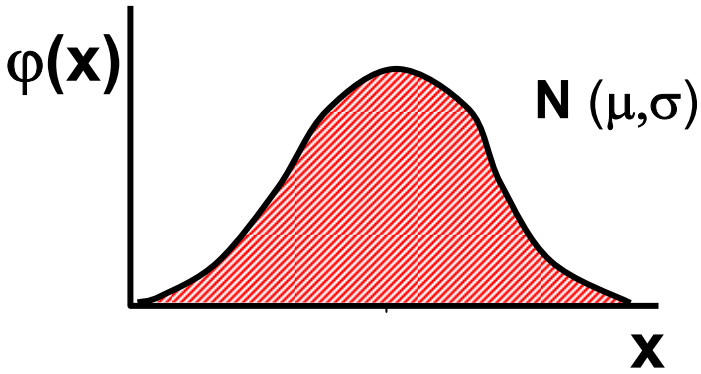
5. Modelová rozložení

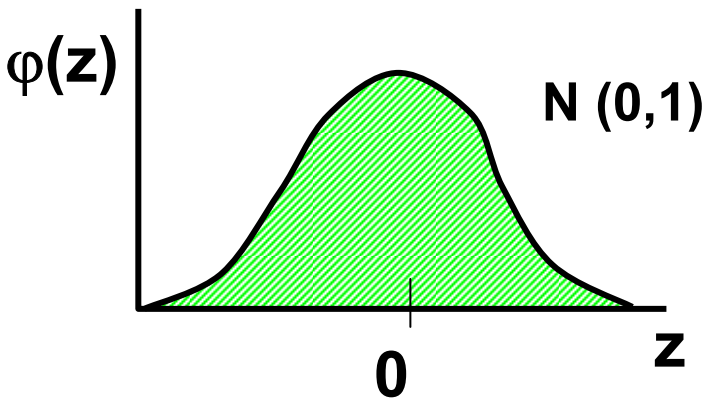
Rozložení hodnot jako model
Příklad - Normální rozložení



$$\varphi(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$z = \frac{x - \mu}{\sigma}$$

Standardizovaná forma

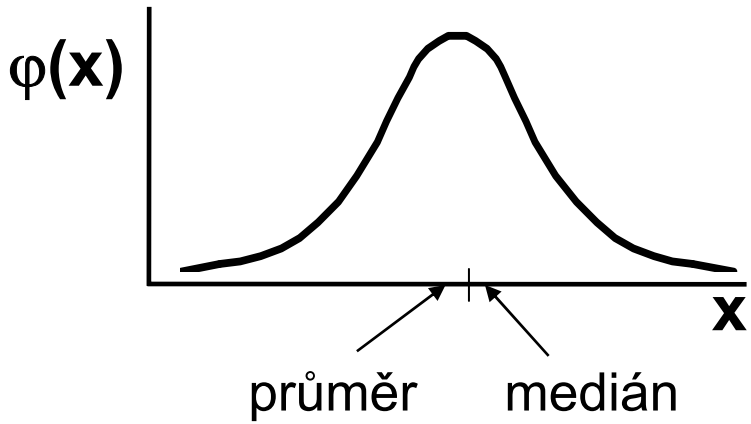


$$\varphi(z) = \frac{1}{\sqrt{2.\pi}}.e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

Parametry charakterizující normální rozložení a jejich význam

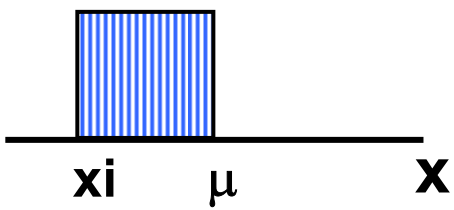
E (x) ~ $\bar{x}$ ~ μ
D (x) ~ s^2 ~ σ^2



a) **μ ~ $\bar{x}$**
průměr - ukazatel středu

b) **σ^2 ~ s^2**
rozptyl

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



c) **σ ~ s**
směrodatná odchylka

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

Pravidlo ± 3s

d) **koeficient variance**

$$c = s / \bar{x}$$

Interpretace parametrů normálního rozložení

Parametr středu

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

Parametr šířky

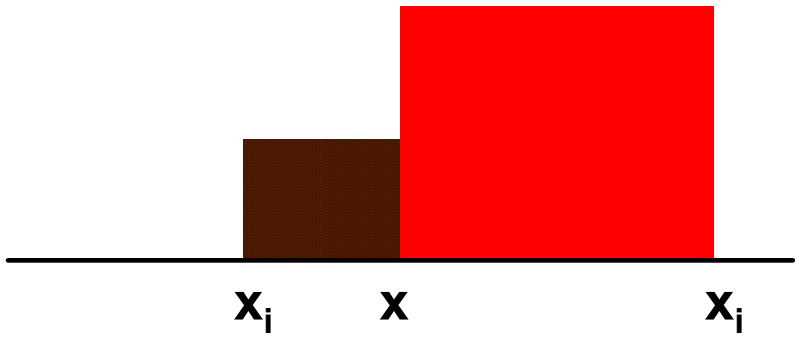
$$D(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}}{n-1} = s^2$$

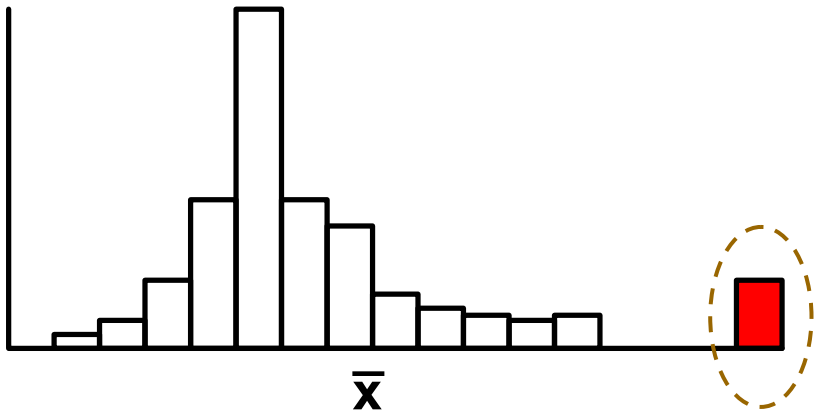
Směrodatná odch. (S.D.)

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

Rozptyl není univerzálním ukazatelem variability

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

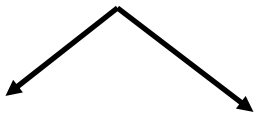




$\Rightarrow$ neúměrně zvýší s^2

Variační koeficient c (koeficient variance)

***Př.: 2* soubory dat -** *koncentrace Zn v rostlinné tkáni*



$\bar{x}_1 = 100$
$s_1 = 10$

$\bar{x}_2 = 10$
$s_2 = 2,6$

$$c_1 = \frac{s_1}{\bar{x}_1} = 0,10$$

$$c_2 = \frac{s_2}{\bar{x}_2} = 0,26$$

Normální rozložení jako model

I. Použitelnost modelu

A) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,4; 3.8

n = 7 opakování

medián = 1,8

průměr = $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}\sum_{i=1}^{7} x_i = \frac{1}{7}(1{,}2+1{,}4+1{,}6+1{,}8+2{,}0+2{,}4+3{,}8) = \frac{1}{7}14{,}2 = 2{,}03$

rozptyl (s^2) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{7}(x_i-2{,}03)^2}{6} = 0{,}766$

sm. odchylka (s) = $\sqrt{s^2} = \sqrt{0{,}766} = 0{,}875$

?

Je předpoklad normálního rozložení oprávněný ?
Jaký předpokládáte možný rozsah hodnot tohoto znaku ?

?

I. Použitelnost modelu

B) X: spojitý znak - hmotnost jedince (myši)

1,2; 1,4; 1,6; 1,8; 2,0; 2,2; 2,4; 3,8; 8,9

n = 9 opakování

medián = 2

průměr = $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}\sum_{i=1}^{9} x_i = \frac{1}{9}(1{,}2+1{,}4+1{,}6+1{,}8+2{,}0+2{,}2+2{,}4+3{,}8+8{,}9) = \frac{1}{9}25{,}3 = 2{,}81$

rozptyl (s^2) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{9}(x_i-2{,}81)^2}{8} = 5{,}79$

sm. odchylka (s) = $\sqrt{s^2} = \sqrt{5{,}79} = 2{,}269$

Jak hodnotíte model u těchto dat ?

Stochastické rozložení jako model

1 Předpoklad: Znak x je rozložen podle daného modelu ✓

2 Znak x je naměřen o n hodnotách s modelovými parametry: $\bar{x}$ a s → **Platnost modelu ?** ?

3 Znak x je převeden na formu odpovídající tabulkovému standardu: →

$$Z_i = \frac{x - \mu}{\sigma}$$

4 Využije se tabelované (modelové) distribuční funkce pro testy o rozložení hodnot x

Normální rozložení jako model - příklad

Tabulky distribuční funkce

- **Data z průzkumu jsou publikována jako:**

Kosti prehistorického zvířete:

n = 2000

průměrná délka = 60 cm

sm. odchylka (s) = 10 cm

✓ **Předpokládáme, že je oprávněný model normálního rozložení**

? Jaká je pravděpodobnost, že by velikost dané kosti překročila velikost 66 cm: P (x > 66) ? $Z = \frac{x - \mu}{\sigma}$

$P(x > 66) = 1 - P(x \leq 66)$ a platí, že $P(X \leq x) = F(X)$

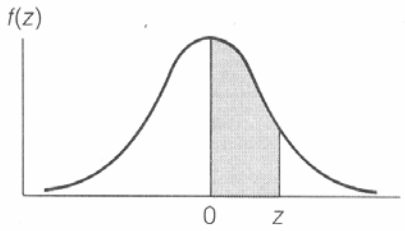
tedy $P(x > 66) = 1 - P(x \leq 66) = 1 - P(\frac{x - m}{s} \leq \frac{66 - 60}{10}) = 1 - F(0,6) = 0,27425$

? Kolik kostí mělo zřejmě délku větší než 66 cm ? $P(x > 66) * n = 0,27425 * 2000 = 548$

? Jaký podíl kostí ležel svou délkou v rozsahu x od 60 cm do 66 cm ?

$P(60 < x < 66) = P\left(\frac{60 - 60}{10} < Z < \frac{66 - 60}{10}\right) = F(0,6) - F(0) = 0,22575$ ➡ 22,6% kostí leží v rozsahu 60-66cm

Normální rozložení jako model - příklad



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Abridged from Table I of A. Hald, *Statistical Tables and Formulas* (New York: John Wiley & Sons, Inc.), 1952. ©1952 by John Wiley & Sons, Inc. Reproduced by permission.

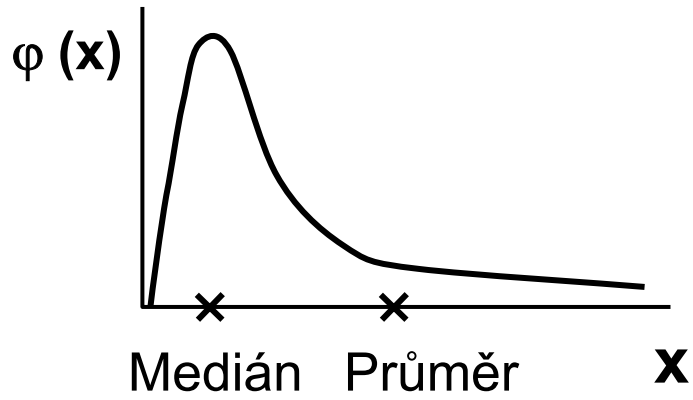
Stručný přehled modelových rozložení I.

Rozložení	Parametry	Stručný popis
Normální	**Průměr (μ)** **Rozptyl (σ^2)**	Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci.
Log-normální	**Medián** **Geometrický průměr** **Rozptyl (σ^2)**	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
Weibullovo	**α - parametr tvaru** **β - parametr rozsahu hodnot**	Změnou parametru a lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhahu LC_{50} nebo EC_{50} u testů toxicity.
Rovnoměrné	**Medián** **Geometrický průměr** **Rozptyl (σ^2)**	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
Triangulární	**$f(x) = [b - ABS(x - a)] / b^2$** **$a - b < x < a + b$**	Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
Gamma	**Parametry distribuční funkce:** **α - parametr tvaru** **β - parametr rozsahu hodnot**	Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. χ^2 rozložení je rozložení typu Gamma. Gamma rozložení s a = 1 je známo jako exponenciální rozložení.

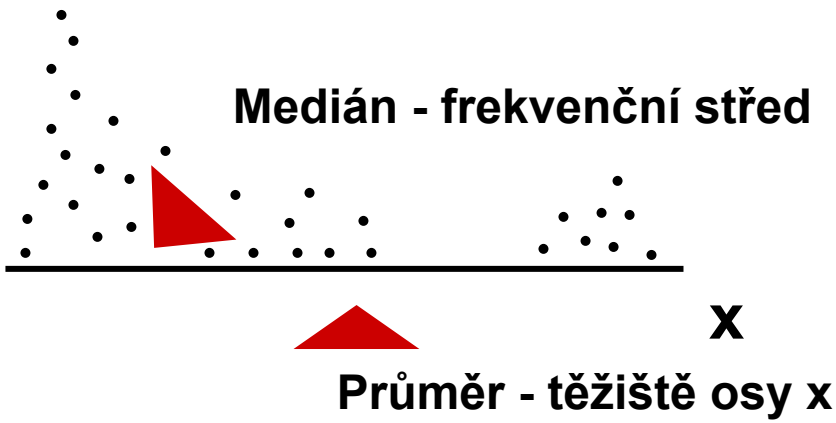
Stručný přehled modelových rozložení II.

Rozložení	Parametry	Stručný popis
Beta	**Parametry distribuční funkce:** **α - parametr tvaru** **β - parametr rozsahu hodnot**	Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu.
Studentovo	**Stupně volnosti - uvažuje velikost vzorku** **Průměr** **Rozptyl**	Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory (n > 100) se limitně blíží k normálnímu rozložení.
Pearsonovo	**Stupně volnosti - uvažuje velikost vzorku**	Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat.
Fisher-Snedecorovo	**Dvojí stupně volnosti - uvažuje velikost dvou vzorků**	Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd.

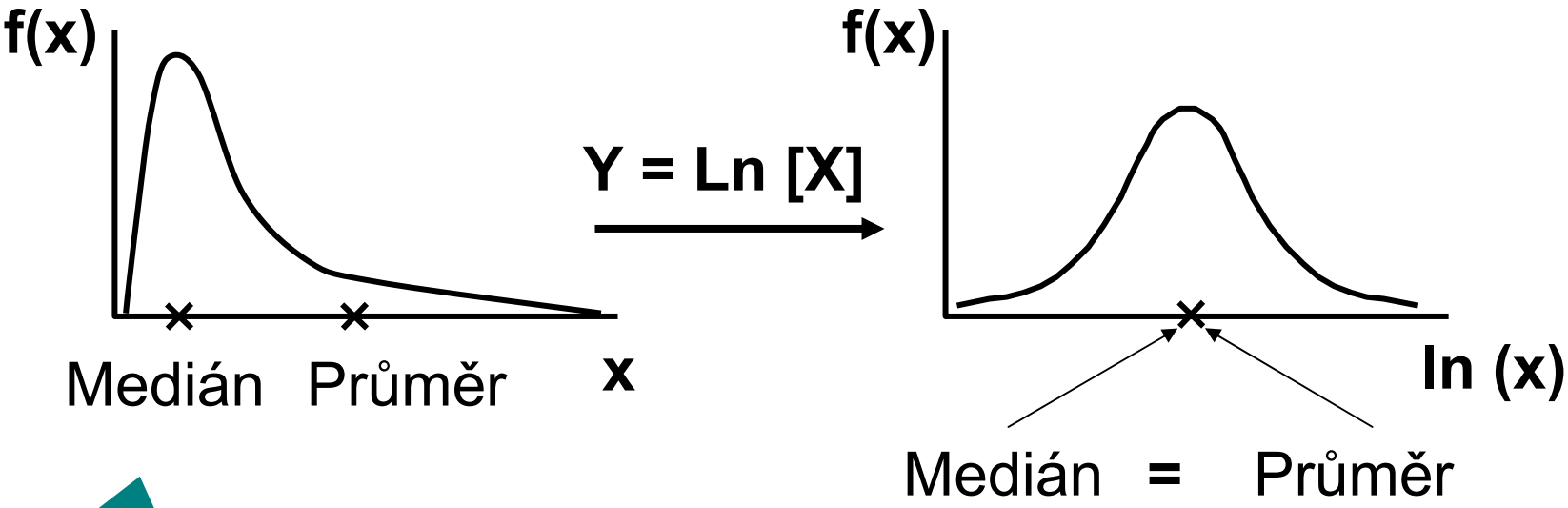
Log-normální rozložení jako častý model reálných znaků



U asymetrických rozložení je medián velmi vhodným alternativním ukazatelem středu



Log-normální rozložení lze jednoduše transformovat



EXP (Y) = Geometrický průměr X

$$\overline{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}$$

$\overline{Y}$ ± Standardní chyba

Transformace dat - legitimní úprava rozložení

✓ **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**

Logaritmická transformace

Logaritmická transformace je velmi vhodná pro data s odlehlými hodnotami na horní hranici rozsahu. Při porovnání průměrů u více souborů dat je pro tuto transformaci indikující situace, kdy se s rostoucím průměrem mění proporcionálně i směrodatná odchylka, a tedy jednotlivé proměnné mají stejný koeficient variance, ačkoli mají různý průměr.

Za takovéto situace přináší logaritmická transformace nejen zeslabení asymetrie původního rozložení, ale také vyšší homogenitu rozptylu proměnných. Pro transformaci se nejčastěji používá přirozený logaritmus a pokud jsou v původním souboru dat nulové hodnoty, je vhodné použít operaci **Y = ln (X+1)**.

Je-li průměr logaritmovaných dat (tedy průměrný logaritmus) zpětně transformován do původních hodnot, výsledkem není aritmetický, ale geometrický průměr původních dat.

Transformace dat - legitimní úprava rozložení

✓ **Základní typy transformací vedou k normalitě rozložení nebo k homogenitě rozptylu**

Odmocninová transformace

Transformace je vhodná pro proměnné mající Poissonovo rozložení, tedy proměnné vyjadřující celkový počet nastání určitého jevu (spíše vzácného) v **n** nezávisle opakovaných pokusech. Obecněji lze tento typ transformace doporučit v případě normalizace dat typu počtu jedinců (buněk, apod.). Jde o transformaci:

$$Y=\sqrt{x} \quad \textbf{nebo} \quad Y=\sqrt{x+1} \quad \textbf{nebo} \quad Y=\sqrt{x}+\sqrt{x+1}$$

Transformace s přičtenou hodnotou 1 jsou efektivní, pokud **X** nabývá velmi malých nebo nulových hodnot. Situace indikující vhodnost odmocninové transformace je také proporcionalita výběrového rozptylu a průměru, tedy obecně jestliže $\mathbf{s^2_x = k}$ (výběrový průměr).

Arcsin transformace

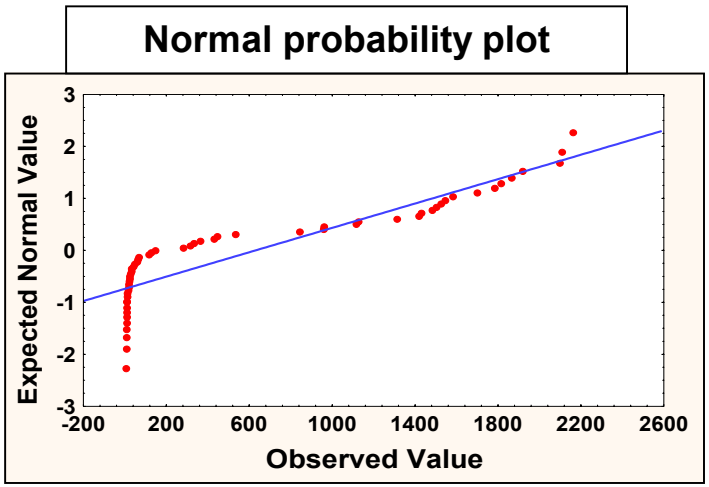
Tzv. **úhlová transformace** - velmi vhodná pro data typu podílů výskytu určitého jevu (znaku) mezi **n** hodnocenými jedinci - tedy pro data mající binomické rozložení. Pokud se určitý znak vyskytuje r-krát mezi **n** možnostmi (jedinci, opakováními), pak lze vyjádřit relativní četnost jeho výskytu jako **p = r/n** s variabilitou **p.(1-p)/n**. Arcsin transformace odstraní ze souborů dat podíly blízké 0 nebo 1, a tak efektivně sníží variabilitu odhadů středu. Transformace však není schopná odstranit variabilitu vyvolanou rozdílným počtem opakování v jednotlivých variantách - v takovém případě lze doporučit provedení vážených transformací dat. Velmi častou formou této transformace je:

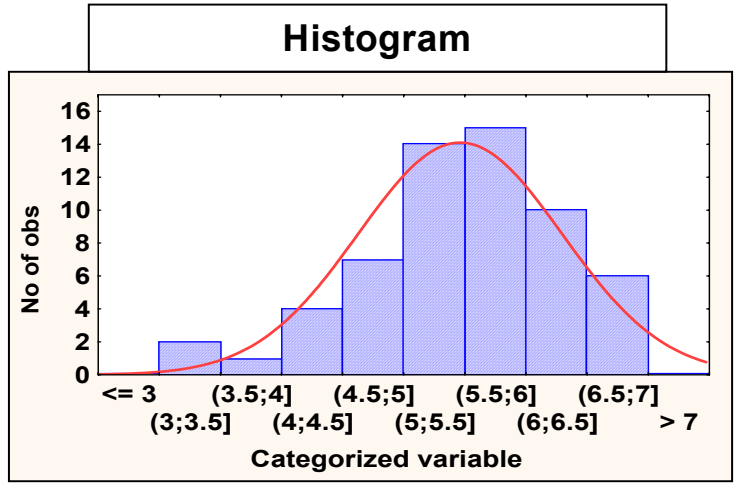
$$Y = \arcsin \sqrt{p}$$

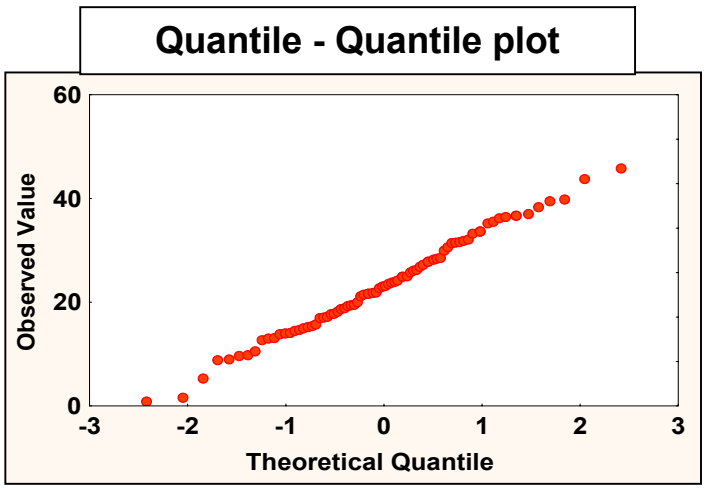
- tedy transformace podílů do hodnot, jejichž sinus je roven druhé odmocnině původních hodnot. Pokud celkový počet jedinců (opakování), mezi kterými je výskyt znaku monitorován, je n < 50, pak lze doporučit velmi efektivní empirická opatření pro transformaci podílů blízkých 0 nebo 1. Pro tento případ lze nahrazovat nulové podíly hodnotou 1/4n a 100 % podíly hodnotou (n-1/4)/n. Pokud se mezi hodnotami vyskytuje větší množství krajních hodnot (menší než 0,2 a větší než 0,8), lze doporučit transformaci:

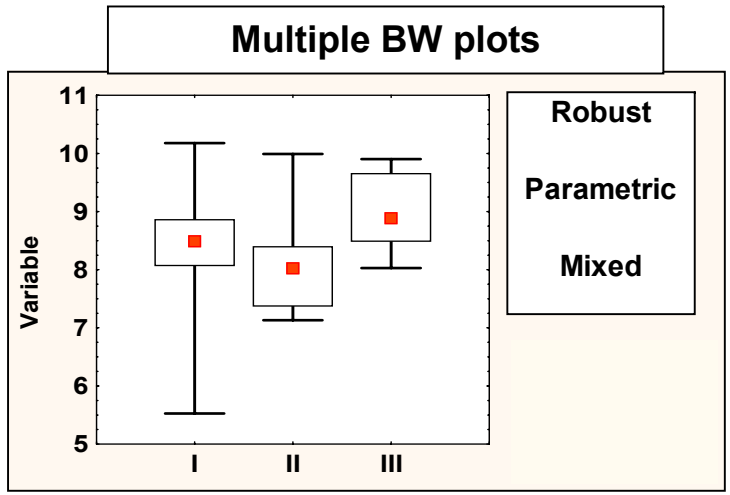
$$Y = \frac{1}{2}\left[\arcsin \sqrt{\frac{x}{n+1}} + \arcsin \sqrt{\frac{x+1}{n+1}}\right]$$

Testy o rozložení, grafický průzkum rozložení









Testy o rozložení: Kolmogorov-Smirnov test, Shapiro-Wilks test, $\chi 2$ test

6. Sumární statistika

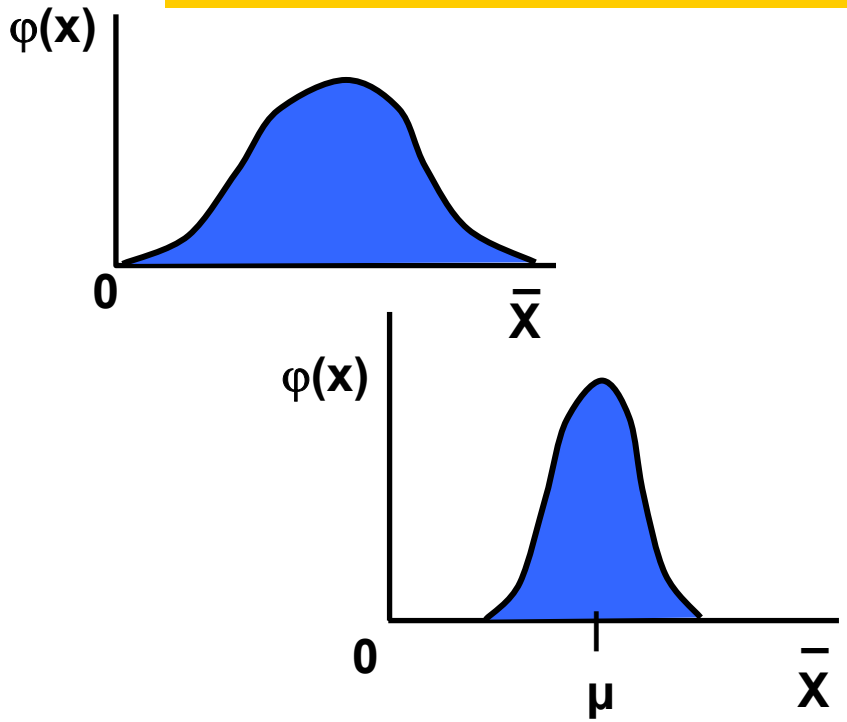
Sumární statistika I

Znak X

- Medián
- Min Max
- kvantily(percentily)
  - horní kvartil
  - dolní kvartil
- Rozsah
- mezikvartilová odchylka

Střed znaku X

- průměr
- SD, SE
- interval spolehlivosti



Sumární statistika II

Posuďte správnost následujících výstupů (X: výška rostlin v cm):

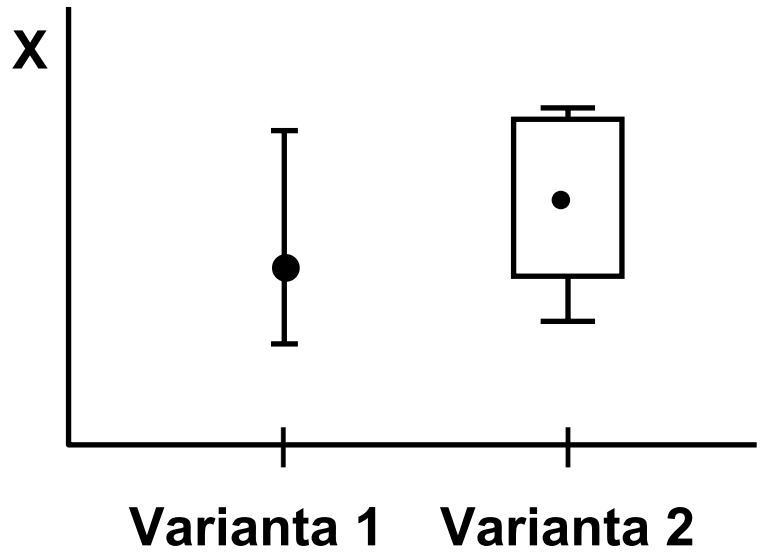
1. $\bar{x} = 20$ $s = 5$
   $M = 22$ Rozsah = 34

2. $\bar{x} = 200$
   Min = 90 Max = 330

3. 25% kvantil: 15
   Medián: 16
   75% kvantil: 48

4. $\bar{x} = 20$ $s = 12$
   $M = 8$

Různá zobrazení v BW Plotech:



Sumární statistika III

Výsledkem průzkumu 23 lokalit s cílem zjistit rozsah zamoření půdy těžkými kovy byli mimo jiné i dvě proměnné udávající koncentraci Zn a Pb v půdě. Následující tabulka uvádí základní statistické parametry těchto proměnných.

a) Vysvětlete význam jednotlivých parametrů.

b) Porovnejte medián s průměrem a pro každou proměnnou udělejte závěr o symetričnosti jejího rozložení.

c) Porovnejte hodnoty jednotlivých kvartilů a usuďte podle nich na symetričnost rozložení proměnných.

d) Má zde variační koeficient stejný význam jako např. u proměnné, která je tvořena výsledky opakovaného stanovení jedné látky v jednom vzorku?

Parametr	Zn	Pb
Průměr	20,97	15,43
Medián	15,1	15,4
Modus	12,8	16
Geometrický průměr	18,17	14,66
Rozptyl	223,69	24,56
Směrodatná odchylka	14,96	4,56
Rozsah	54,4	20,6
Spodní kvartil	12,9	11,1
Horní kvartil	19,9	17,6
Mezikvartilová odchylka	7	6,5
Šikmost	2,55	0,54
Špičatost	5,82	0,3
Variační koeficient	71,32	32,12

Sumární statistika III

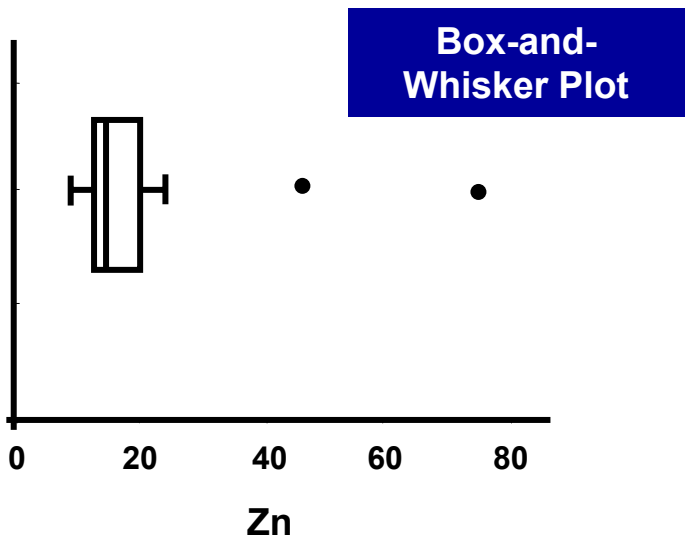
a) Testování normality proměnných z příkladu 6. (Zn, Pb) Kolgomorov-Smirnovovým testem poskytlo následující výsledky :

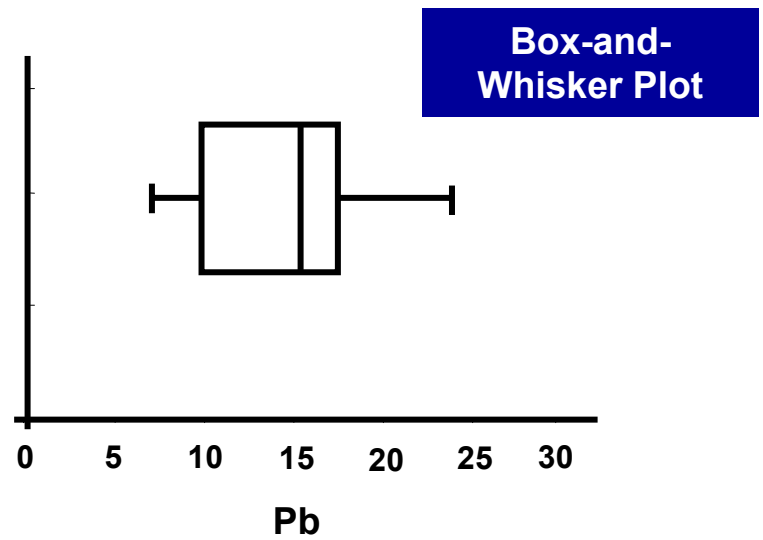
$$Zn : D_{\max} = 0{,}326$$

$$Pb : D_{\max} = 0{,}125$$

Porovnejte tato čísla s tabelovanými kritickými hodnotami a uveďte hladinu významnosti pro zamítnutí nulové hypotézy.

b) Velmi užitečným způsobem zobrazování rozložení proměnných je následující graf (opět pro proměnné Zn a Pb). Porovnejte grafy se statistickým rozborem proměnných uvedeným v příkladě 6.

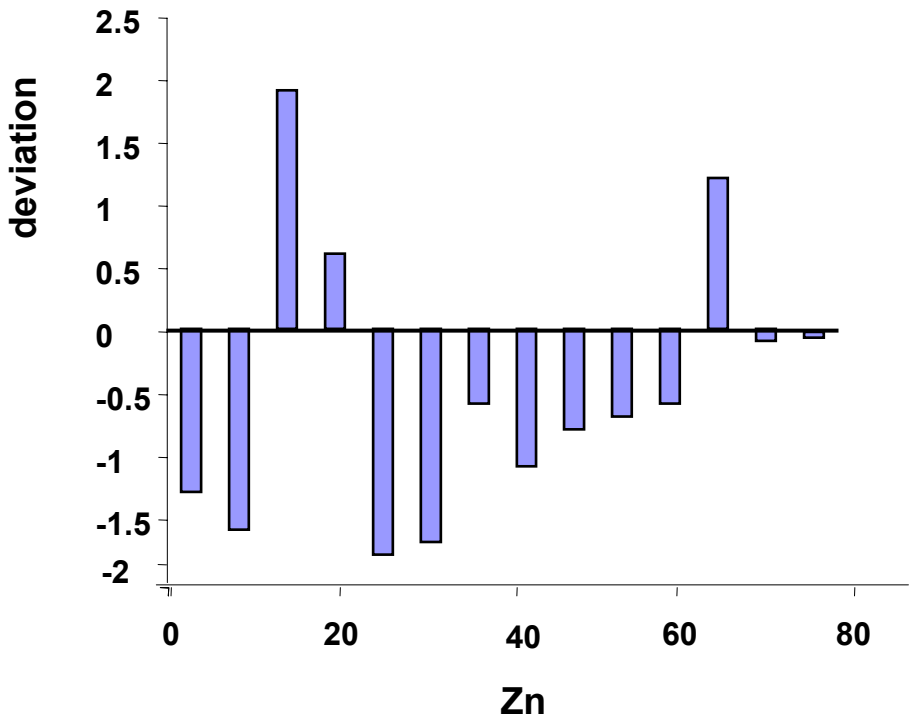




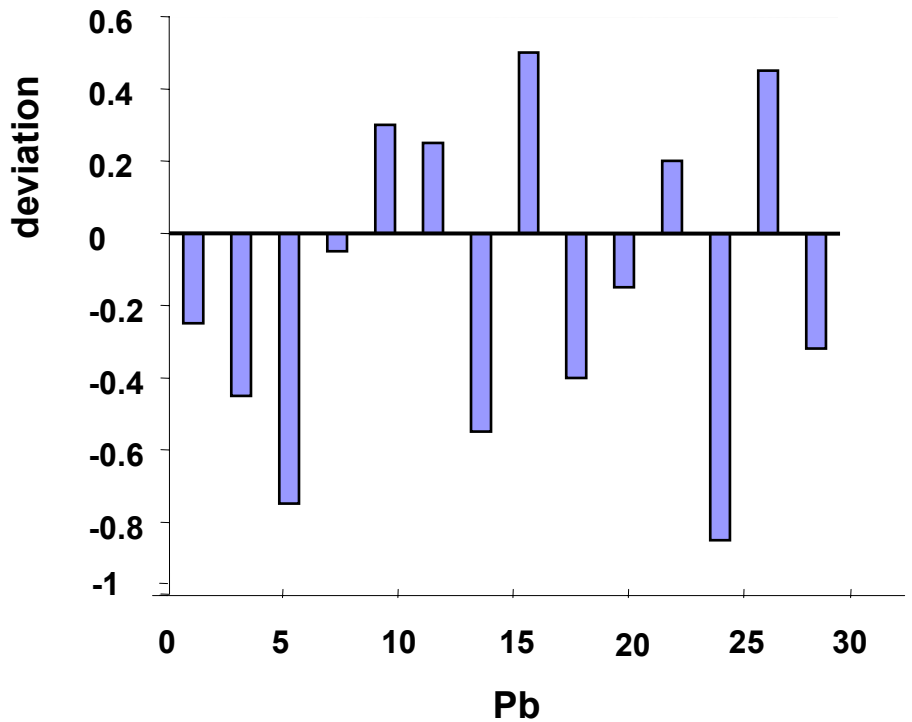
Sumární statistika I

Následuje přehled jednoduchých grafů, které umožňují posouzení normality proměnných. Porovnejte jejich vypovídací schopnost (opět pro proměnné Zn a Pb).

Rootgram

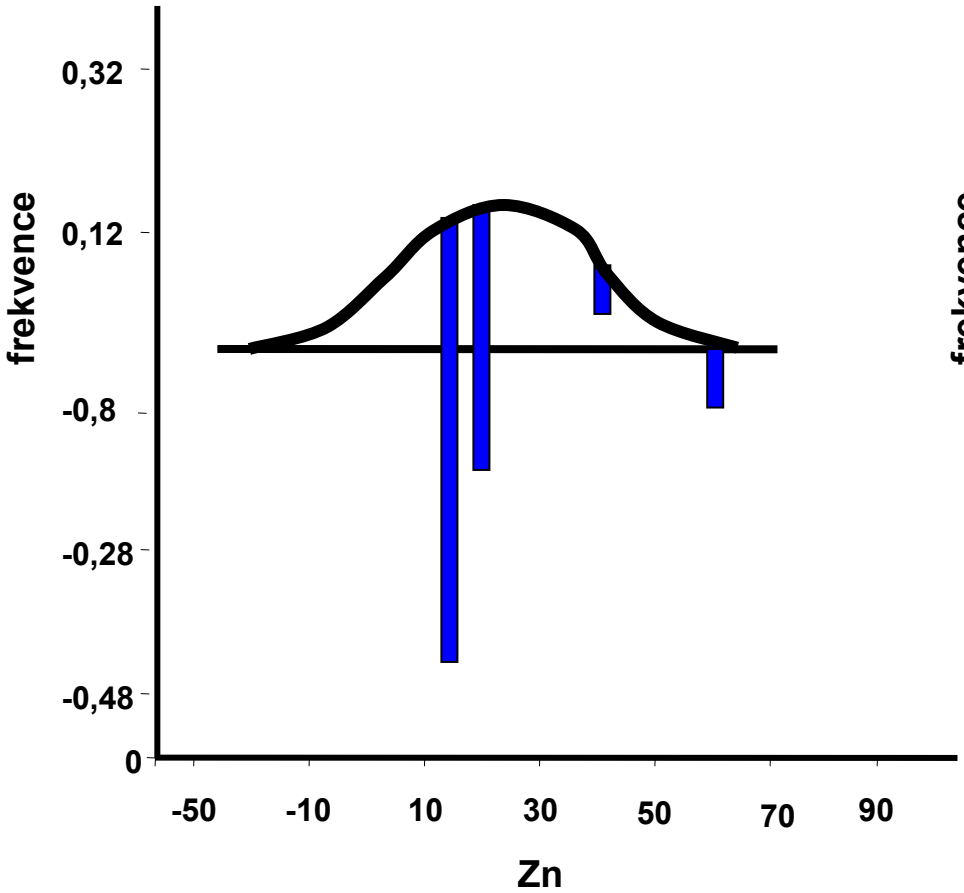


Rootgram

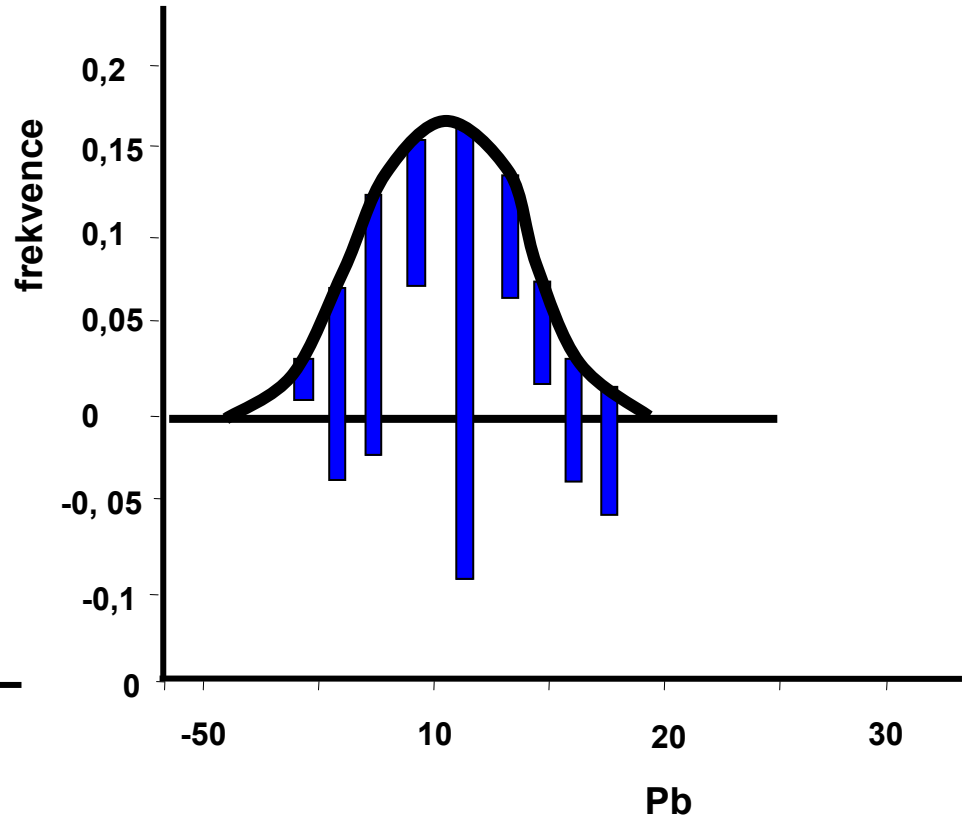


Sumární statistika I

Hanging Histobars.

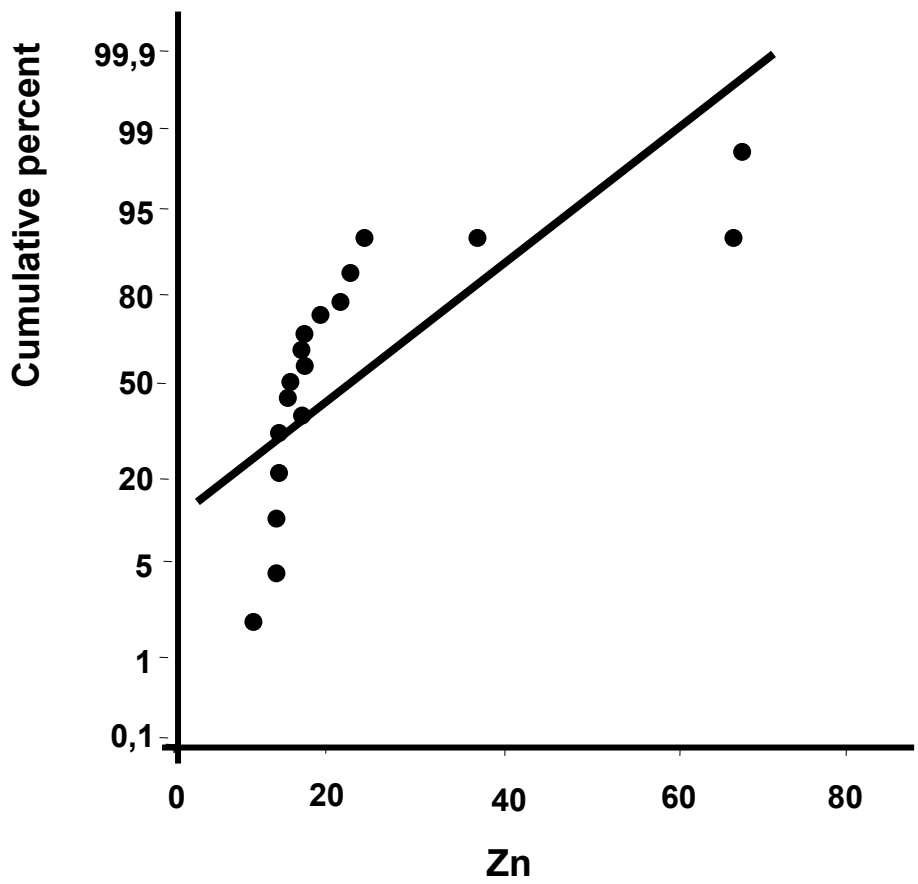


Hanging Histobars.

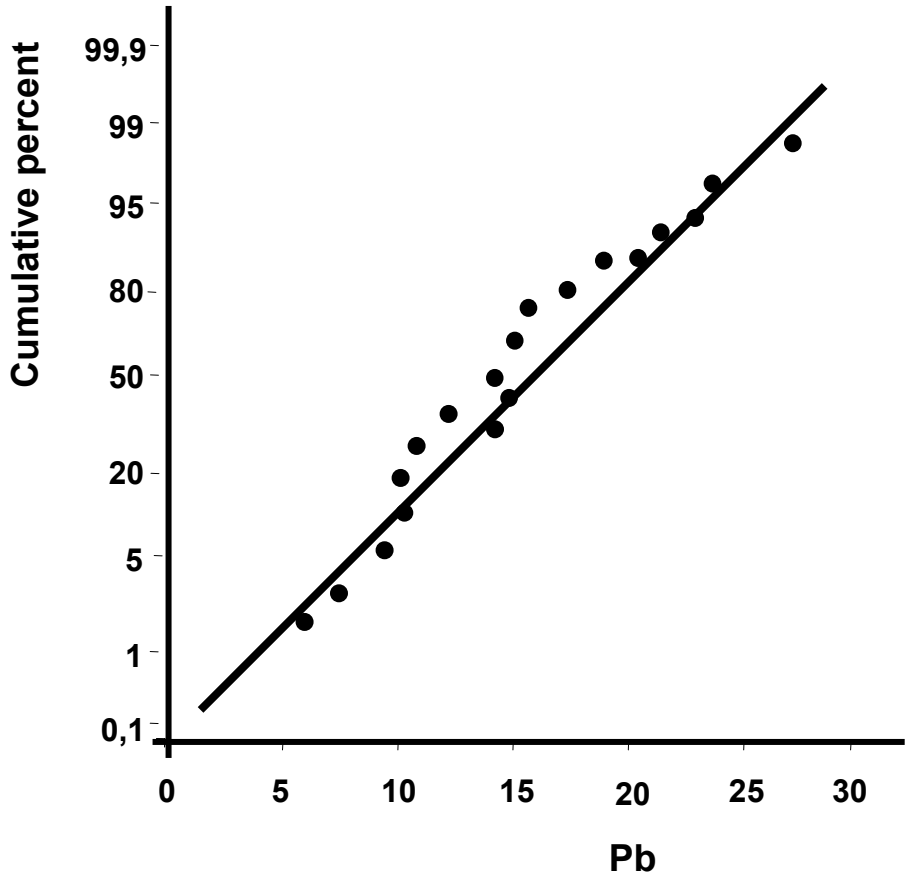


Sumární statistika I

Normal Probability Plot

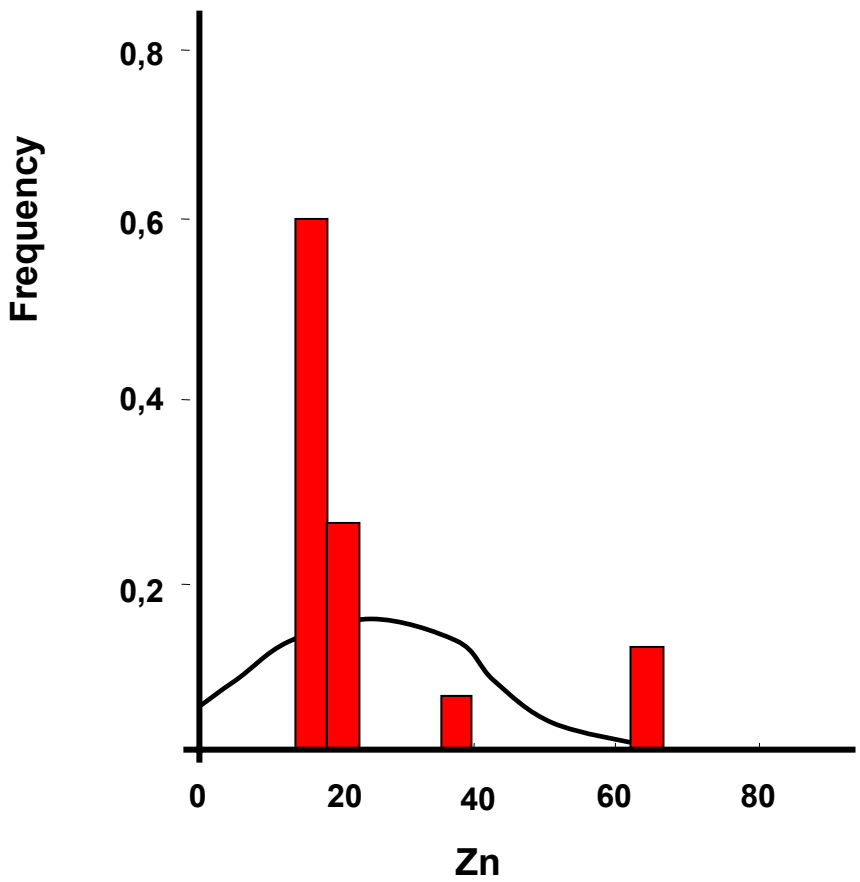


Normal Probability Plot

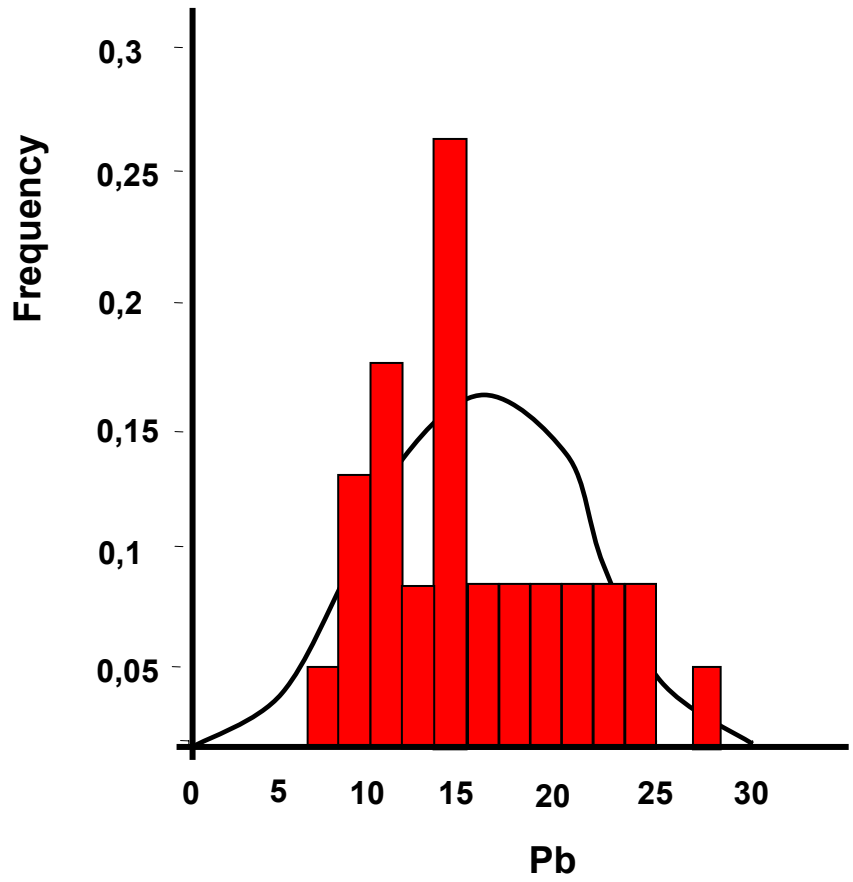


Sumární statistika I

Frequency Histogram



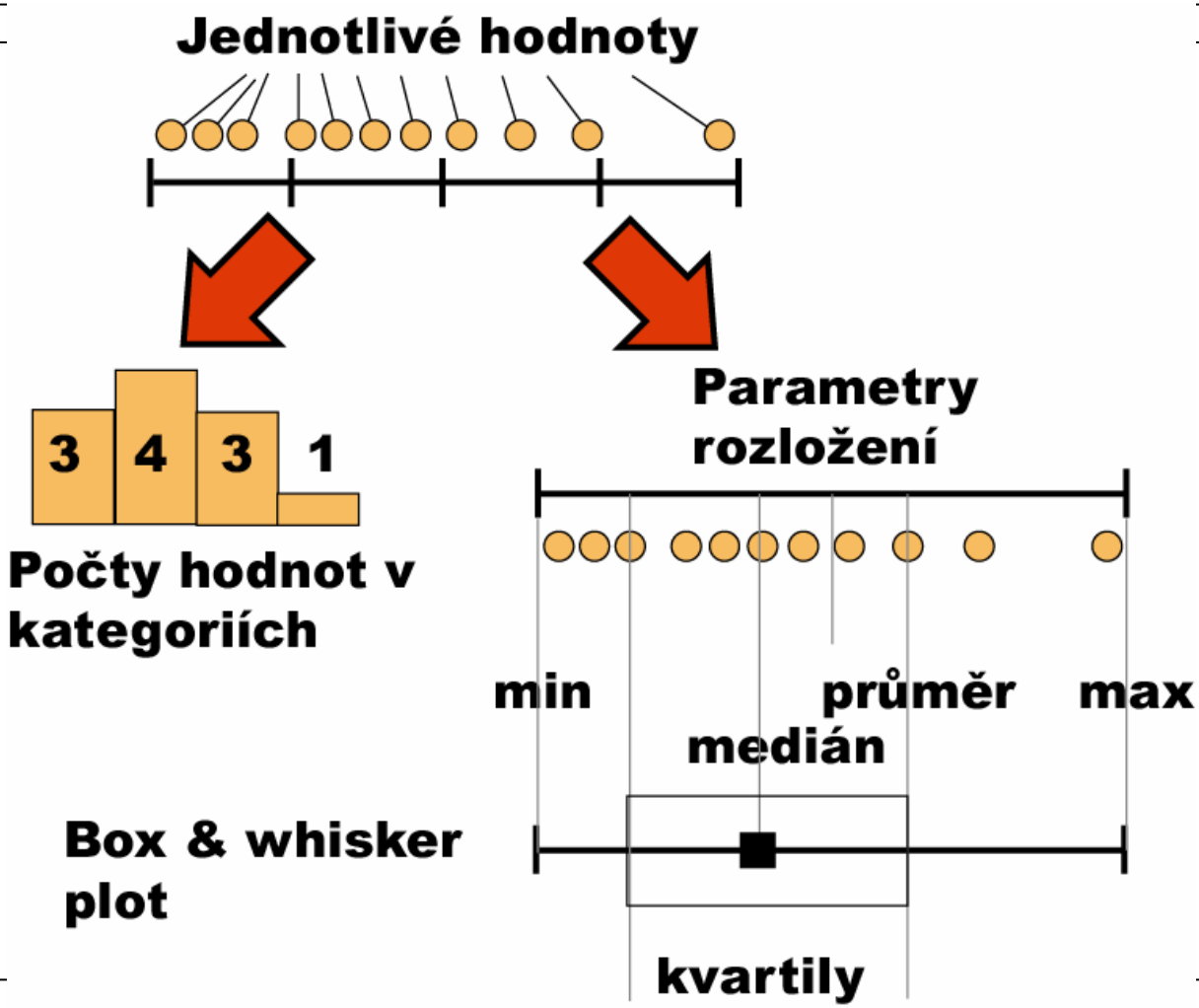
Frequency Histogram



Typy proměnných

- **Kvalitativní/kategorická**
  - binární - ano/ne
  - nominální - A,B,C … několik kategorií
  - ordinální - 1<2<3 …několik kategorií a můžeme se ptát, která je větší
- **Kvantitativní**
  - nespojitá – čísla, která však nemohou nabývat všech hodnot (např. počet porodů)
  - spojitá – teoreticky jsou možné všechny hodnoty (např. krevní tlak)

Řada dat a její vlastnosti

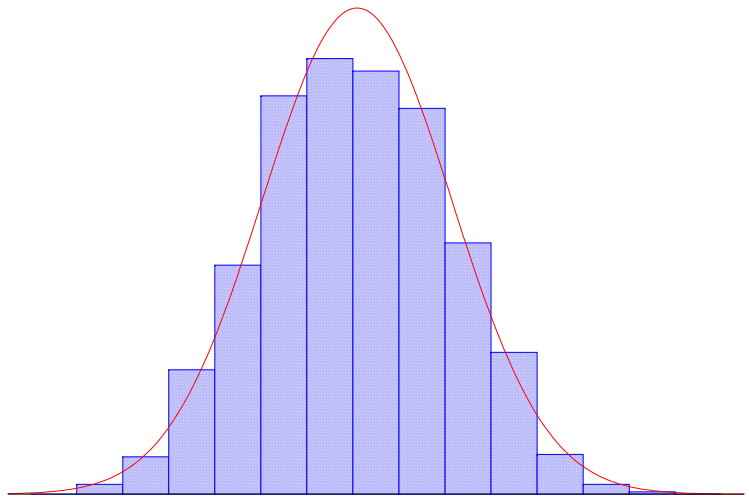


Frekvenční rozložení

Kategorie	Četnost
B	5
C	8
D	1

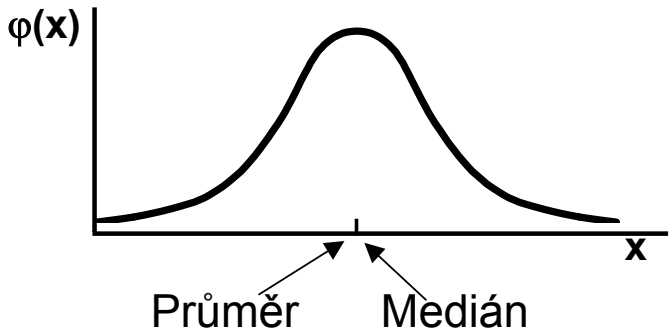
Kvalitativní data
Tabulka s četností jednotlivých kategorií.

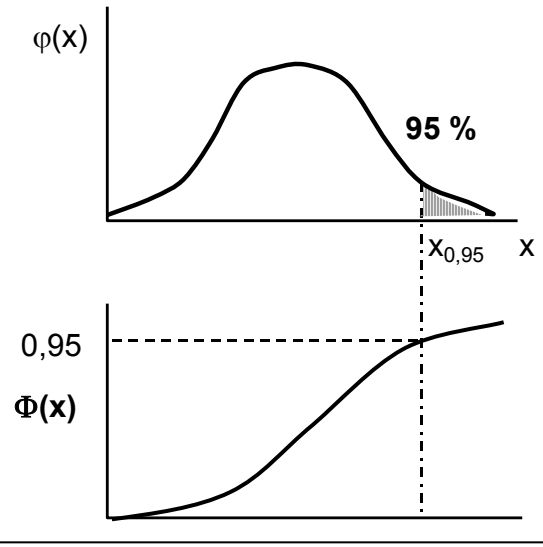
Kvantitativní data
Četnost hodnot rozložení v jednotlivých intervalech.



Parametry rozložení

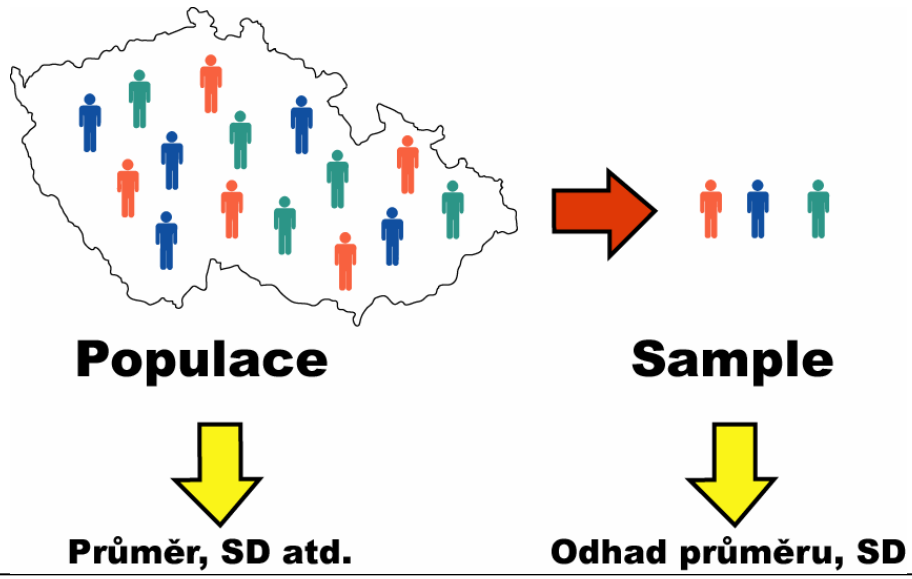
- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - Středu (medián, průměr, geometrický průměr)
  - Šířky rozložení (rozsah hodnot, rozptyl, směrodatná odchylka)
  - Tvaru rozložení (skewness, kurtosis)
  - Kvantily rozložení – kolik % řady dat leží nad a pod kvantilem





Populace a vzorek

- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (**sample**) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme odhady parametrů rozložení



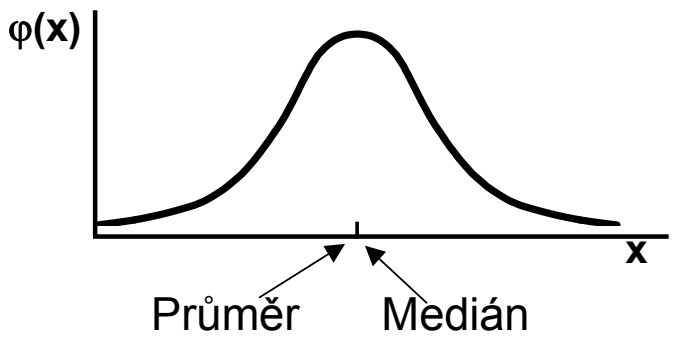
Ukazatele středu rozložení I

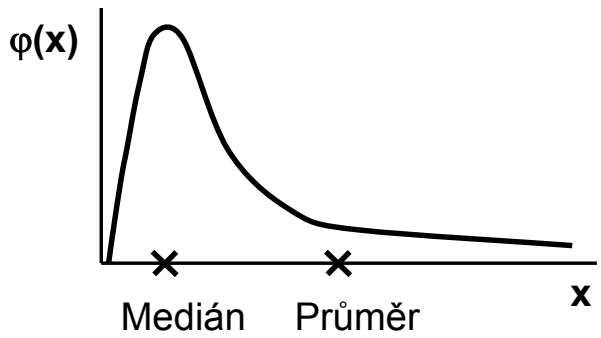
- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde $\mathbf{x_i}$ jsou jednotlivé hodnoty a **n** jejich počet

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

- **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem

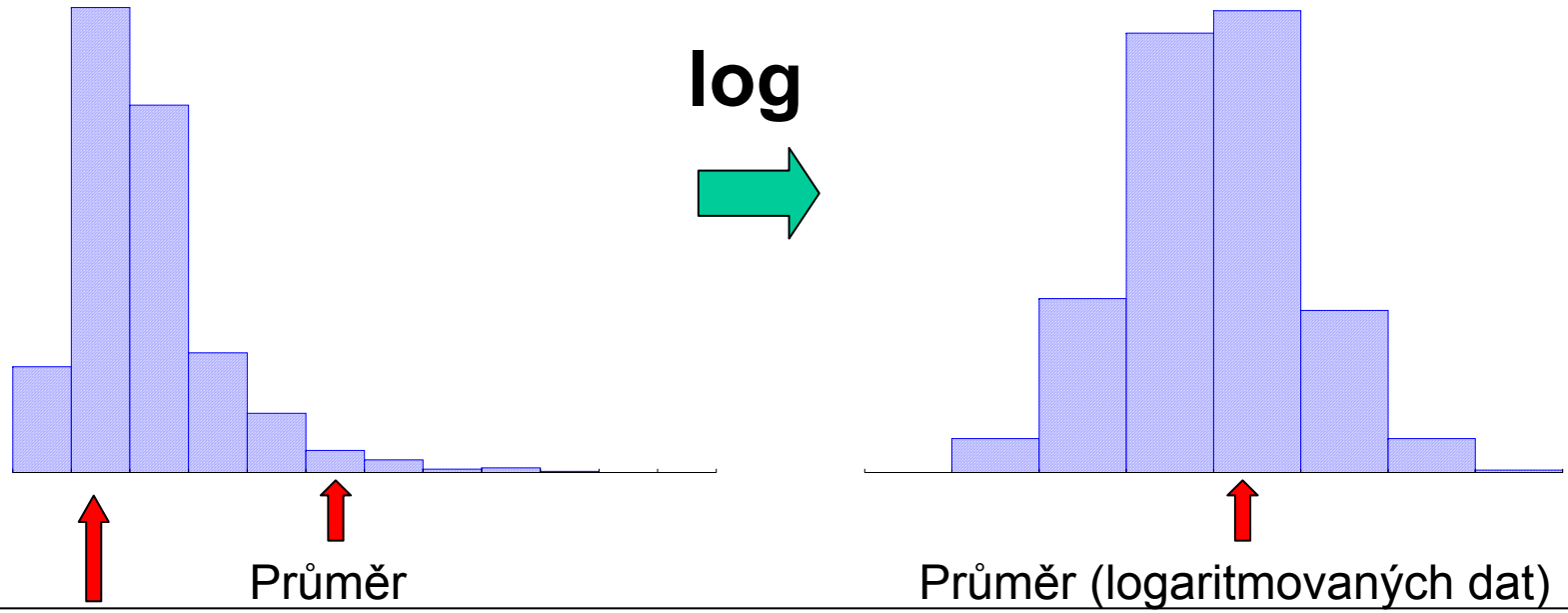
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné





Ukazatele středu rozložení II.

- Geometrický průměr – antilogaritmus průměru logaritmovaných dat, je vhodný pro doleva asymetrická data (lognormální rozložení), která jsou v biologii velmi častá, jeho hodnota v podstatě odpovídá mediánu
- Takto asymetrická data je možné převést logaritmickou transformací na normální rozložení



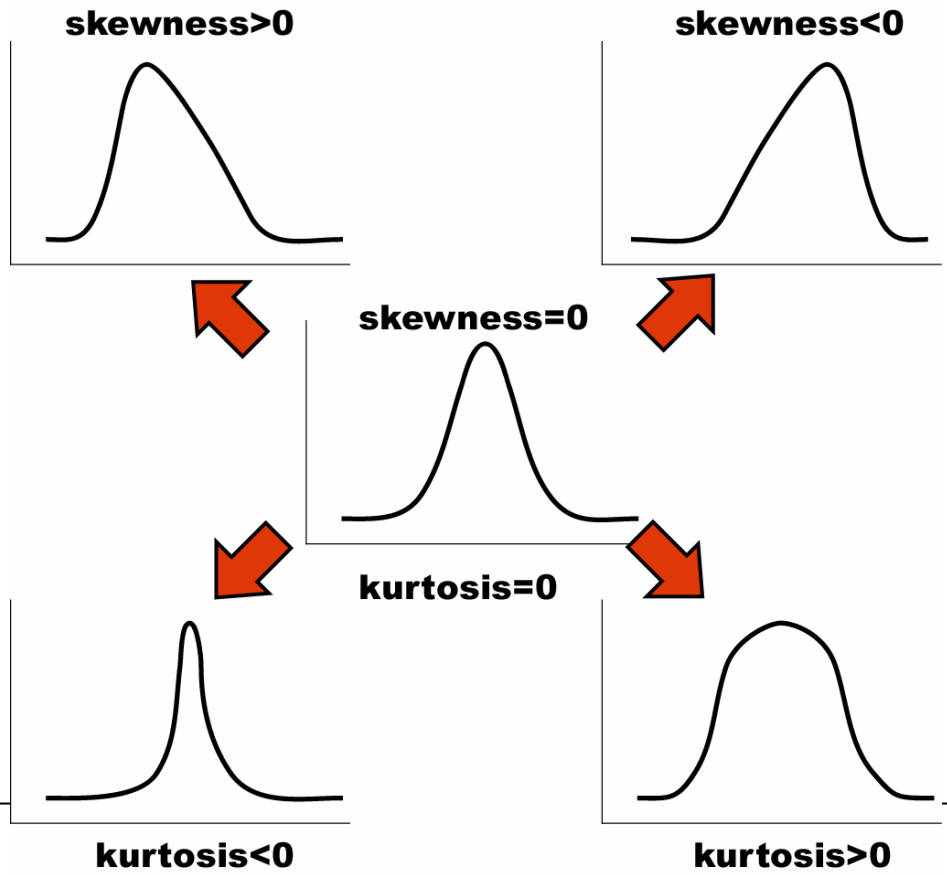
Medián, geometrický průměr

Ukazatele šířky rozložení

- **Rozptyl** je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru. $s^2 = \frac{\sum (x_i - x)^2}{n-1}$
- Obdobně jako u průměru je jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení
- **Směrodatná odchylka** je druhá odmocnina z rozptylu
- **Koeficient variance** - podíl SD ku průměru (u normálního rozložení by se 95% hodnot mělo vejít do průměr ⌖ 3 SD), pokud je SD větší než 1/3 průměru jsou teoreticky pravděpodobné záporné hodnoty v rozložení – ukazatel problémů s normalitou dat

Ukazatele tvaru rozložení

- **Skewness** – ukazatel „šikmosti“ rozložení, asymetrie rozložení
- **Kurtosis** – ukazatel „špičatosti/plochosti“ rozložení

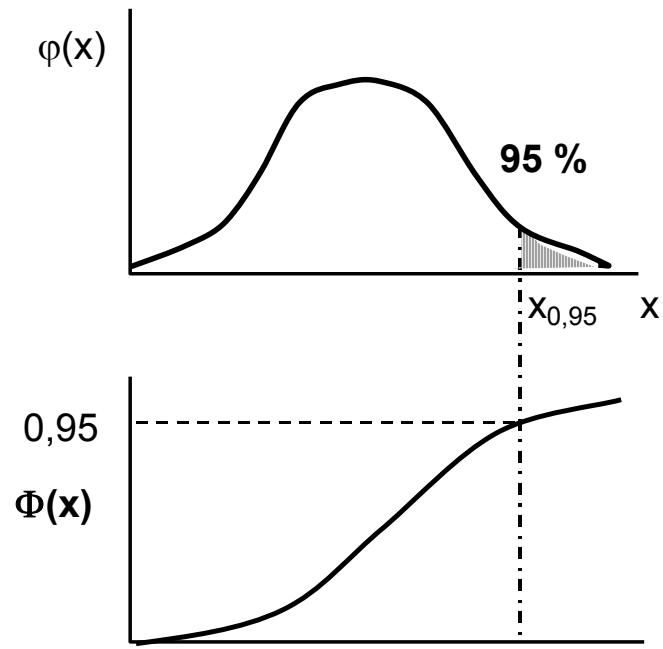


Další parametry rozložení

- **Počet hodnot –** důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Střední chyba odhadu průměru** - je založena na směrodatné odchylce rozložení a **počtu hodnot**, vlastně jde o směrodatnou odchylku rozložení průměru. Říká jak přesný je náš výpočet průměru. Čím větší počet hodnot rozložení, tím je náš odhad skutečného průměru přesnější.
- **Suma hodnot**
- **Modus** – nejčastější hodnota, vhodný např. při kategoriálních datech
- **Minimum, maximum**
- **Rozsah hodnot**
- **Harmonický průměr** - převrácená hodnota průměru převrácených hodnot (vždy platí harmonický průměr < geometrický průměr < aritmetický průměr)

Distribuční funkce

- Definice kvantilu dle distribuční funkce - Kvantil rozložení ($X_{0,95}$) je číslo, jehož hodnota distribuční funkce je rovna pravděpodobnosti, pro kterou je kvantil definován ($\Phi(x)$ … distribuční funkce), tj. pokud vezmeme nějaký bod rozložení a porovnáme jej s tímto bodem (kvantilem), máme 95% pravděpodobnost, že bude menší než hodnota kvantilu ($X_{0,95}$).
- Pomocí distribuční funkce můžeme určit jaký podíl hodnot rozložení je menší než daná hodnota – využití při statistických testech



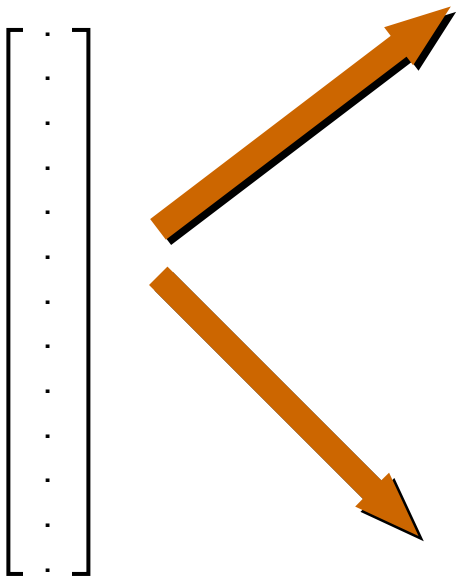
7. Strategie sumarizace a zviditelnění dat

Zviditelnění dat a jeho zásadní strategie

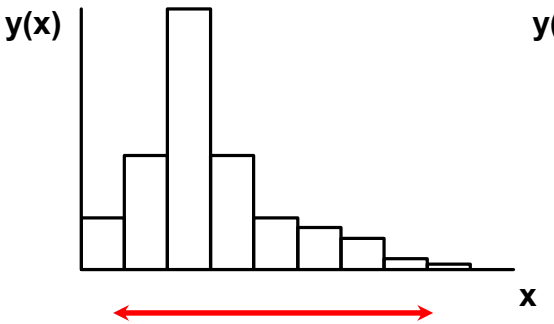
Popis

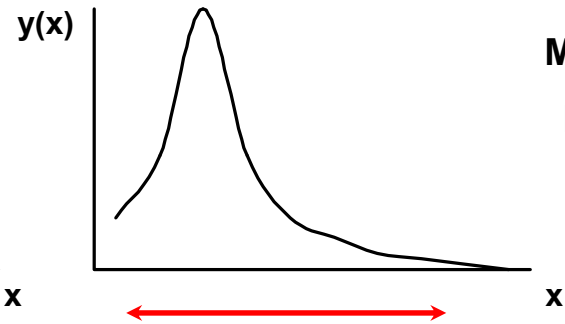
Naměřená data

x



A. Zviditelnění reálných dat – výběrové rozložení



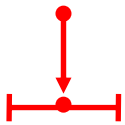


MIN / MAX

Kvantily

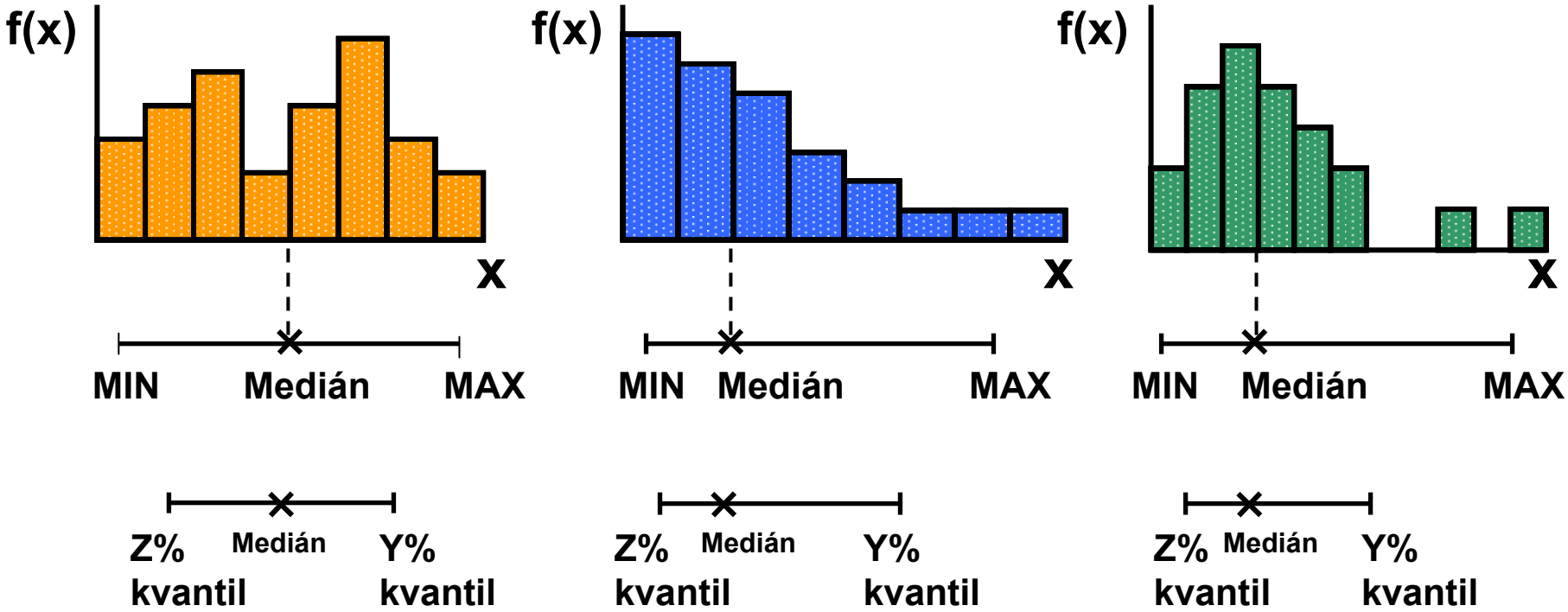
Komparace

B. Sumarizace odhadem „zástupců“ primárních dat



Odhad a jeho spolehlivost

Formální popis tvaru rozložení



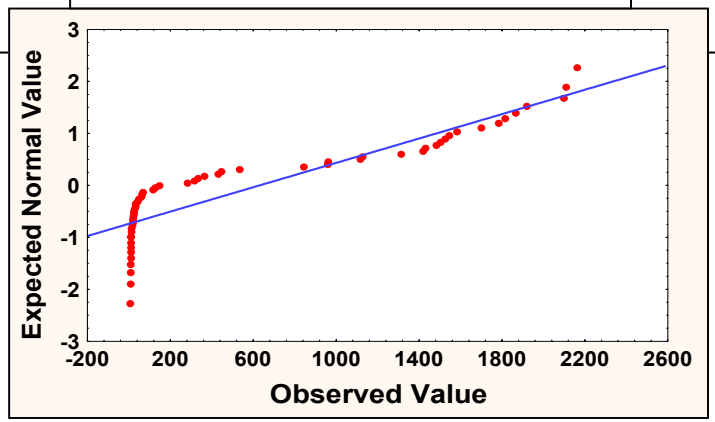
Medián = 50 % kvantil = frekvenční střed

MAX - MIN = rozsah (range)

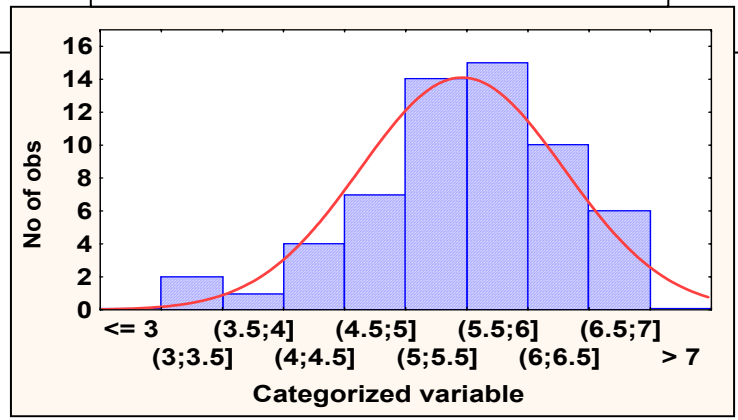
Modus = nejčastější hodnota

Testy o rozložení, grafický průzkum rozložení

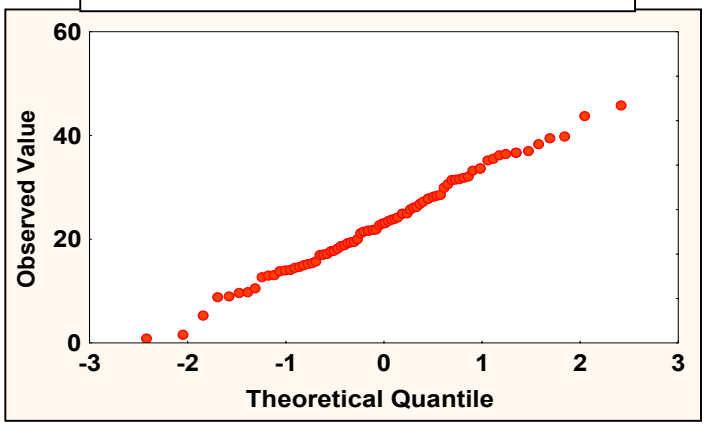
Normal probability plot



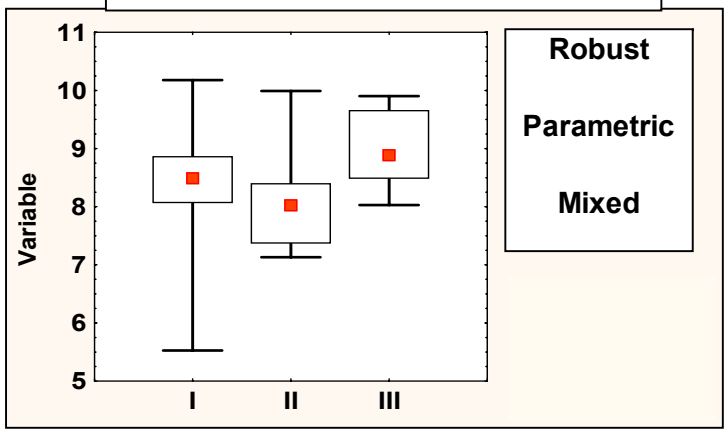
Histogram



Quantile - Quantile plot



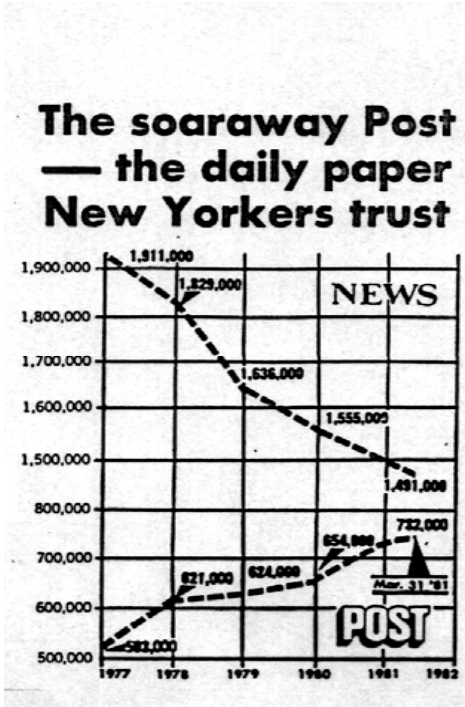
Multiple BW plots

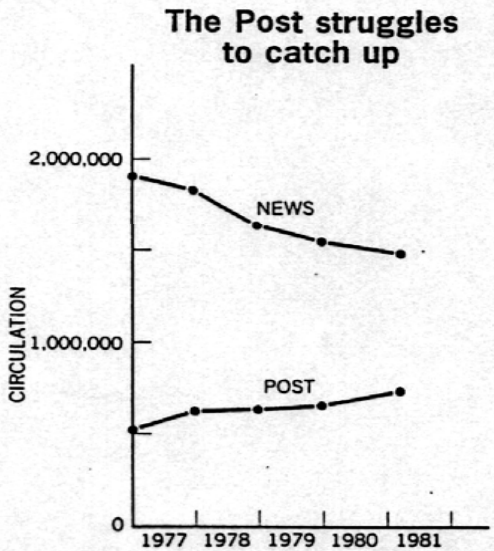


Testy o rozložení: Kolmogorov-Smirnov test, Shapiro-Wilks test, c2 test

Přehlednost a zviditelnění dat je základním stavebním kamenem analýz

Možný problém





!