Statistické zpracování dat

Metodologie sociálních výzkumů

Úvod do statistiky

Pojem statistika, vymezení:

- Statistika - věda
- Statistika - statisticky vyjádřené šetření
- Dvě definice:
  - **Statistika** je obor, který se zabývá shromažďováním, analýzou a interpretací údajů získaných z pozorování, experimentů a šetření.
  - **Statistika** je věda, která se zabývá získáváním informací z numerických dat.

Pojetí statistiky

- původní význam slova statistika souvisí se státem, s jeho administrativním spravováním (zaznamenávání údajů k vojenským a daňovým účelům)
- Statistika vychází jako matematická věda především z počtu pravděpodobnosti a teorie her.
- Studuje převážně hromadné jevy.

Statistická jednotka

Jednotlivý člen populace (základního souboru) se nazývá **statistická jednotka** (základní prvek, na kterém pozorujeme projev určité hromadné události). Statistickou jednotkou může být osoba, firma, prodejna, území, věc, událost apod.

Znaky

Vlastnosti, sledované na prvcích (jednotkách statistického šetření) výběru či populace, nazýváme znaky (veličiny) nebo proměnné. Proměnné nabývají hodnoty.

Příklad:

statistická jednotka	proměnná	hodnota proměnné
učitel	škola, roky praxe	ZŠ Ostrava, 25 let
žák	pohlaví, věk (roky)	chlapec, 14 let
vedomostní test	skóre	21 bodů

Druhy znaků (proměnných)

Rozlišujeme:

- kvalitativní, jsou-li varianty zkoumané vlastnosti dány slovním vyjádřením,
- kvantitativní, jsou-li varianty vyjádřeny číslem. Člení se dále na:
  - spojité - jednotlivé varianty znaku mohou nabývat jakékoliv hodnoty z určitého intervalu nebo rozmezí (výška, hmotnost apod.)
  - diskrétní (nespojité) - varianty znaku jsou vyjádřeny oddělenými čísly (počet onemocnění, počet zemřelých apod.)

Úrovně měření proměnných (stupnice, škály)

1. **Nominální** (kategorická)
2. **Pořadová** (ordinální)
3. **Intervalová** (kardinální)
4. **Poměrová** (absolutní)

Úrovně měření (škálování) proměnných jsou seřazeny od nejnižší k nejvyšší úrovni. Intervalová a poměrová stupnice představují metrickou úroveň.

Statistické metody zpracování hodnot proměnné se uplatňují v návaznosti na úroveň měření proměnné.

Nižší úrovně měření

Nominální stupnice - Objekt je zařazen vždy do jedné z více možných vzájemně se vylučujících kategorií (tříd, skupin). Například: národnost, krevní skupina, místo bydliště, škola, rodinný stav, pohlaví.

- Dichotomická (binární) proměnná - je specifický případ nominální proměnné, kde objekt patří vždy pouze do jedné ze dvou možných kategorií. Například: pohlaví, patří/nepatří do skupiny, časový okamih před akcí/po akci.

Pořadová stupnice - hodnoty znaků jsou seřazeny do pořadí, přičemž čísla jim přiřazená odrážejí toto uspořádání, ale neposkytují žádnou informaci o vzdálenosti mezi nimi, např. bolest hlavy, invalidita, spokojenost se službami, hodnocení prospěchu žáků učitelem.

Vyšší úrovně měření

Intervalová stupnice - Umožňují seřazení objektů, ale i jejich kvantifikaci a porovnání velikosti rozdílů mezi nimi. Intervalová proměnná musí mít určenou jednotku měření a v ní musí být vyjádřeny všechny její hodnoty. Nulová pozice je věcí volby, např. teplota ve stupních Celsia. Příklady intervalové škály: teplota (°C), výsledky z testů (počet bodů).

Poměrová stupnice - nulová pozice je pevně dána a vyjadřuje naprostou nepřítomnost měřené vlastnosti. Příklady poměrové škály: hmotnost v kg, výška v cm, měření času.

Členění statistiky

- **popisná**
  - základní charakteristika získaných dat
- **induktivní (testovací)**
  - charakterizace určitého výběru (vzorku populace), ze které usuzujeme na vlastnosti celého základního souboru

Popisná statistika

Popisná (deskriptivní) statistika se zabývá uspořádáním souborů, jejich popisem a účelnou sumarizací.

- Účelem je:
  - zpřehlednit získaná data (tabulky, grafické znázornění),
  - charakterizovat výběry pomocí kvantitativních charakteristik (míry středu, míry variability, míry tvaru).

Induktivní statistika

- umožňuje z pozorovaných dat vytvářet obecné závěry s udáním stupně jejich spolehlivosti,
- výpočet stupně spolehlivosti závěrů je objektivní, neboť je založen na poznatcích teorie pravděpodobnosti a nezávisí na subjektivním názoru hodnotitele.

Příklad využití induktivní statistiky

- Při sledování účinku léku na hodnotu krevního tlaku u nemocných s hypertenzní chorobou nelze přeměřit všechny nemocné v naší populaci,
- místo celého souboru nemocných se vyšetří jen určitý vzorek nemocných s hypertenzní chorobou (výběr),
- z pozorování se snažíme odvodit závěry pro všechny pacienty trpící hypertenzní chorobou (populaci), kteří jsou použitým vzorkem reprezentováni.

Deskriptivní (popisná) statistika

Jednorozměrná distribuce

Účel deskriptivní statistiky:

- Zpřehlednit získaná data (tabulky, grafické prostředky)
- Charakterizovat výběry pomocí kvantitativních charakteristik (míry středu, míry variability, míry tvaru)

Jednorozměrná distribuce – zjišťování rozdělení (distribuce) hodnot jednotlivé proměnné, označuje se i jako třídění prvního stupně

Rozdělení (distribuce) četností

- tabulka rozdělení četností - ukazuje, kolikrát byly pozorovány jednotlivé hodnoty či kolik pozorování padlo do určitých intervalů
- Obsahuje naměřené hodnoty znaku (proměnné), dále relativní četnosti a kumulativní četnosti.

Rozdělení četností – příklad (kvantitativní znak)

- U 70 žen byl změřen hemoglobin s přesností 0,1 g/100 ml (minimální a maximální hodnota je označena *):

10.2, 13.7, 10.4, 14.9, 11.5, 12.0, 11.0, 13.3, 12.9, 12.1, 9.4, 13.2, 10.8, 11.7, 10.6, 10.5, 13.7, 11.8, 14.1, 10.3, 13.6, 12.1, 12.9, 11.4, 12.7, 10.6, 11.4, 11.9, 9.3, 13.5, 14.6, 11.2, 11.7, 10.9, 10.4, 12.0, 12.9, 11.1, *8.8, 10.2, 11.6, 12.5, 13.4, 12.1, 10.9, 11.3, 14.7, 10.8, 13.3, 11.9, 11.4, 12.5, 13.0, 11.6, 13.1, 9.7, 11.2, *15.0, 10.7, 12.9, 13.4, 12.3, 11.0, 14.6, 11.1, 13.5, 10.9, 13.1, 11.8, 12.2

Tabulka rozdělení četností

Hladina hemoglobinu v g/100 ml	Počet	Relativní četnost v %	Kumulativní rel. četnost v %
8,0-8,9	1	1,4	1,4
9,0-9,9	3	4,3	5,7
10,0-10,9	14	20,0	25,7
11,0-11,9	19	27,1	52,9
12,0-12,9	14	20,0	72,9
13,0-13,9	13	18,6	91,4
14,0-14,9	5	7,1	98,6
15,0-15,9	1	1,4	100,0
Celkem	70	100,0	-

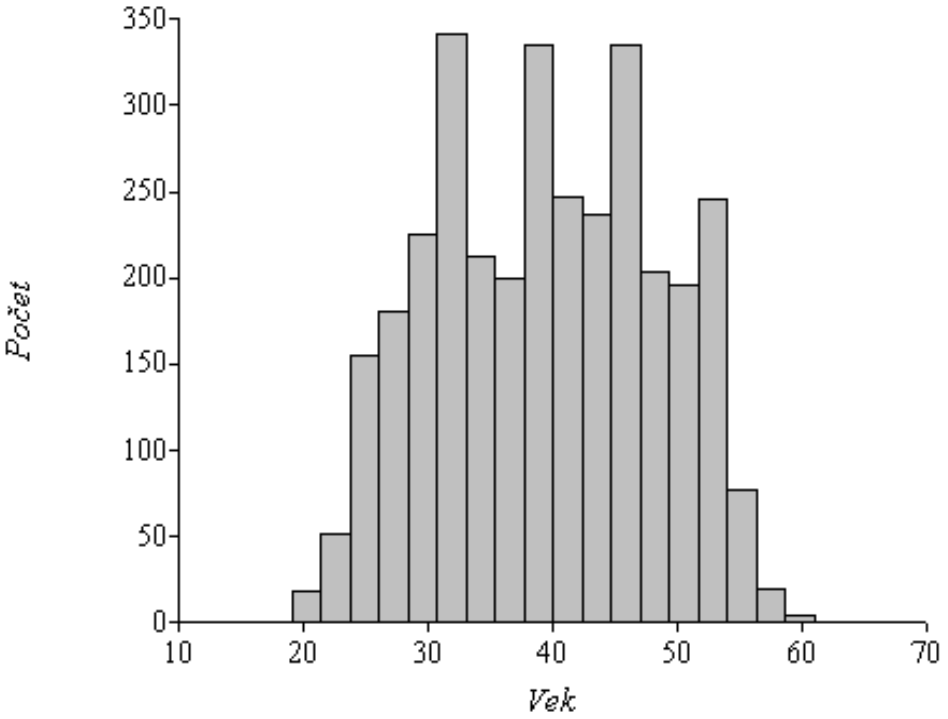
Histogram

- histogram se používá ke znázornění rozdělení absolutních nebo relativních četností - sloupce (obdélníky) jsou vždy vertikální a jejich výška odpovídá četnosti (absolutní nebo relativní)

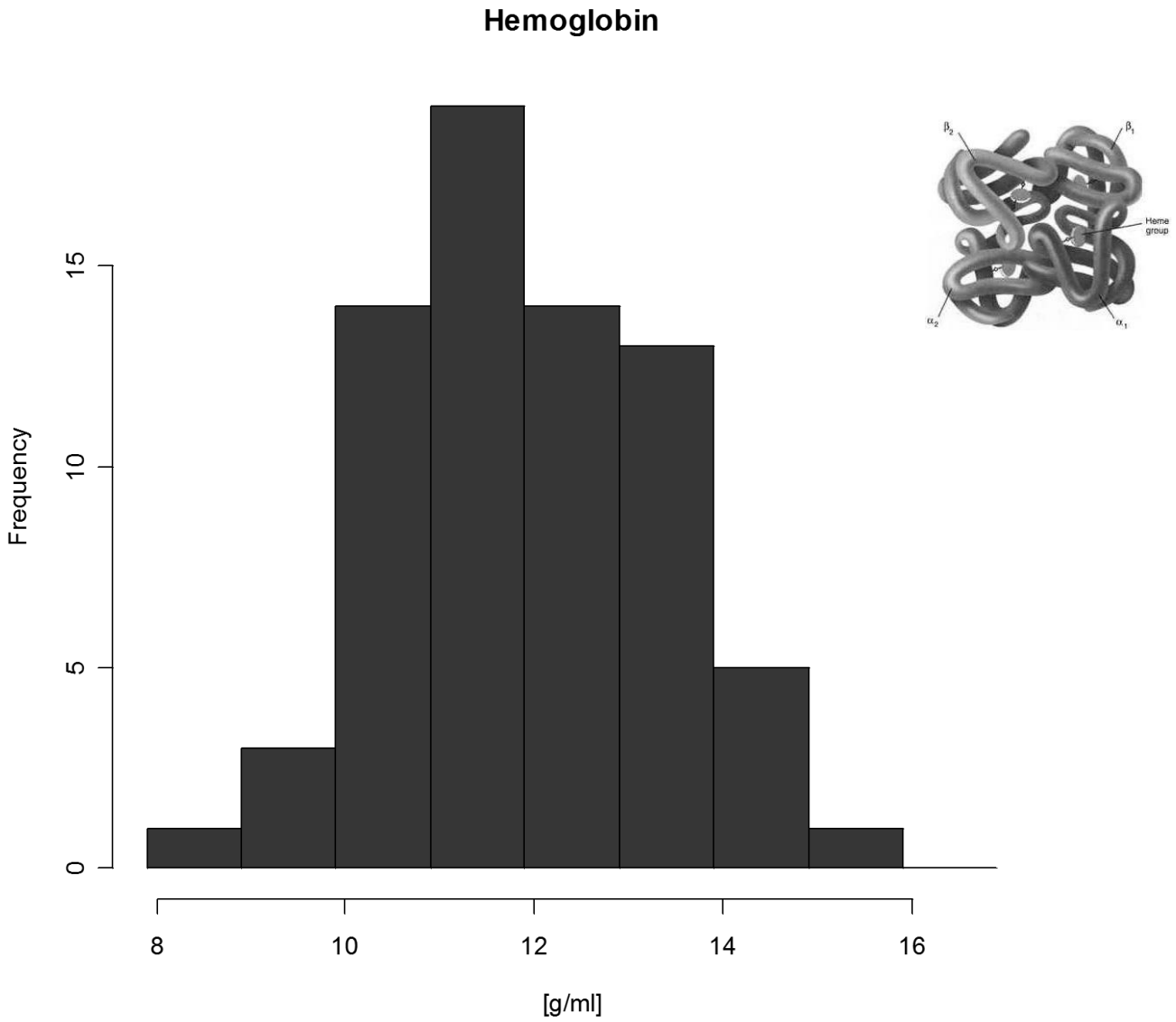
Histogram

Je typ sloupcového grafu určený pro znázornění rozdělení intervalové proměnné.

Graf vyjadřuje četnosti hodnot proměnné.

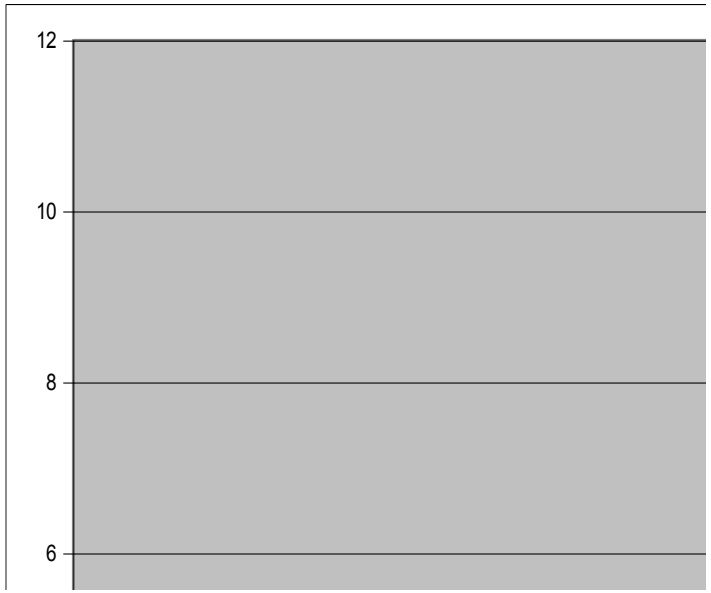


Histogram



Zobrazení dat

- Sběrná tabulka, tabulka četností,
- histogram četností

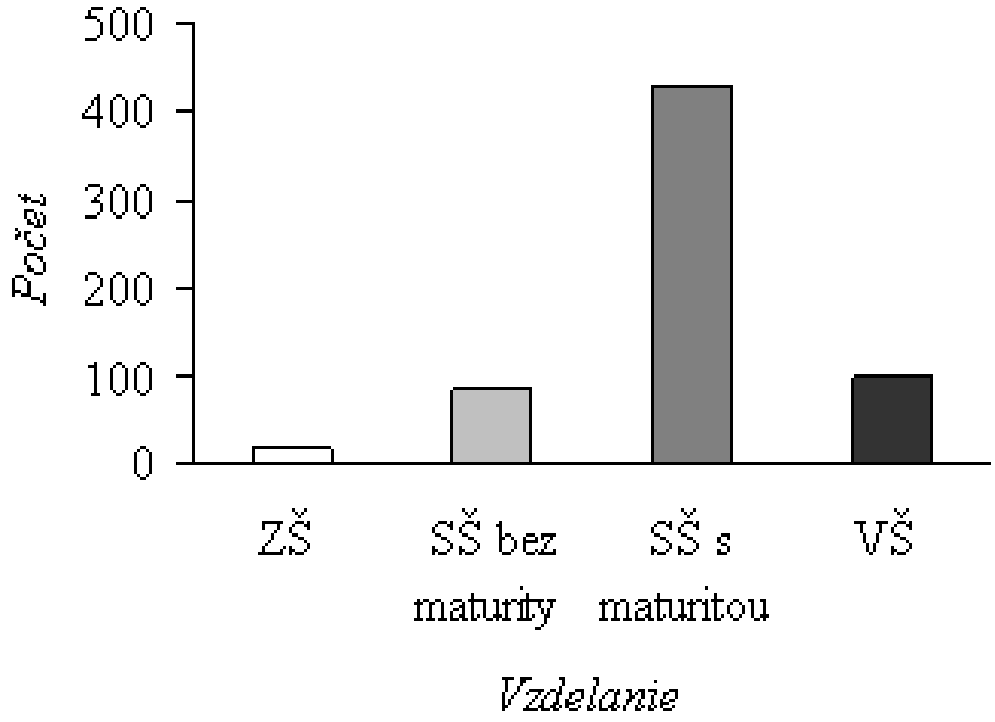
originální data	setříděná data	histogram
115	<100: 0	
135	100-110: 1	
120	111-120: 0	
140	121-130: 2	
125	131-140: 4	
130	141-150: 8	
150	151-160: 4	
145	161-170: 11	
.	>171: 0	
.		
.		

Tabulka četností – příklad rozdělení kategorické (kvalitativní) proměnné

vzdělání	četnost	procento
ZŠ	17	2,7
SŠ bez maturity	83	13,3
SŠ s maturitou	428	68,4
VŠ	98	15,6

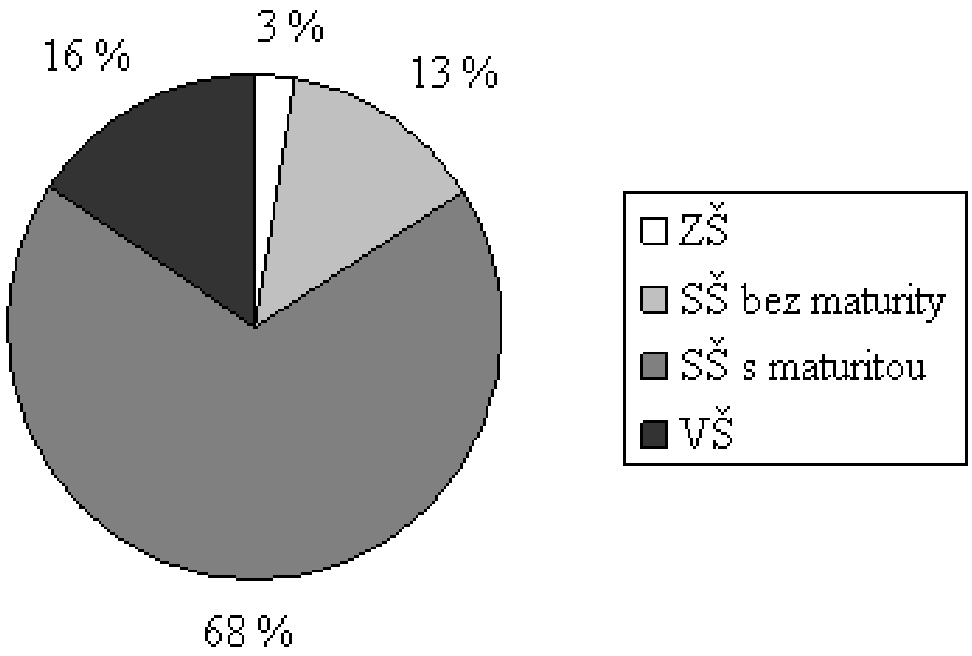
Sloupcový graf

Grafické zobrazení absolutních četností frekvenční tabulky.



Koláčový graf

Grafické znázornění procent (relativních četností) frekvenční tabulky.



Numerické (číselné) charakteristiky

1. **Míry polohy** (aritmetický průměr, medián, modus)
2. **Míry variability** (variační rozpětí, interkvartilové rozpětí, rozptyl, standardní odchylka, variační koeficient)
3. **Míry tvaru** (šikmost, špičatost)

Míry polohy

- měří polohu statistického souboru na ose x a mají stejný rozměr jako samotná pozorování:
  - průměr – „těžiště“ (citlivý na odlehlá pozorování)
  - medián – „prostřední“ hodnota uspořádaného souboru (má smysl jen pro kvantitativní a ordinální veličiny.
  - modus – nejčetnější hodnota (důležitý pro kvalitativní, zejména nominální znaky)

Míry polohy

- **Aritmetický průměr** – najznámější střední hodnota, počítá se jako součet všech hodnot dělený jejich počtem

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

Míry polohy

- **Medián** představuje střední hodnotu souboru, který je seřazen od nejmenší po největší hodnotu. V případě sudého počtu hodnot je medián aritmetický průměr hodnot na místech n/2 a n/2+1.
- Kromě mediánu se stanovují i **kvartily** (dělí soubor na 4 části) a **percentily** (dělí soubor na 100 částí). Medián je druhý kvartil resp. padesátý percentil.
- **Modus** představuje nejčastěji se vyskytující hodnotu proměnné.

Míry polohy (míry středu, střední hodnoty)

Použití vzhledem k úrovni měření:

- nominální znaky → **modus**
- ordinální znaky → **medián**
- intervalové znaky → **aritmetický průměr**

Míry variability

- Variabilita – proměnlivost hodnot znaku (proměnné) v statistickém souboru

- rozpětí - rozdíl mezi nejvyšší a nejnižší hodnotou v datech (závisí na extrémních hodnotách)
- kvartilové rozpětí – rozdíl 3. a 1. kvartilu
- rozptyl (variance) - průměrná kvadratická odchylka od průměru
- směrodatná odchylka (standardní odchylka) – odmocnina z rozptylu

Míry variability

- **Variační rozpětí** je jednoduchou mírou variability.

  Vypočítá se jako rozdíl mezi největší a nejmenší hodnotou v souboru.

  $R = x(max) – x(min)$

- **Kvartilové rozpětí** – představuje rozdíl mezi třetím a prvním kvartilem. Reprezentuje oblast středních 50 procent hodnot proměnné.

Míry variability

- **Rozptyl** (s^2) je často využívanou mírou variability. Rovná se průměrnému čtverci odchylek hodnot od průměru.
- Čím je rozptyl větší, tím více se údaje odchylují od průměru
- Počítá se:

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

Míry variability

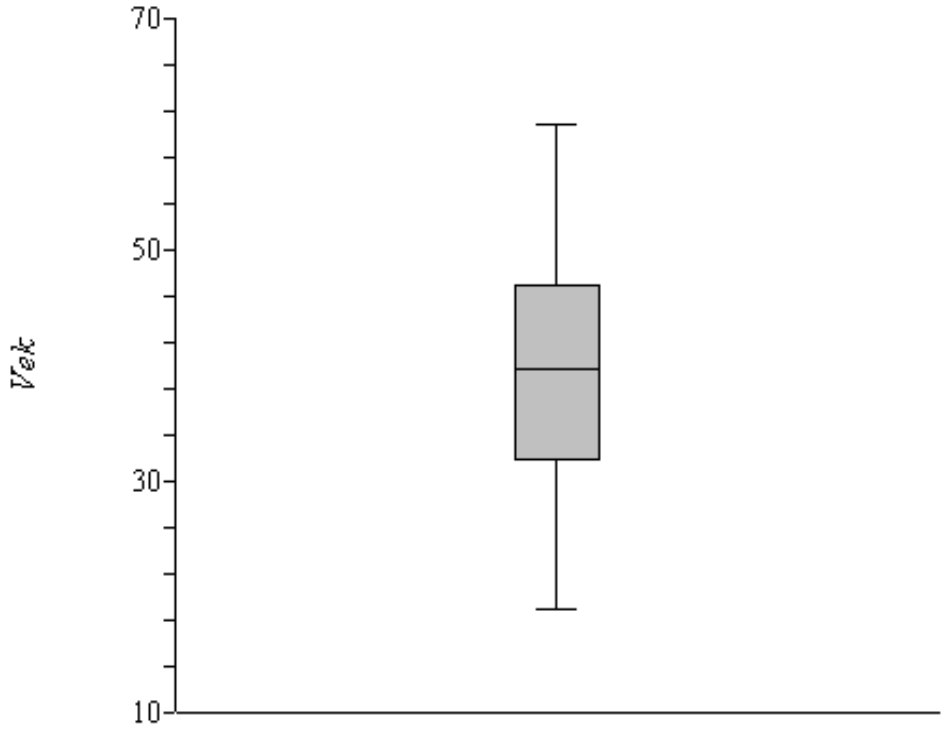
- **Směrodatná (standardní) odchylka** (s, SD) je odmocněný rozptyl, čím se odstraňuje vliv umocňování při výpočtu hodnoty rozptylu.
- Je to průměrný rozdíl mezi hodnotami a průměrem při ignorovaní znamének.
- Počítá se jako odmocnina rozptylu:

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

Krabicový graf

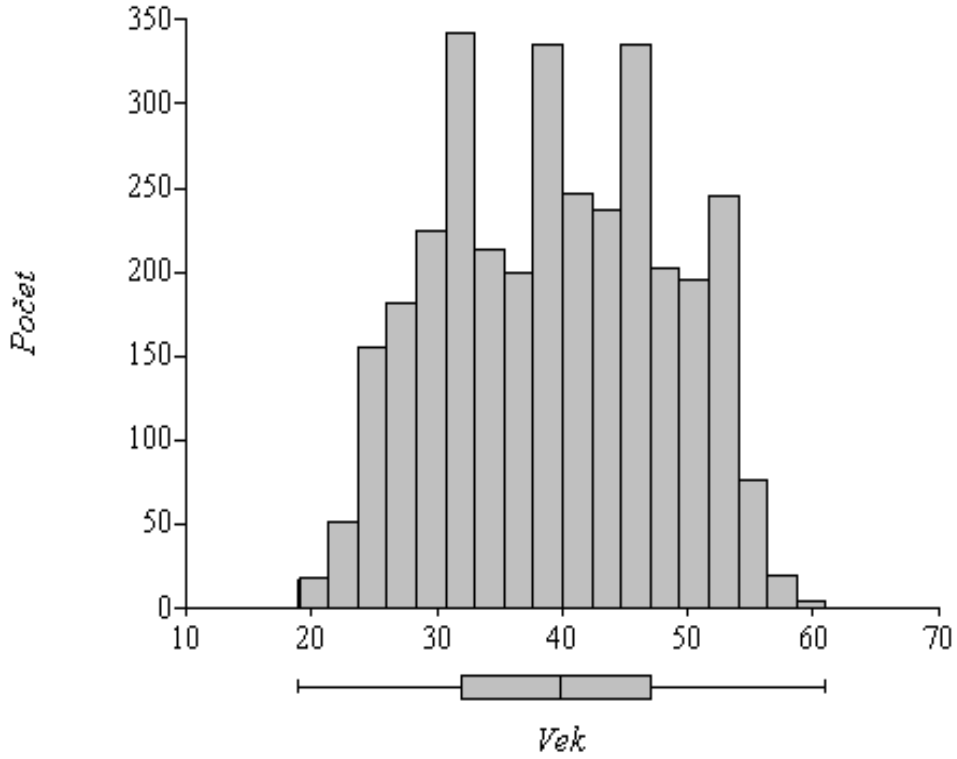
Horizontální čára představuje medián, horní hrana krabice 75. percentil a dolní hrana 25. percentil. Délka obdélníka představuje středních 50% hodnot souboru.

Dolní anténa – minimální hodnota, horní anténa – maximální hodnota.



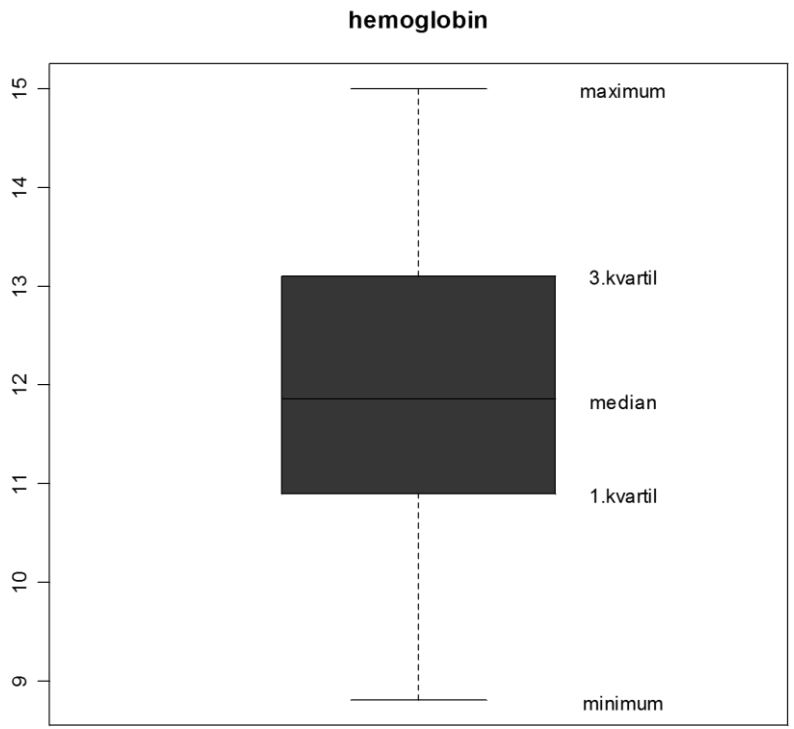
Histogram + krabicový graf

Tímto znázorněním se zvyšuje množství informací, které graf obsahuje.



Krabicový graf (box plot)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max
8.80	10.93	11.85	11.98	13.08	15.00

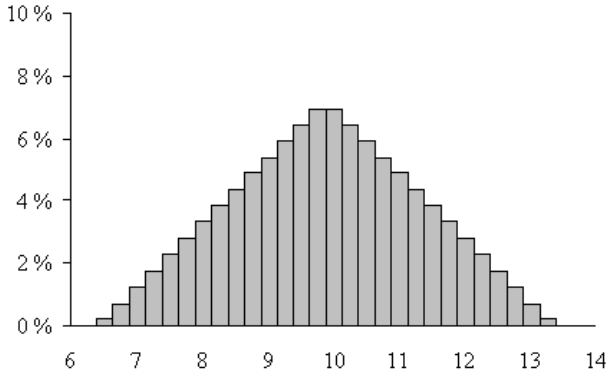


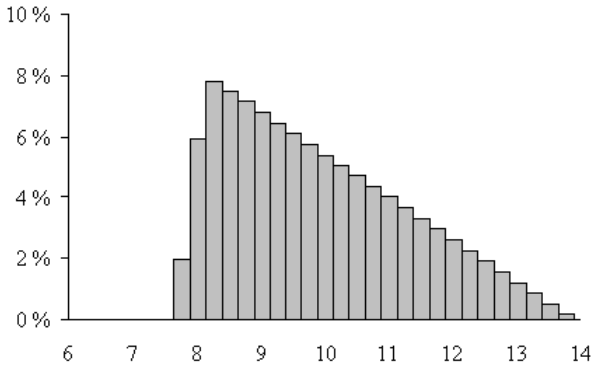
Stem-and-leaf plot

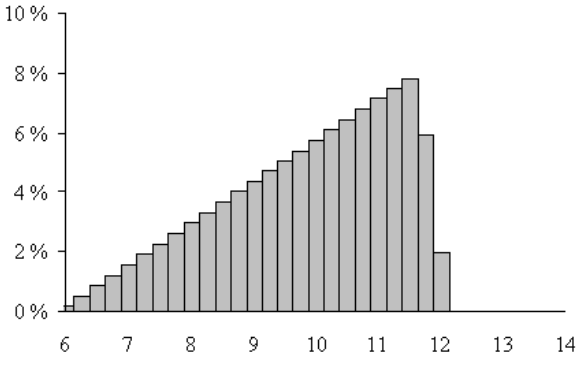
```
 8 | 8
 9 | 347
10 | 22344566788999
11 | 00112234456677889 9
12 | 00111235579999
13 | 0112334455677
14 | 16679
15 | 0
```

Míry tvaru

- **Šikmost** – měří směr a stupeň asymetrie rozdělení proměnné.
- Kladná (pravostranná šikmost) hodnota znamená, že většina hodnot je menší jako průměr, záporná hodnota (levostranná šikmost) znamená, že většina hodnot je větší jako průměr.

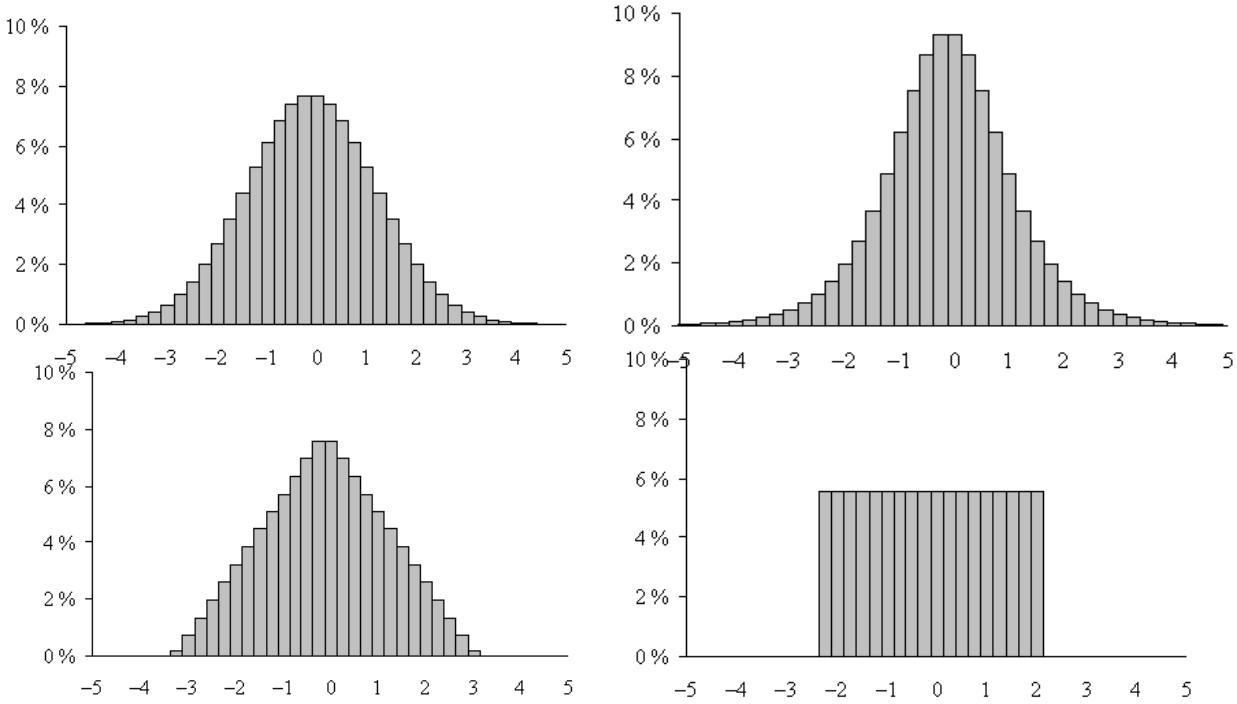






Míry tvaru

- **Špičatost** – měří hustotu chvostů rozdělení proměnné, t.j. charakterizuje výskyt extrémně vysokých a extrémně nízkých hodnot



Koeficient šikmosti (S_1) – je mírou asymetrie, bezrozměrné číslo, vyhodnocuje se :

$S_1 = 0$ symetrické rozdělení

$S_1 > 0$ pozitivní (pravostranná) asymetrie

$S_1 < 0$ negativní (levostranná) asymetrie

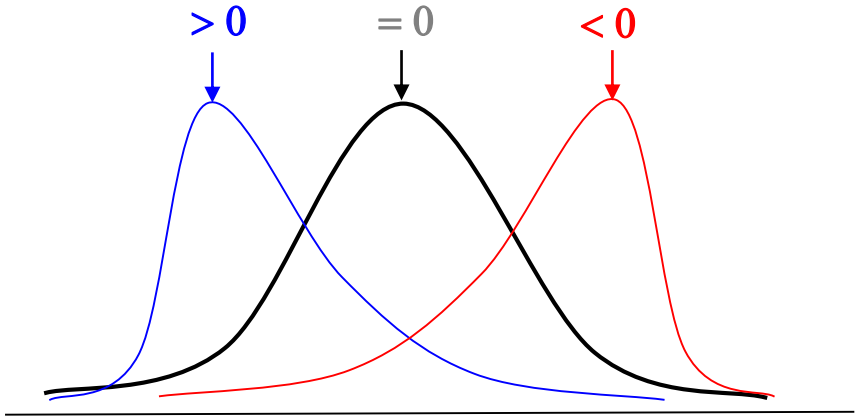
Koeficient špičatosti (S_2) – je mírou strmosti, bezrozměrné číslo

$S_2 = 0$ normální rozdělení

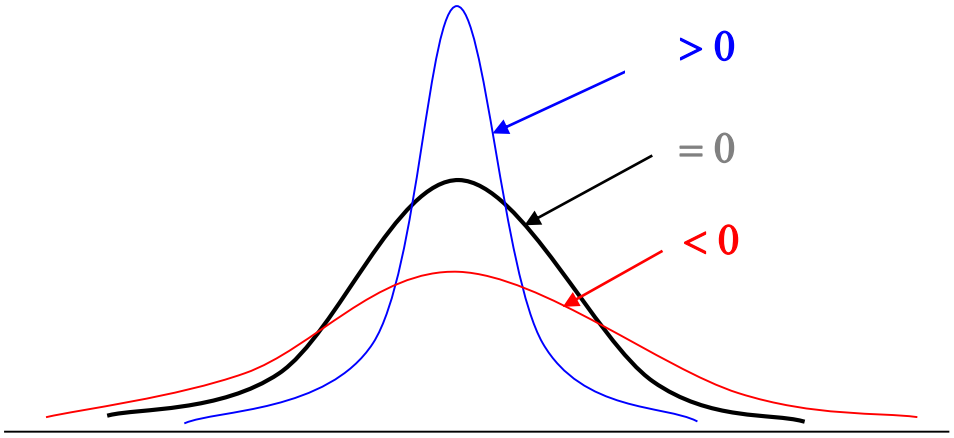
$S_2 < 0$ plochší rozdělení

$S_2 > 0$ špičatější rozdělení

Koeficient šikmosti

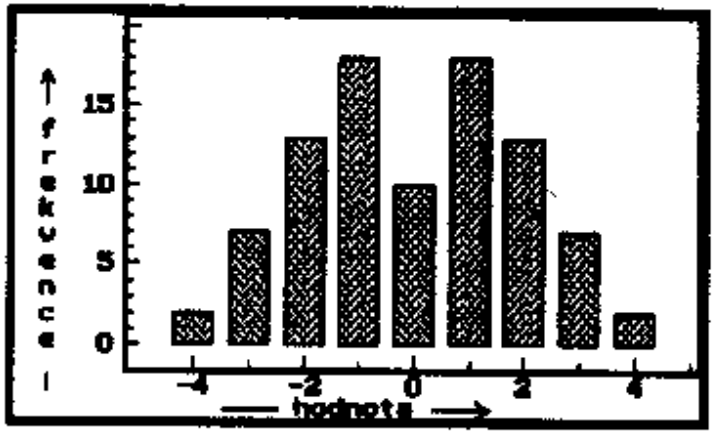


Koeficient špičatosti

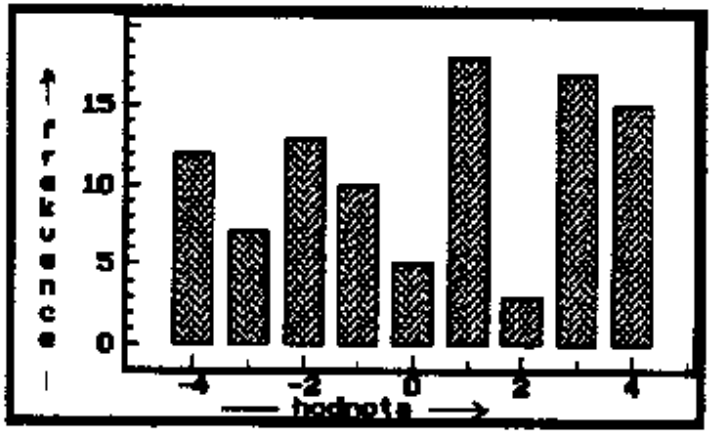


Symetrie, variabilita

symetrie

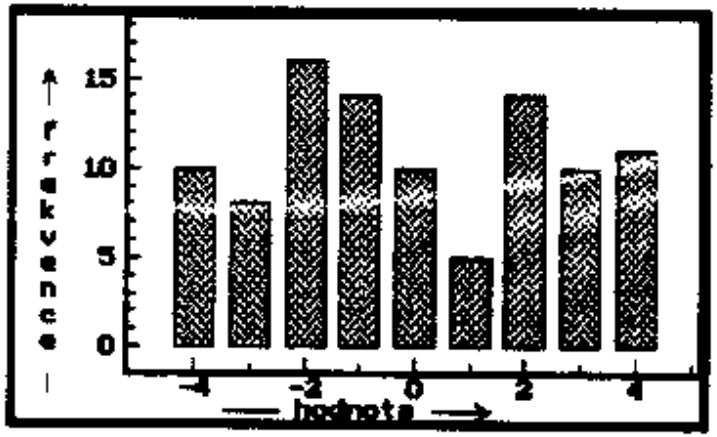


Obr I.A.4a Symetrické rozdělení

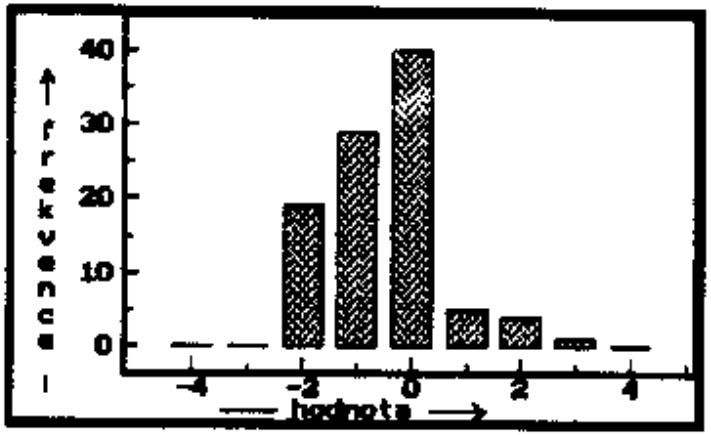


Obr. I.A.4b Asymetrické rozdělení

variabilita



Obr. I.A.4c Velká variabilita

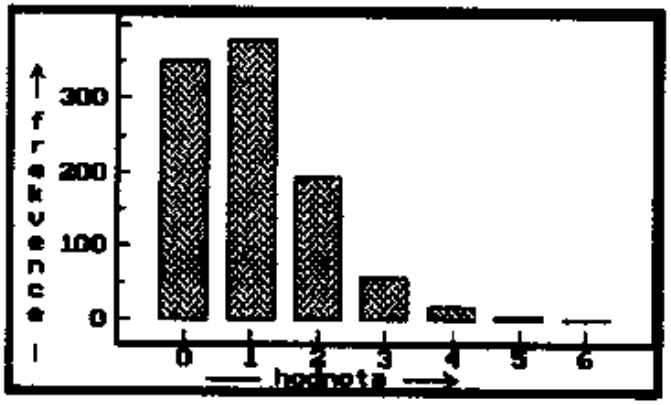


Obr. I.A.4d Malá variabilita

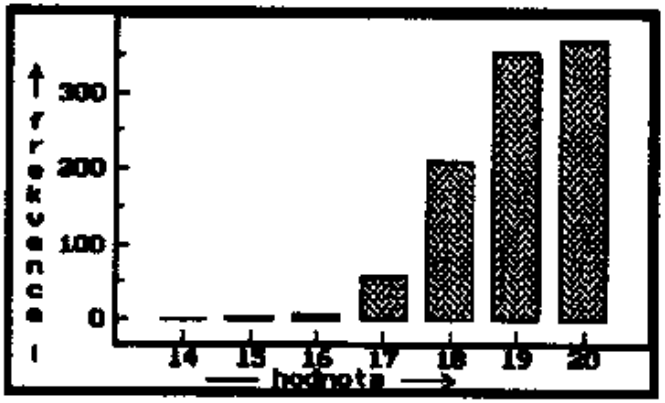
[Hanousek, Charamza 1992: 21]

Šikmost a špičatost

šikmost

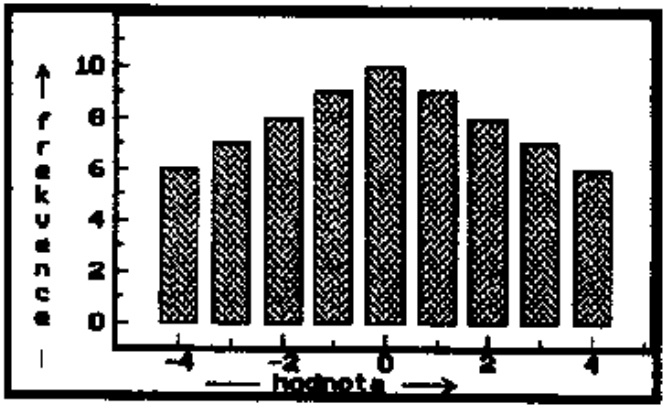


Obr. I.A.5a Kladně sešikmené

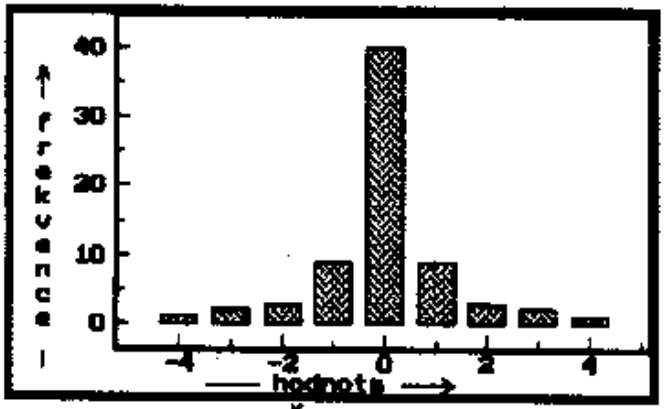


Obr. I.A.5b Záporně sešikmené

špičatost



Obr. I.A.5c Ploché rozdělení



Obr. I.A.5d Špičaté rozdělení

[Hanousek, Charamza 1992: 21]

Dvourozměrné distribuce

Charakteristika:

- Distribuce (rozdělení) dvou proměnných – třídění druhého stupně
- Zaměřujeme se na sledování vztahů dvou proměnných (např. vztah výšky a váhy, finančního ohodnocení a spokojenosti v práci)
- Tři typy vztahů: shoda, protiklad, nezávislost

Možnosti prezentace

- Tabulkové vyjádření – dvourozměrné tabulky (kontingenční tabulky), krostabulace
- Grafické znázornění – bodový graf
- Numerické charakteristiky – korelační a kontingenční koeficienty

Uspořádání dvourozměrné tabulky

ZÁVISLÁ - vysvětlovaná	NEZÁVISLÁ - vysvětlující			
		pohlaví		
	spokojenost	**muž**	**žena**	Celkový součet
	1 (nespokojen)	5 (41 %)	2 (22 %)	7
	2	5 (41 %)	1 (11 %)	6
	3 (spokojen)	2 (16 %)	6 (66 %)	8
	Celkový součet	12 (100 %)	9 (100 %)	21

Nejčastěji bývá **závislá** proměnná nalevo v řádcích a **nezávislá** (vysvětlující) ve sloupcích.

Druhy relativních četnosti

- Relativní **řádkové** četnosti = vypočteno vzhledem k součtu v příslušném řádku
- Relativní **sloupcové** četnosti = vypočteno vzhledem k součtu v příslušném sloupci
- Relativní **celkové** četnosti = vypočteno vzhledem k rozsahu celého souboru

Interpretace tabulek

závislá proměnná = je v hypotéze ovlivňována, způsobována (nejčastěji je v řádcích)

nezávislá(é) proměnná = vysvětluje, ovlivňuje závislou

V kategoriích nezávislé proměnné ukazujeme kompletní (100 %) distribuci závislé proměnné.

Pozor! **Směr kauzality je vždy věcí teorie, nelze ji určit z dat samotných.**

Souvislost znaků v tabulce

- Seskupení vysokých hodnot na diagonále tabulky naznačuje, že existuje souvislost mezi proměnnými.

- Souvislost ale může mít i jinou formu, např. v každém sloupci jsou pozorování nahromaděna do jediného pole, jehož pozice je pro každý sloupec jiná.

Zjišťování vztahů

- graficky (bodový graf)
- korelační koeficient
- lineární regrese

Korelace

Vztah (souvislost, asociace) mezi dvěma proměnnými, jejichž hodnoty jsou uspořádány ve dvojicích.

Sílu tohoto vztahu určuje korelační koeficient (Pearson, Spearman, Kendall).

Lineární i nelineární vztahy.

Není úplným popisem dat ani při velmi silném vztahu.

Neznamená sama o sobě příčinný (kauzální) vztah.

Umožňuje odvodit procento společné variability.

Korelace

- Máme dvě proměnné (označují se *X* a *Y*) - ptáme se, zda jsou nezávislé, nebo existuje mezi nimi souvislost (korelace) a jak je silná.

  - Součinový korelační koeficient (Pearson) – r
  - Pořadový korelační koeficient (Spearman) - r_s

Korelační koeficient

Nabývá hodnot od -1 do +1

- Když se r = 1, leží všechny body na přímce, a ta má rostoucí charakter.
- Když se r = -1, leží všechny body na přímce, a ta má klesající charakter.
- Když se r = 0, proměnné jsou nezávislé.
- Asociace je kladná, když r je větší než nula.
- Asociace je záporná, když r je menší než nula.

Korelační koeficient

- Korelační koeficient může mít kladnou nebo zápornou hodnotu. To je vyjádření pozitivního nebo negativního vztahu proměnných. Pokud s růstem hodnot první proměnné rostou i hodnoty druhé proměnné mluvíme o pozitivní souvislosti (např. vztah úsilí a výkonu). Pokud naopak s růstem hodnot proměnné „x“ hodnoty proměnné „y“ klesají mluvíme o negativní souvislosti (např. vztah únavy a výkonu).
- Hodnoty korelačního koeficientu od 0,1 do 0,3 znamenají slabý vztah proměnných. Hodnoty nad 0,3 po 0,6 středně silný vztah. Hodnoty nad 0,6 silný vztah dvou proměnných. Uvedené platí jak pro kladné tak i pro záporné hodnoty korelací.

Koeficient determinace

Druhá mocnina korelačního koeficientu

Udává procento společné variability sledovaných proměnných, tedy kolik procent variability proměnné Y

může být vysvětleno variabilitou proměnné X a naopak.

Druhy korelačních koeficientů

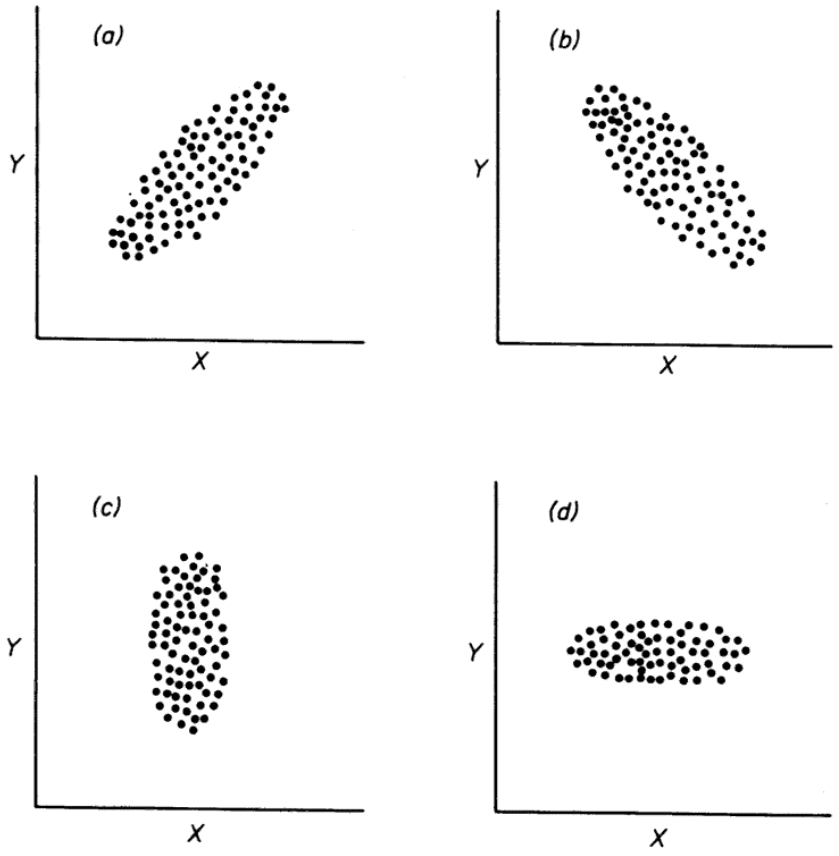
Pearsonův k.k. (r):

- Mírou síly vztahu dvou spojitých proměnných
- Vyjadřuje pouze sílu lineárního vztahu
- Ovlivněn odlehlými hodnotami

Spearmanův k.k. (rho, rs):

- Koreluje se pořadí hodnot jednotlivých proměnných
- Zachycuje nejen lineární, ale obecně rostoucí nebo klesající vztahy
- Resistentní vůči odlehlým hodnotám.

Grafické znázornění korelace dvou proměnných – bodový graf



Jednoduchá lineární korelace: positivní korelace (a), negativní korelace (b), bez korelace (c a d)