

2. Základní typy dat



Spojitá a kategoriální data
Základní popisné statistiky
Frekvenční tabulky
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.

Typy proměnných (dat)

Binární = dummy data

Proměnná, která může nabývat pouze dvou hodnot. Bývá definovaná odpovědí na otázku (např. TRUE × FALSE, 1 × 0).

Nominální = kategoriální data

Proměnná, která může nabývat počtu hodnot ($n \in \mathbb{N}$), pro které neexistuje přirozené pořadí (např. barvy vzorků).

Ordinální data

Nominální proměnná, pro kterou ale existuje jasné pořadí kategorií (např. velikost oděvů S, M, L, XL).

Kardinální data

Kardinální proměnné odpovídají počtům něčeho. Hodnoty jsou od sebe stejně vzdálené (např. počet dětí v rodině).

Intervalová data

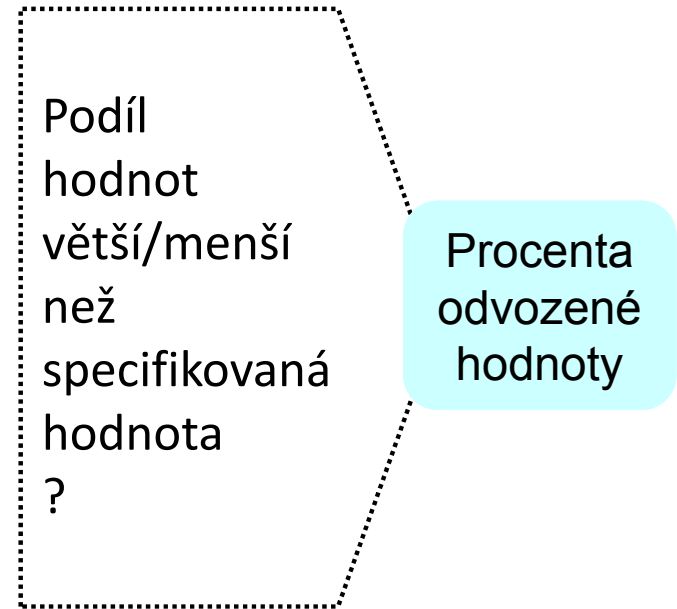
Obvykle spojitá proměnná, u které lze určit rozdíl mezi kategoriemi – často jde o vzdálenost od 0 (např. teplota ve °C, čas).

Poměrová data

Intervalová proměnná, u které má smysl určovat podíly jednotlivých kategorií (např. hmotnost, vzdálenost).

Jak vznikají informace ?

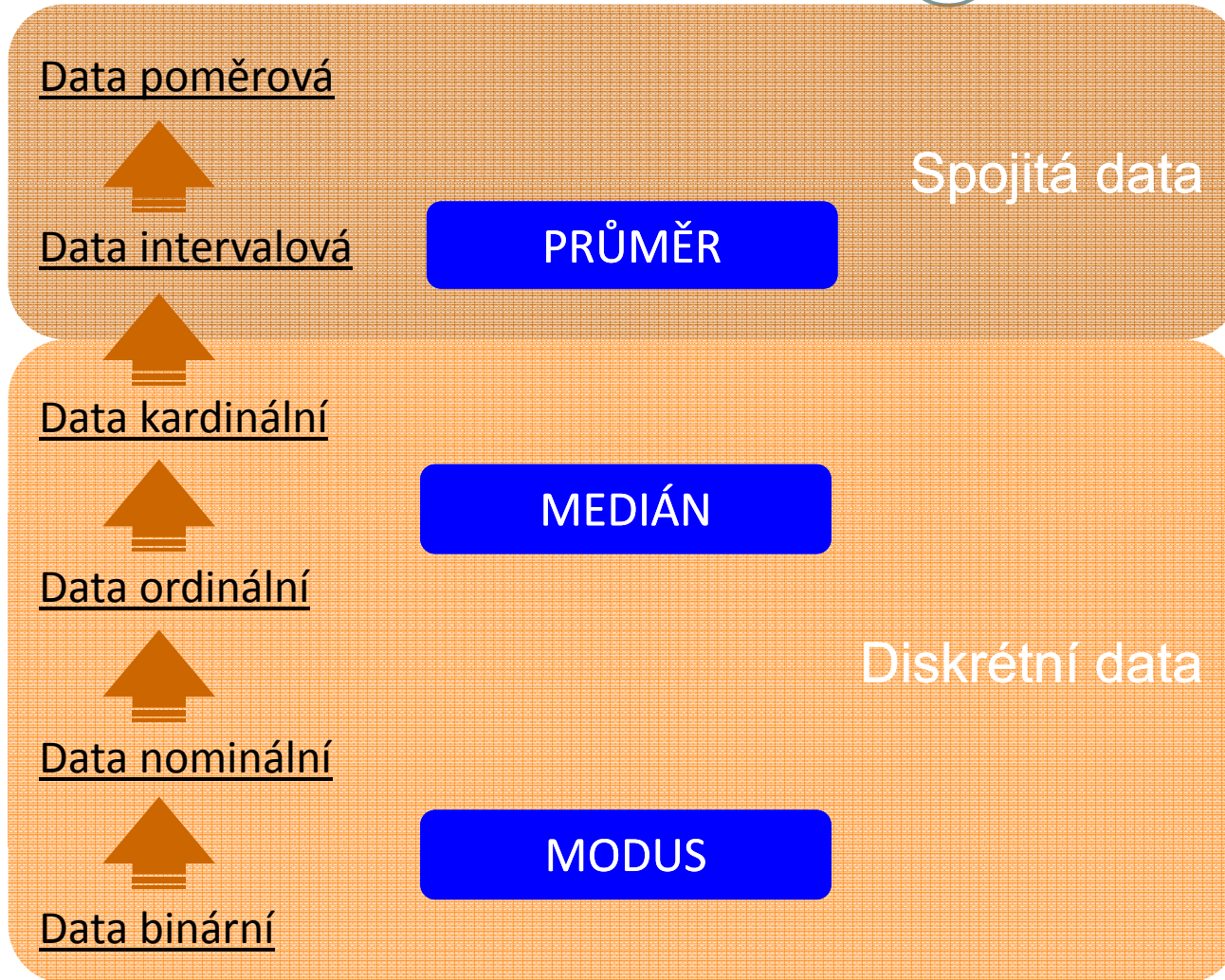
– různé typy dat znamenají různou informaci



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

Jak vznikají informace ?

– různé typy dat znamenají různou informaci



$Y = f$

X

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

Základní soubor × výběr (vzorek)



- Parametry základního souboru jsou obvykle dané, ale neznáme je (např. průměr, směrodatná odchylka).
- Pro odhad parametrů základního souboru používáme tzv. výběrové charakteristiky založené na našem omezeném výběru (vzorku).

Rozptyl (základní):

Rozptyl (výběrový):

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

Jak vznikají informace ?

– základní popisné statistiky



Data:

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

p-tý kvantil

$$q_p = x_j: |\{x_k: x_k \leq x_j\}| = p \cdot n$$

Průměr (výběrový):

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

Medián:

$$\tilde{x} = x_j: |\{x_k: x_k \leq x_j\}| = \frac{n}{2}$$

Rozptyl (výběrový):

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

Modus:

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

Směrodatná odchylka (výběrová):

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

JAK vznikají informace ?

- opakovaná měření informují rozložením hodnot

Y: frekvence
-
absolutní / relativní

KOLIK se
naměřilo

y

y

x

x

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
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



$n(x)$

30 -

20 -

10 -

0

1

2

3

x

$p(x)$

0,3 -

0,2 -

0,1 -

0

1

2

3

x

$N(x)$

60 -

40 -

20 -

0

1

2

3

x

$F(x)$

0,6 -

0,4 -

0,2 -

0

1

2

3

x

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

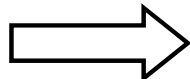
Histogram

Plocha: $n(l) / n$



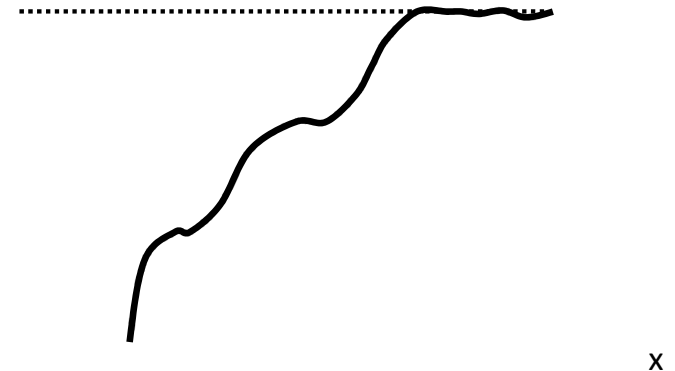
20 40 60 80 100 x

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

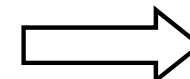


Intervalová
hustota
četnosti

Výběrová distribuční funkce



$F(x)$



Intervalová
relativní
kumulativní
četnost

Počet zvolených tříd a velikost souboru určují kvalitu výstupu



k = 10 tříd

k = 5 tříd

1,5 2,0 2,5 3,0 3,5 4,0 4,5 5,0

1 2 3 4 5

k = 20 tříd

1,0

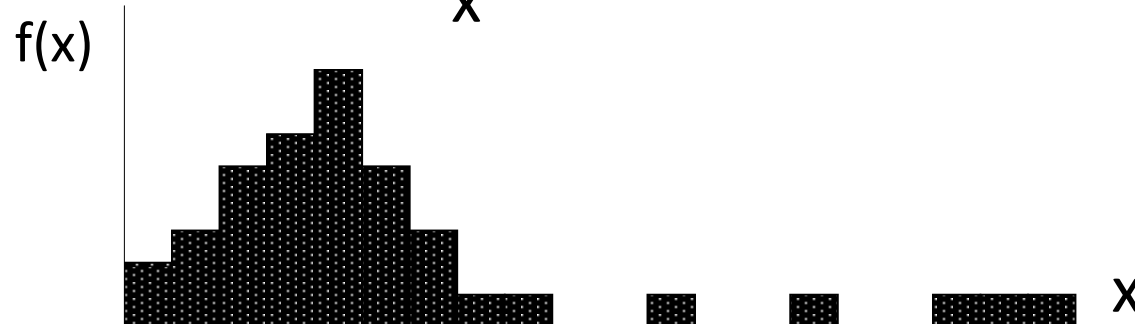
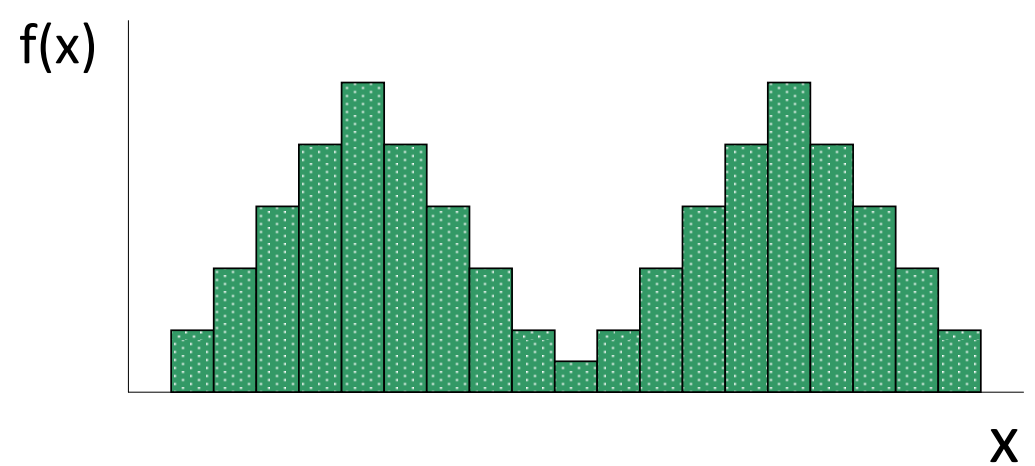
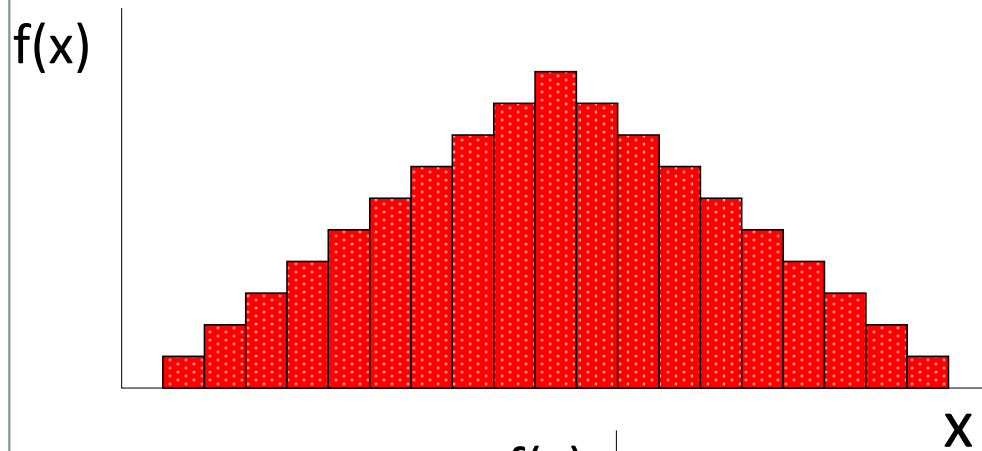
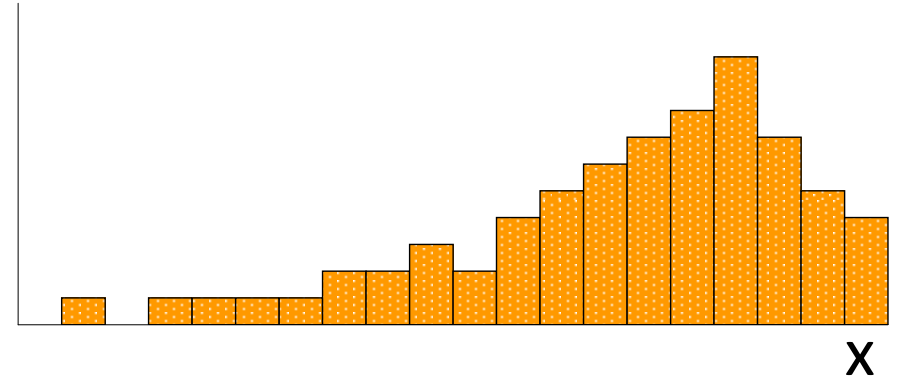
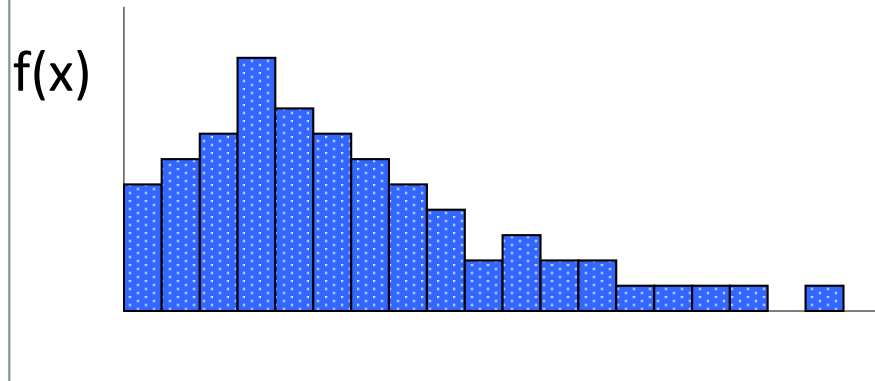
2,0

3,0

4,0

5,0

Histogram vyjadřuje tvar výběrového rozložení



Příklad: věk účastníků vážných dopravních nehod



Správný histogram ?

Kategorie na ose x nemusí být ekvidistantní.

Frekvence

Věk (roky)

Správný histogram ?

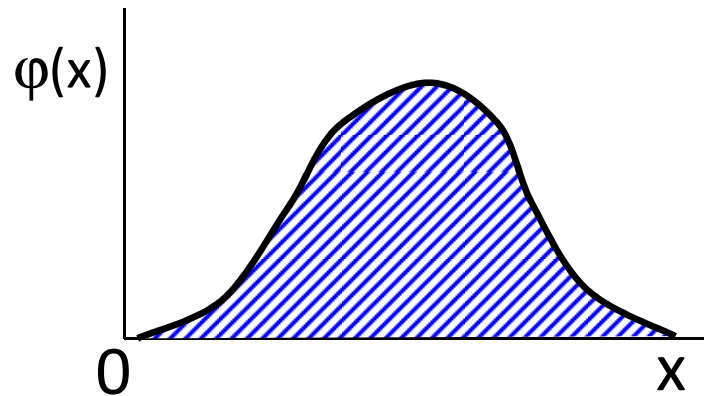
Plocha histogramu odpovídá počtu případů (pokud jde o pravděpodobnost, je plocha 1).

<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

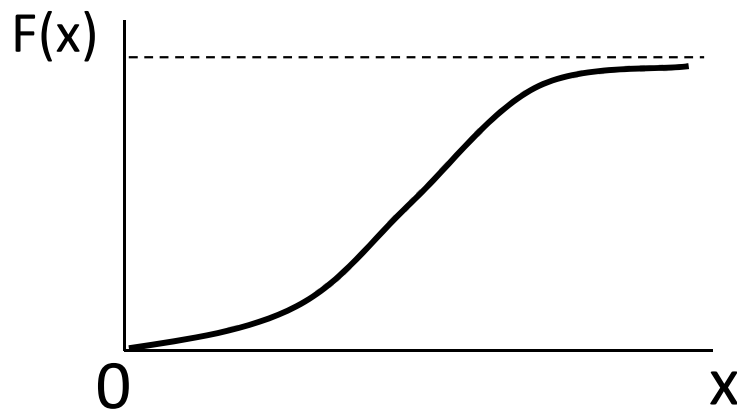
Frekvence po roce věku

Věk (roky)

Pojem ROZLOŽENÍ - příklad spojitých dat



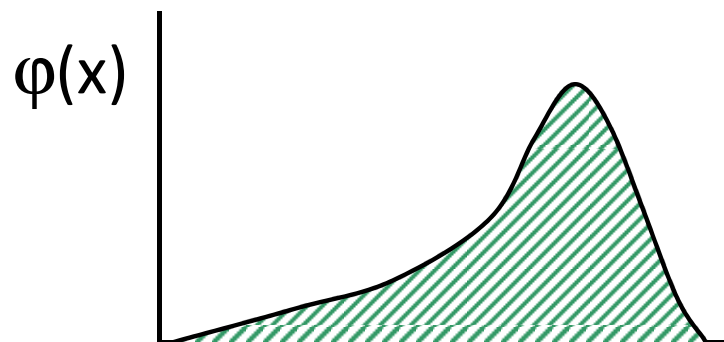
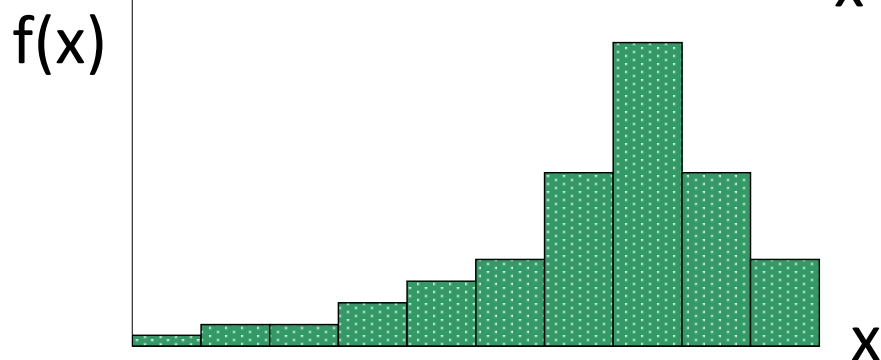
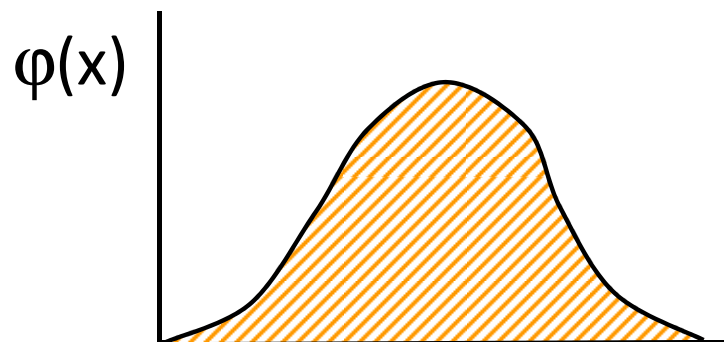
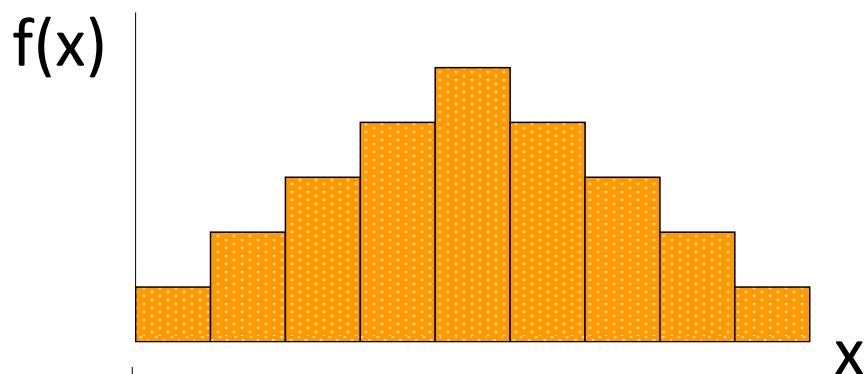
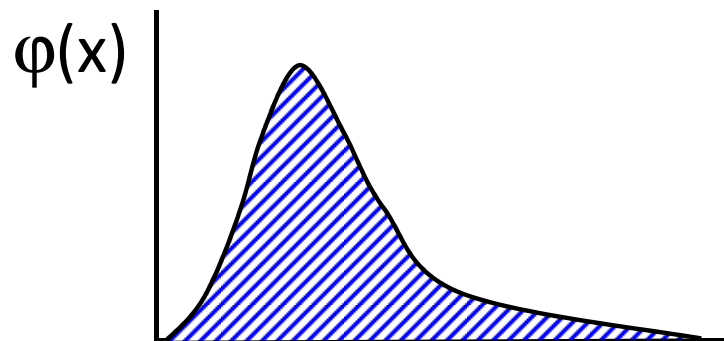
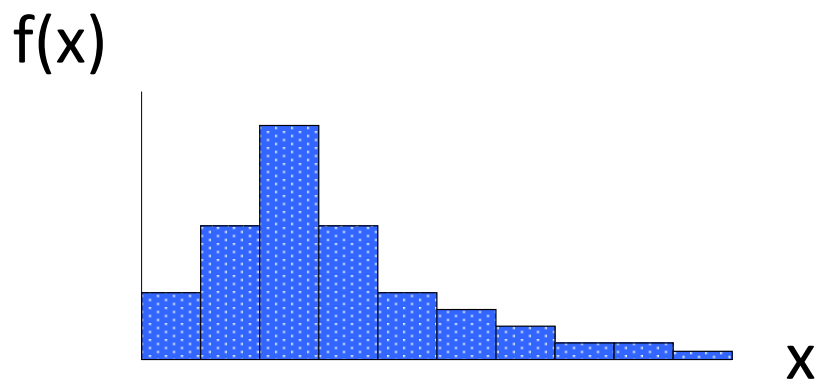
Rozložení



Distribuční funkce

Je - li dána
distribuční
funkce,
je dáno rozložení

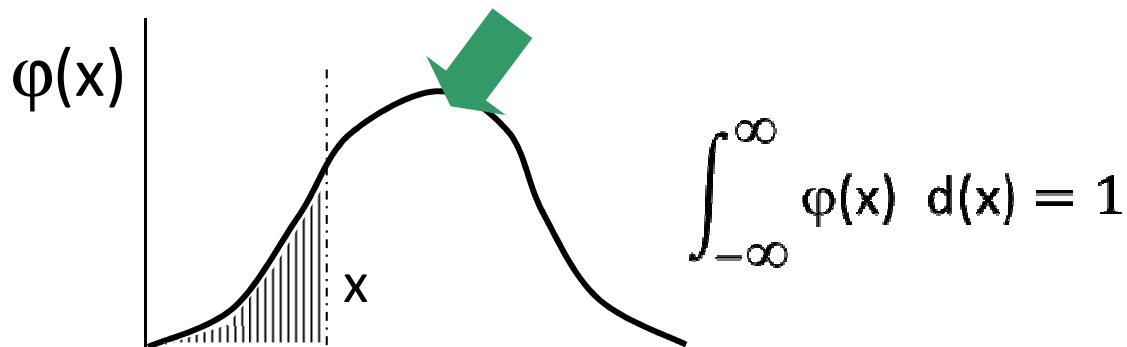
Výběrové rozložení hodnot lze modelově popsat a odhadnout tak pravděpodobnost výskytu X



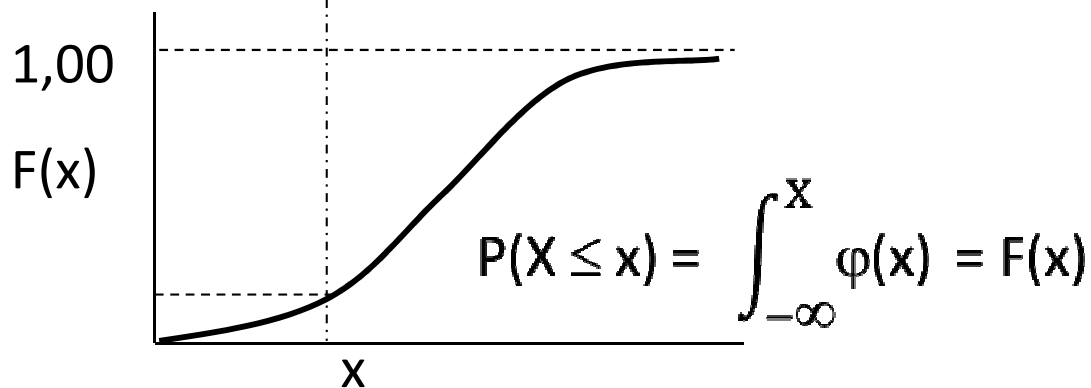
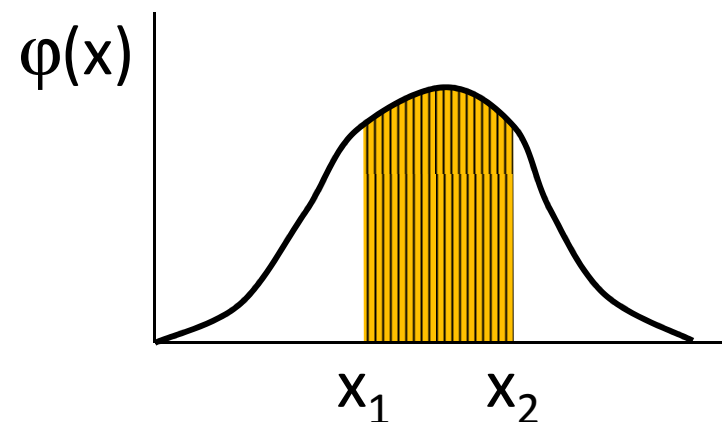
Distribuční funkce jako užitečný nástroj pro práci s rozložením



Plocha = relativní četnost



$F(x)$: Pravděpodobnost, že se X vyskytne v intervalu $(-\infty; x)$.



$$P(X \in (x_1; x_2)) = \int_{x_1}^{x_2} \varphi(x) = F(x_2) - F(x_1)$$

$\Phi(x)$... distribuční funkce

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 ?

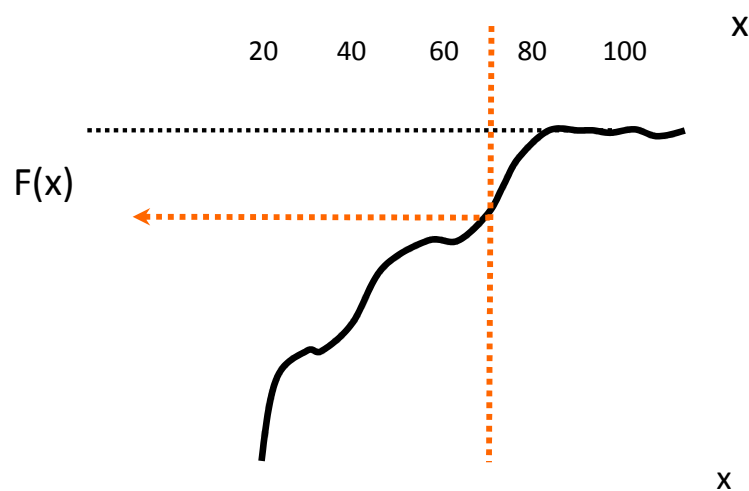
- frekvenční sumarizace spojitých dat

Grafické výstupy z frekvenční tabulky – spojitá data

$f(x)$



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



KVANTIL

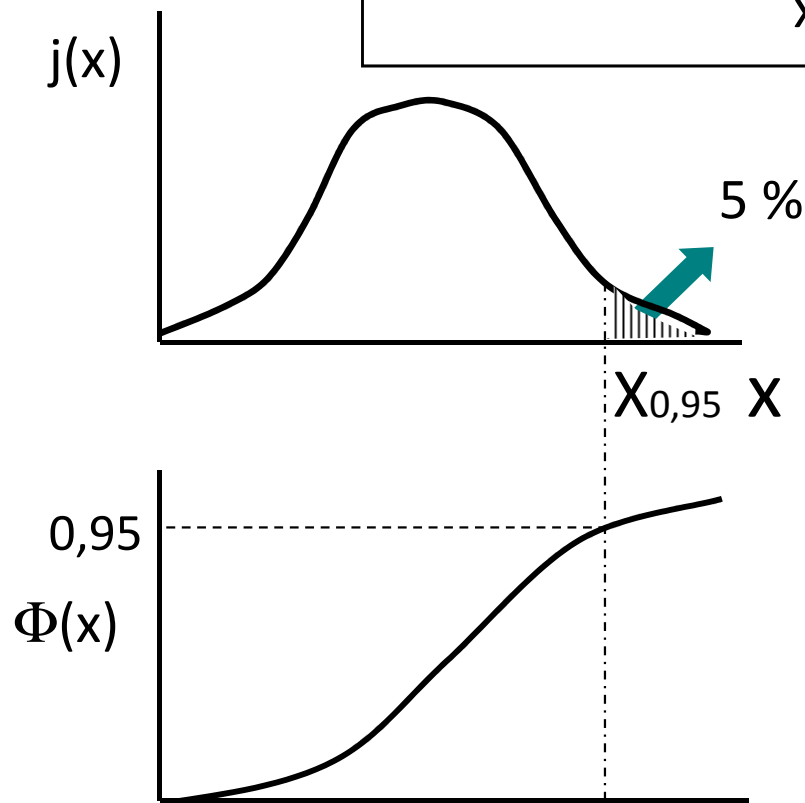
$X_{0.1}; X_{0.9}; X_{0.5}; X_{\theta}$

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



$\theta = 0,95$... pravděpodobnost

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



$F(x_\theta) = \theta$



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

Jakékoliv číslo na ose x je kvantilem*

* za předpokladu omezeného definičního oboru distribuční funkce