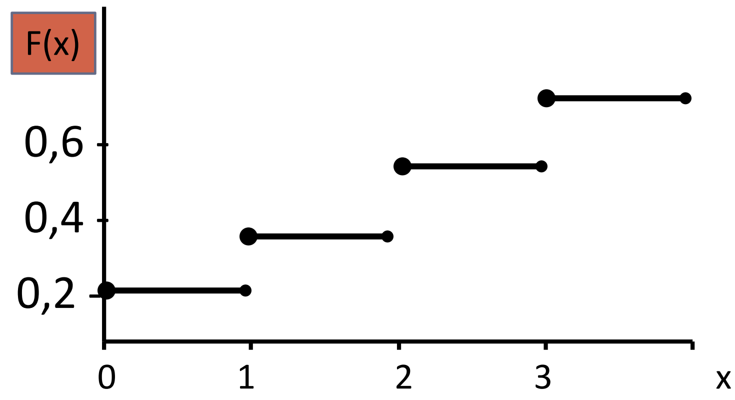
p(x) – relativní četnost; p(x) = n(x) / n
F(x) – kumulativní relativní četnost hodnot nepřevyšujících x; F(x) = N(x) / n

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)

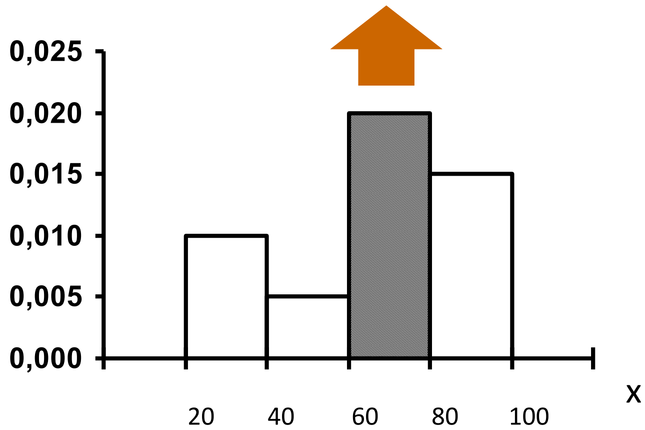
interv	d(I)	n(I)	n(I)/n	N(x'')	F(x'')
<20, 40)	20	20	0,2	20	0,2
<40, 60)	20	10	0,1	30	0,3
<60, 80)	20	40	0,4	70	0,7
<80, 100)	20	30	0,3	100	1,0

d(I) – šířka intervalu
n(I) – absolutní četnost
n(I) / 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

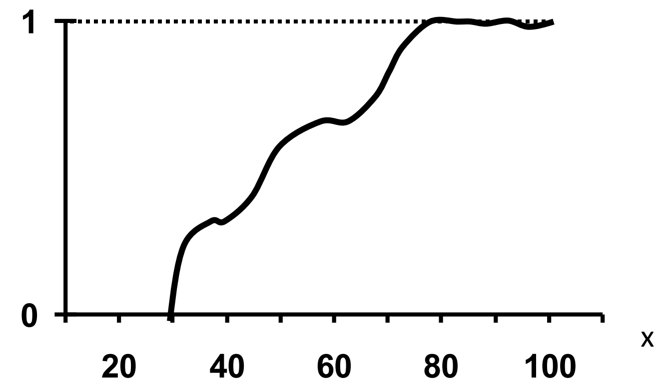
Histogram

Plocha: n(I) / n



$$f(x) = \frac{n(I)\,/\,n}{d(I)}$$ ⇨ Intervalová hustota četnosti

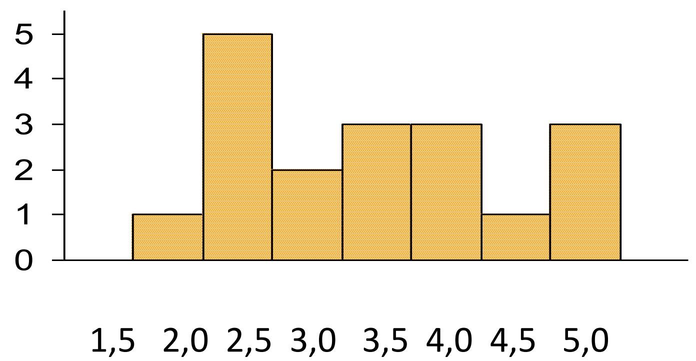
Výběrová distribuční funkce



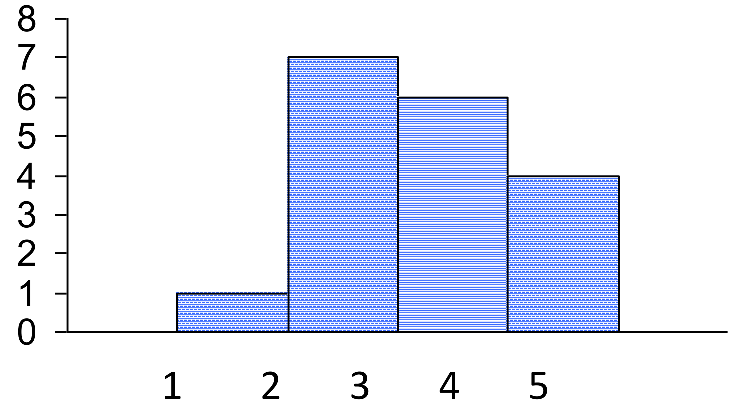
$F(x)$ ⇨ Intervalová relativní kumulativní četnost

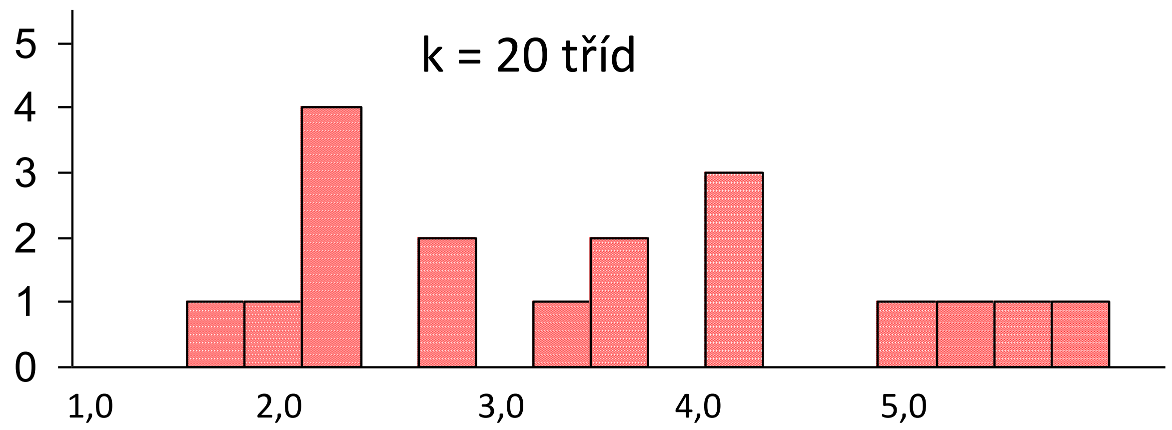
Počet zvolených tříd a velikost souboru určují kvalitu výstupu

k = 10 tříd

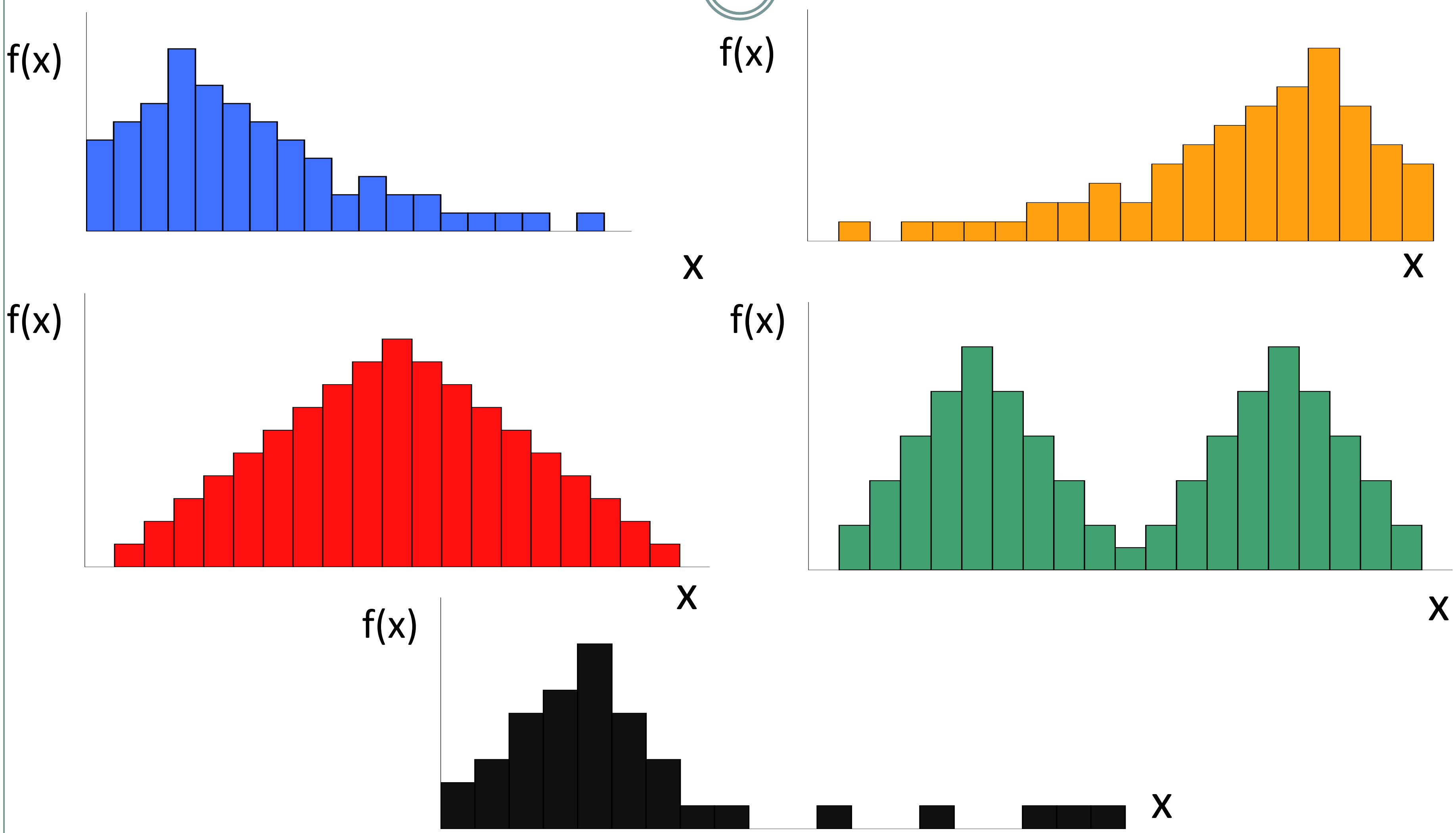


k = 5 tříd





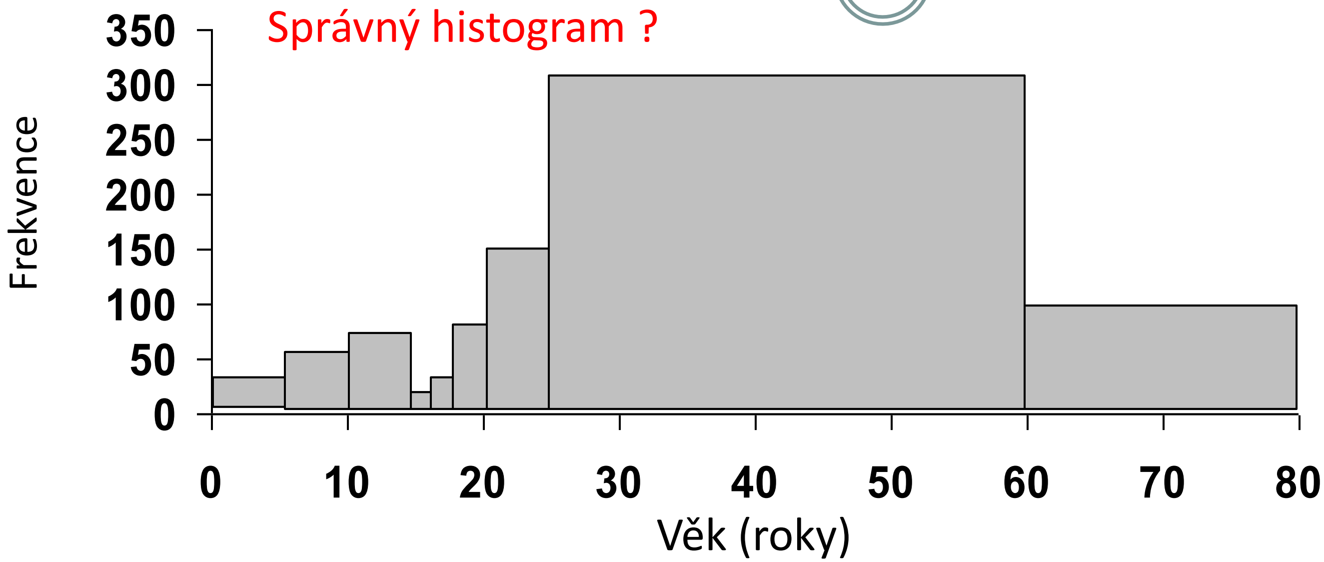
Histogram vyjadřuje tvar výběrového rozložení



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

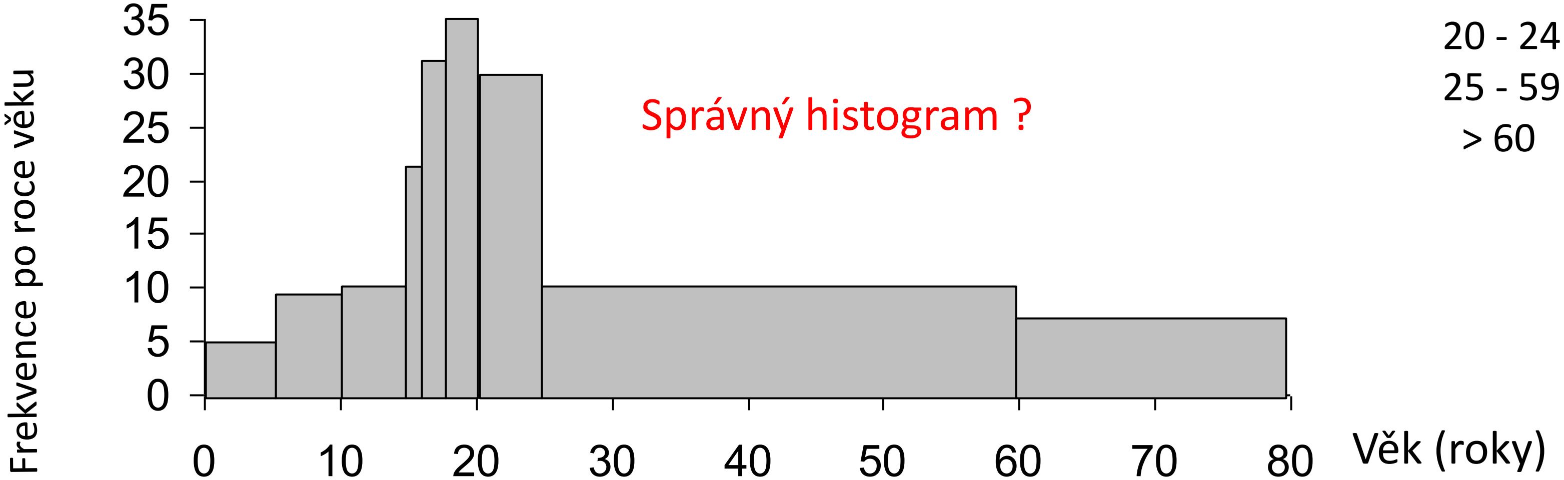
Příklad: věk účastníků vážných dopravních nehod

Správný histogram ?

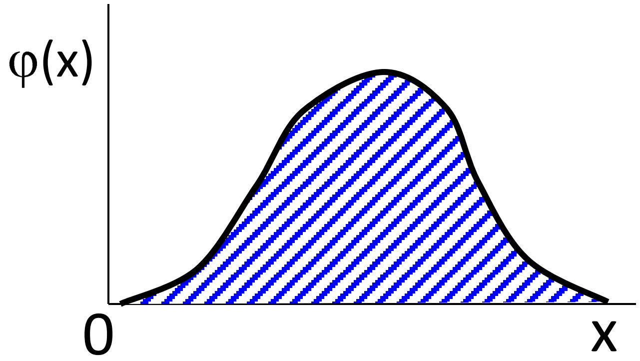


Věk	f
0 - 4	28
5 - 9	46
10 - 15	58
16 - 19	20
20 - 24	114
25 - 59	316
> 60	103

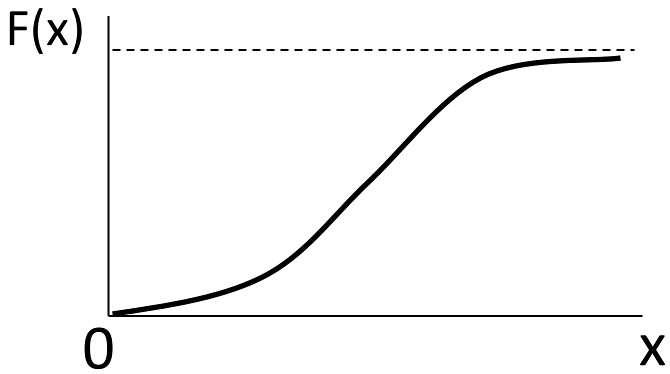
Správný histogram ?



Pojem ROZLOŽENÍ - příklad spojitých dat



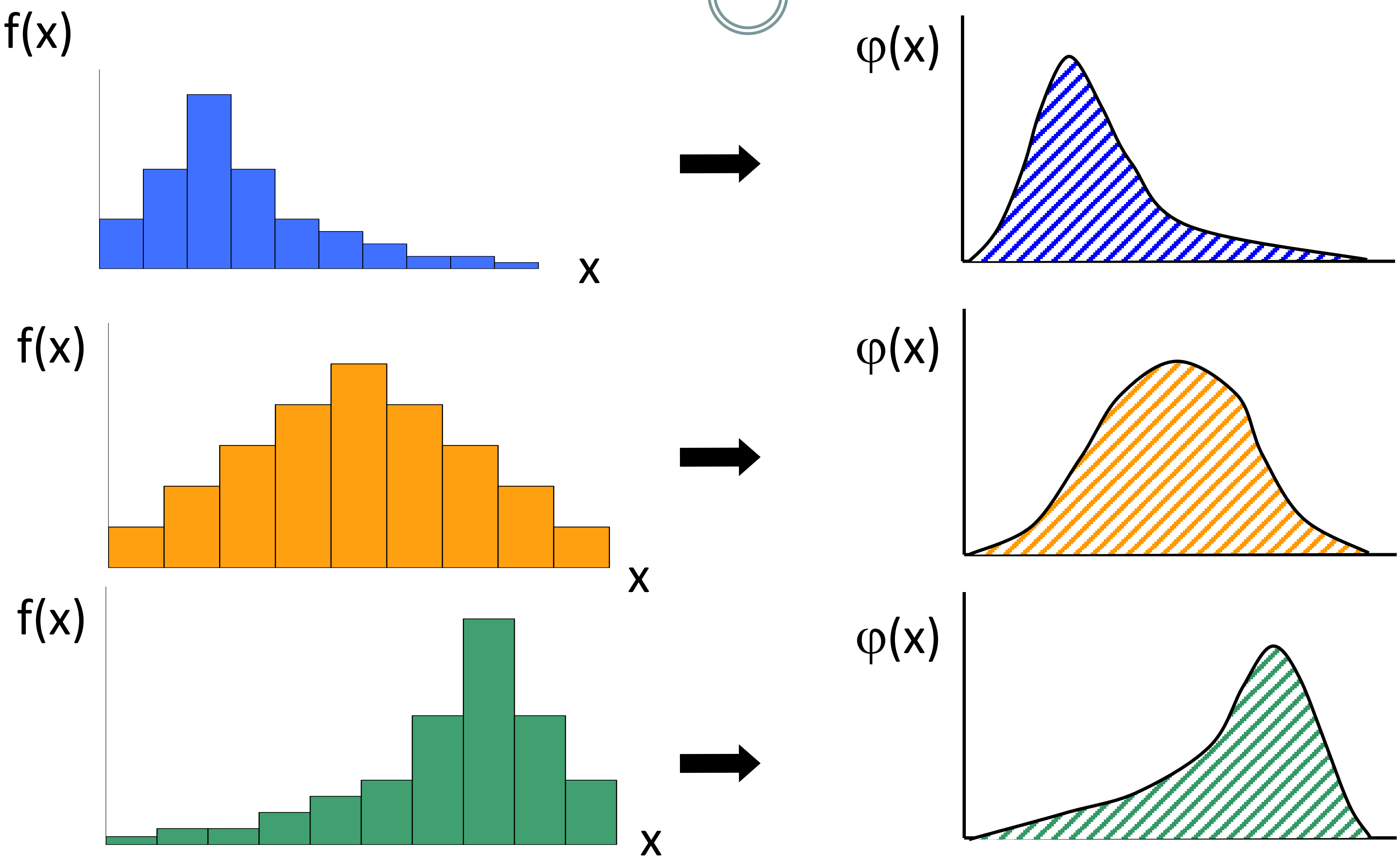
Rozložení



Distribuční
funkce

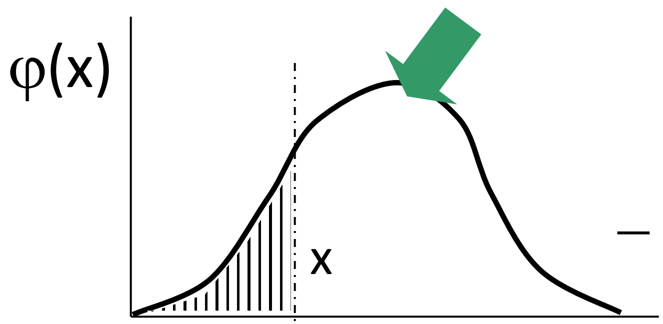
Je - li dána
distribuční
funkce,
je dáno rozložení

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



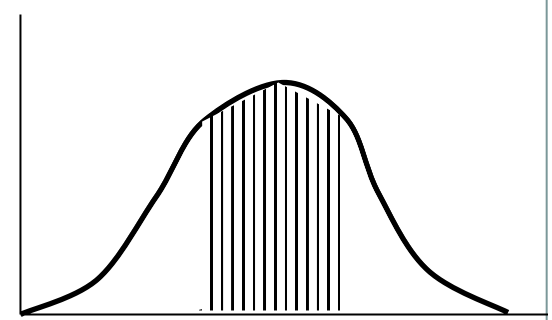
Distribuční funkce jako užitečný nástroj pro práci s rozložením

Plocha = relativní četnost

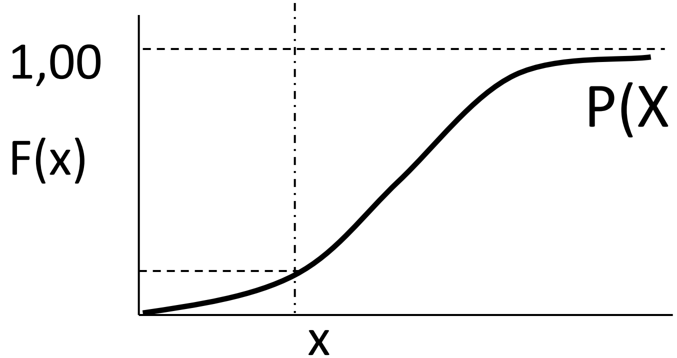


$$\int_{-\infty}^{\infty} \varphi(x)\ d(x) = 1$$

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



M



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

$\Phi(x)$ … distribuční funkce

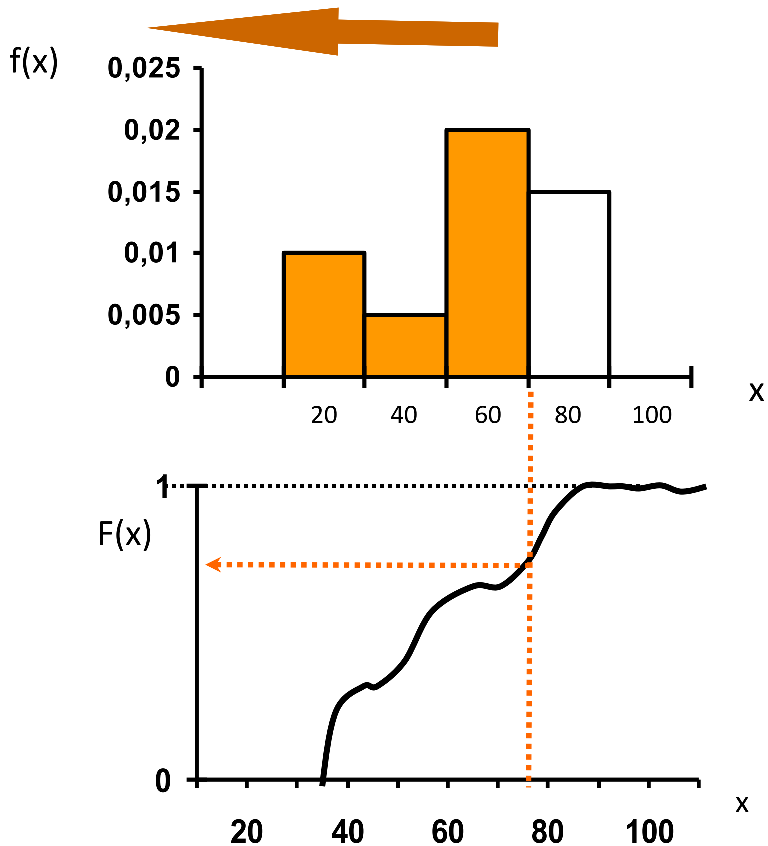
$P(X \leq \ = \int_M \varphi(x)\ d(x)$

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 <u>pravděpodobnostní zařazení</u> každé jednotlivé hodnoty

KVANTIL

$X_{0.1}$; $X_{0.9}$; $X_{0.5}$; X_{θ}

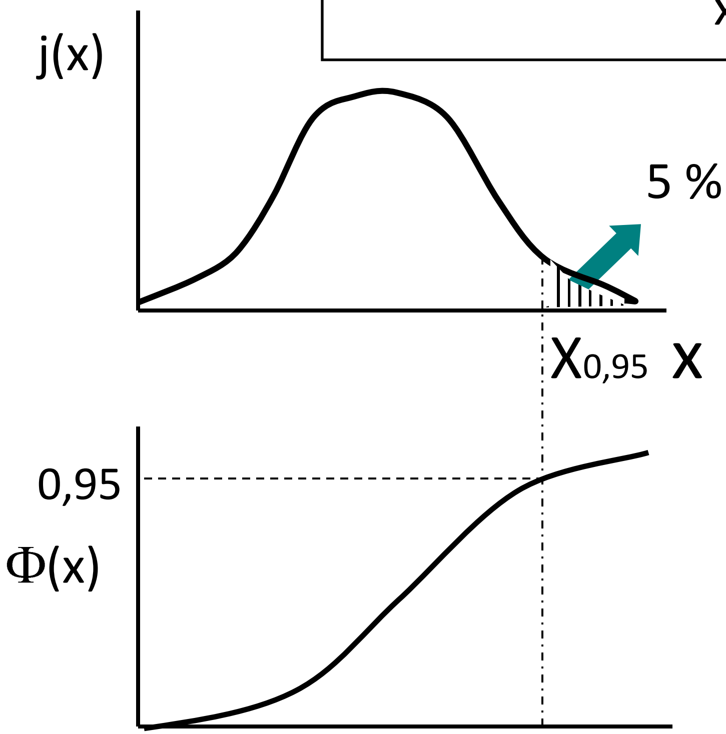
Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, J. Kalina

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

θ = 0,95 ... Pravděpodobnost

<u>Hledáme:</u> $P(X \leq \quad) = 0{,}95 = \theta$

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



F (x =

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

Jakékoliv číslo na ose x je kvantilem