# Metodologie pro Informační studia a knihovnictví 2

## Modul 5: Popis nekategorizovaných dat

### Co se dozvíte v tomto modulu?

- Kdy používat modus, průměr a medián.
- Co je to směrodatná odchylka.
- Jak popsat distribuci dat.
- Jak zobrazovat spojité proměnné.

### Obsah


Nekategorizované proměnné ..... 2
Aritmetický průměr ..... 2
Minimum, maximum a rozpětí ..... 3
Rozptyl a směrodatná odchylka ..... 4
Percentily ..... 6
Zobrazování kardinálních dat ..... 8

## Nekategorizované proměnné

**Nekategorizované proměnné** jsou ty proměnné, které mohou nabývat všech hodnot z daného intervalu. Může jednat o plat, věk, počet obyvatel města, délku pracovní zkušenosti v měsících…

## Aritmetický průměr

**Aritmetický průměr** je třetí mírou centrální tendence. U kardinálních dat lze jako míry centrální tendence využívat všechny tři:

- modus,
- medián,
- aritmetický průměr.

Aritmetický průměr je ukazatelem „průměrné" hodnoty, nemusí být ale vždy ukazatelem nejvhodnějším – vhodné je jej kombinovat s mediánem. Aritmetický průměr je totiž velmi citlivý na extrémní hodnoty. I jedna extrémní hodnota může výrazně posunout aritmetický průměr.

*Příklad: V roce 2010 byl podle serveru Platy.cz průměrný měsíční plat 23 300 Kč. Medián byl však na hodnotě 21 000 Kč. Znamená to, že průměr vychýlil menší počet jedinců s výrazně vyšším platem.*

| Průměrný měsíční plat (v Kč) | Medián (Kč) | Rozdíl (v %) |
|---|---|---|
| 23 300 | 21 000 | 11% |

Zdroj: Platy.cz

Pro připomenutí:


**Modus** se používá, pokud:

- rozdělení má více vrcholů,
- chceme zjistit nejčastější hodnoty.

**Medián** používáme, pokud:

- jsou data ordinální nebo kardinální,
- chceme znát střed rozložení dat,
- (v kombinaci s průměrem) pokud soubor obsahuje extrémní hodnoty,
- jestliže je rozložení dat zešikmené.

**Aritmetický průměr** je vhodné používat, pokud

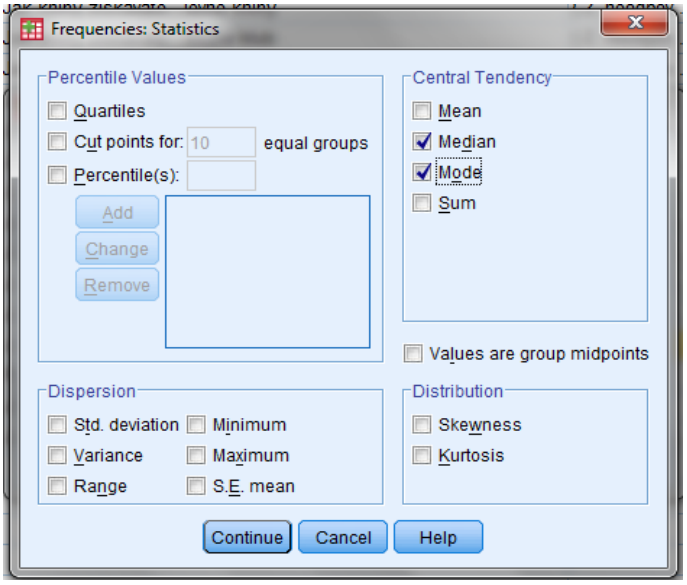
- jsou data kardinální,
- rozložení je symetrické,
- chceme použít statistické testy. (Hendl 2009)

### *Aritmetický průměr v Excelu*

- Příkaz **PRŮMĚR**

### *Aritmetický průměr v SPSS*

Pro zjištění hodnot měr centrální tendence v SPSS zadáte Analyze → Descriptive Statistics → Frequencies → **Statistics → Mean, Median, Mode**




## Minimum, maximum a rozpětí

První charakteristiky nekategorizovaných dat, na které se díváme už při fázi čištění dat, jsou **minimální** a **maximální hodnoty**. Z nich také snadno spočítáme **rozpětí**.

**Rozpětí** je nejjednodušší míra variability a snadno se vypočítá jako rozdíl mezi nejvyšší a nejnižší hodnotou.

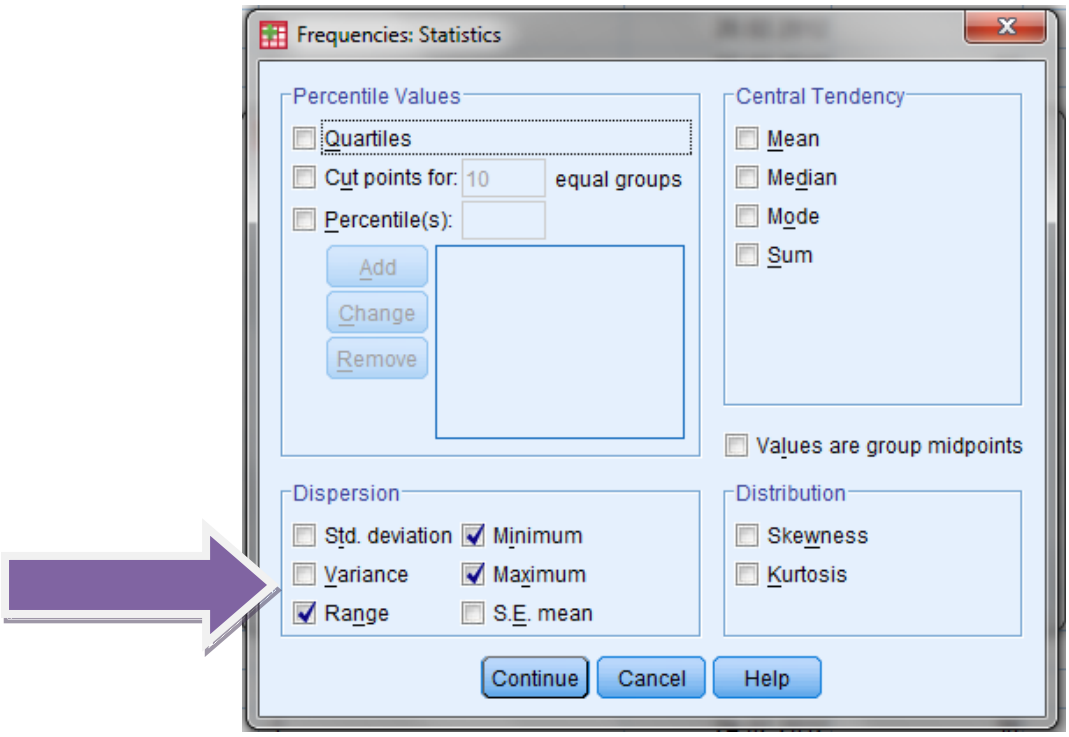
> *Např. Je-li minimální hodnota 18 a maximální 1024, rozpětí hodnot proměnné v souboru je 106.*

### *Minimum, maximum a rozpětí v Excelu*

- Příkaz **MIN(oblast hodnot)**
- Příkaz **MAX(oblast hodnot)**
- Rozpětí jako rozdíl hodnot MAX a MIN

### *Minimum, maximum a rozpětí v SPSS*

Vypočítání rozpětí můžete v SPSS zadat tímto řetězcem: **Analyze – Frequencies – Statistics:**




## Rozptyl a směrodatná odchylka

**Rozptyl** je definován jako střední hodnota kvadrátů odchylek od střední hodnoty (průměru). Vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem její střední hodnoty. Při průměrování odchylek dělíme číslem n-1.

S rozptylem úzce souvisí **směrodatná odchylka**. Ta se vypočítá jako odmocnina z rozptylu. Vrací tedy míru rozptýlenosti do měřítka původních dat. V podstatě nám říká, uvnitř jakého intervalu okolo průměru leží zvolené procento případů – tedy čím je směrodatná odchylka menší, tím lépe pro aritmetický průměr.

Hendl (2009) srozumitelně vysvětluje, jak dochází k výpočtu směrodatné odchylky:

1. Nejprve si vypočítáme všechny odchylky od průměru (např. při hodu kostkou vždy spočítáme odchylku konkrétní hozené hodnoty od celkového průměru).
2. Umocnění na druhou převede záporné odchylky na kladná čísla. Zároveň zvýrazní váhu extrémnějších odchylek.
3. Sečteme kvadratických odchylek.
4. Dělením číslem n-1 získáme průměrnou kvadratickou odchylku.
5. Odmocnina (v případě směrodatné odchylky) převede výsledek do původního měřítka dat.

Pro názornost si pojďme ukázat příklad, který dobře znáte – hodnocení vyučujících na KISKu a směrodatnou odchylku tohoto hodnocení.

```
Zajímavost předmětu              není vůbec zajímavý .***X(*)**... je velmi zajímavý
Přínosnost předmětu              není vůbec přínosné ***X*(*)*.... je velmi přínosné
Obtížnost obsahu                       velmi snadný .....(.)**X** velmi obtížný
Náročnost na přípravu                  velmi snadný .....(*)X*... velmi obtížný
Dostupnost studijních zdrojů  velmi špatně dostupné .....(.).**X* velmi dobře dostupné
Jak učitel učí                         velmi špatný .***X(*)**... vynikající
Učitel jako odborník                 není odborníkem .....(.)***X* je odborníkem
```

```
Zajímavost předmětu              není vůbec zajímavý .....(.)...*X je velmi zajímavý
Přínosnost předmětu              není vůbec přínosné .....(.)..*X* je velmi přínosné
Obtížnost obsahu                       velmi snadný **X**(.)..... velmi obtížný
Náročnost na přípravu                  velmi snadný *X**.(.)..... velmi obtížný
Dostupnost studijních zdrojů  velmi špatně dostupné .....(.)**X** velmi dobře dostupné
Jak učitel učí                         velmi špatný .....(.)...*X vynikající
Učitel jako odborník                 není odborníkem .....(.)....X je odborníkem
```

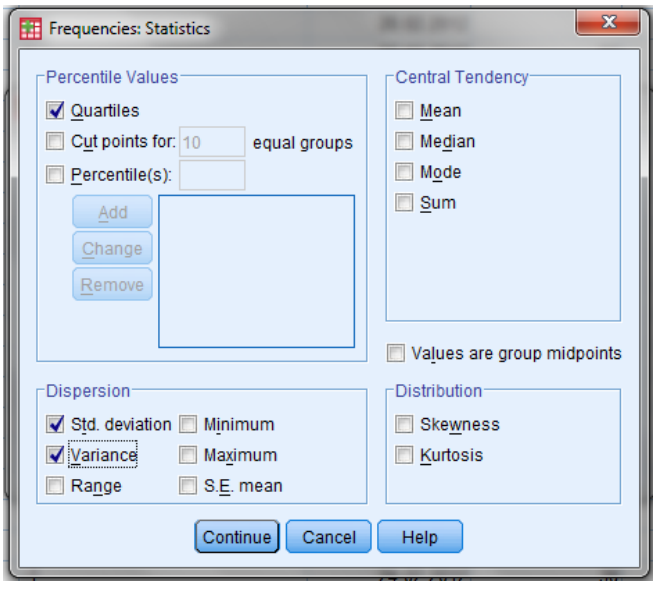
*Průměrné hodnocení proměnné „Učitel jako odborník" je u obou vyučujících podobné – jeden vyučující má průměrné hodnocení 9, druhý má průměrné hodnocení 10. Směrodatná odchylka (zvýrazněná hvězdičkami) nám ale poskytne rychlou další informaci – říká nám, jak moc se hodnocení všech respondentů pohybovalo kolem průměru. Vidíme, že zatímco v druhém případě se hodnocení výjimečně shodovalo a studující se shodli na tom, že učitel je skutečný odborník, v prvním případě nebyla shoda zdaleka tak veliká.*

### *Rozptyl a směrodatná odchylka v Excelu*

- rozptyl – příkaz **VAR**
- směrodatná odchylka – příkaz **SMODCH.VÝBĚR**

### *Rozptyl a směrodatná odchylka v SPSS*

Vypočítání rozptylu a směrodatné odchylky můžete v SPSS zadat tímto řetězcem: **Analyze – Frequencies – Statistics:**



## Percentily

Percentil x je hodnota, pro kterou platí, že x procent případů má hodnotu menší nebo rovnu percentilu x.

Nejčastěji se využívají:

- **MEDIÁN** (x50)
- **KVARTILY** (x25, x50, x75)
- **DECILY** (x10, x20, x30, x40, x50, x60, x70, x80, x90)

*Například vás může zajímat, jak jsou rozloženy příjmy obyvatel v horním a spodním percentilu. Tato informace spolu s mediánem ukazuje, jak moc jsou rozevřené pomyslné nůžky mezi „horní" a „spodní" vrstvou společnosti.*

**Jak vysoký je medián proti průměrné mzdě? (ve vybraných zemích OECD)**

| Země | spodních 10 % | medián | horních 10 % |
|---|---|---|---|
| Švédsko | 56 % | 89,8 % | 150,9 % |
| Finsko | 62,3 % | 89,5 % | 147,9 % |
| Kanada | 44,6 % | 89,1 % | 166,9 % |
| Dánsko | 60,9 % | 89 % | 150,4 % |
| Norsko | 63,2 % | 88,9 % | 149 % |
| Japonsko | 52,4 % | 87,6 % | 162,7 % |
| Nový Zéland | 51,2 % | 87,2 % | 160,6 % |
| Německo | 43,4 % | 87 % | 165,7 % |
| **Česko** | **49,3 %** | **85,2 %** | **153,1 %** |
| Itálie | 56,1 % | 85,1 % | 156,6 % |
| Švýcarsko | 56,6 % | 84,9 % | 153,4 % |
| Belgie | 60,4 % | 84,5 % | 153,4 % |
| Nizozemí | 51,7 % | 84 % | 158,8 % |

*Zdroj: [http://finexpert.e15.cz/jak-se-lisi-prumerna-mzda-a-median](http://finexpert.e15.cz/jak-se-lisi-prumerna-mzda-a-median)*

### *Percentil v Excelu*

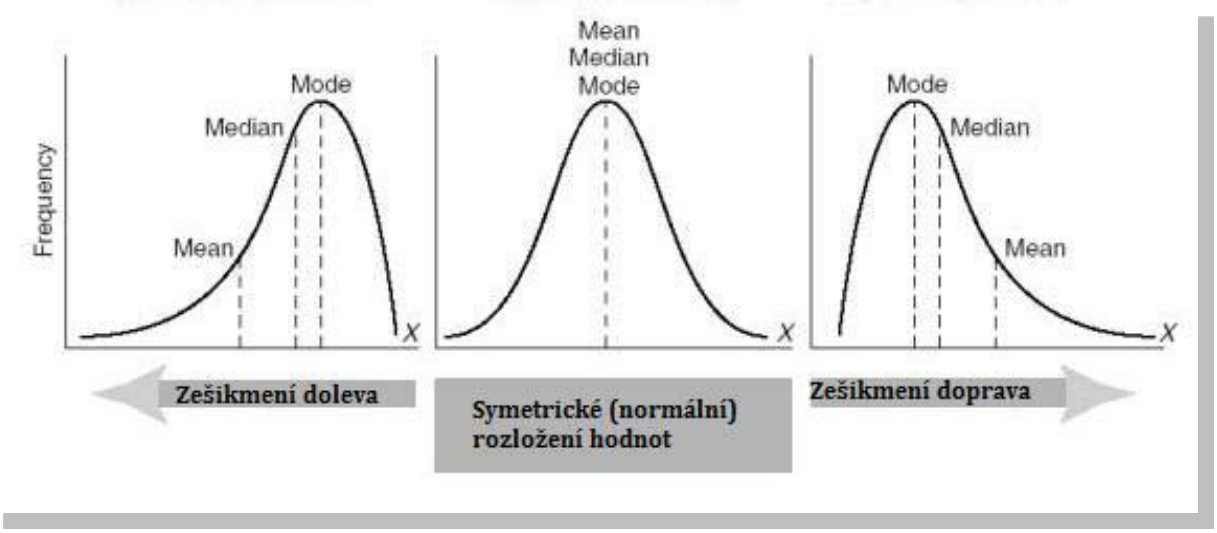
- Příkaz **PERCENTIL** (rozpětí dat;hodnota percentilu z intervalu 0-1)
  *Tedy např. percentil 50 můžeme zapsat jako =PERCENTIL(A1:A30;0,5)*

### *Percentil v SPSS*

Vypočítání rozptylu a směrodatné odchylky můžete v SPSS opět zadat tímto řetězcem: **Analyze – Frequencies – Statistics (políčko Percentile Values)**.

## Šikmost a špičatost

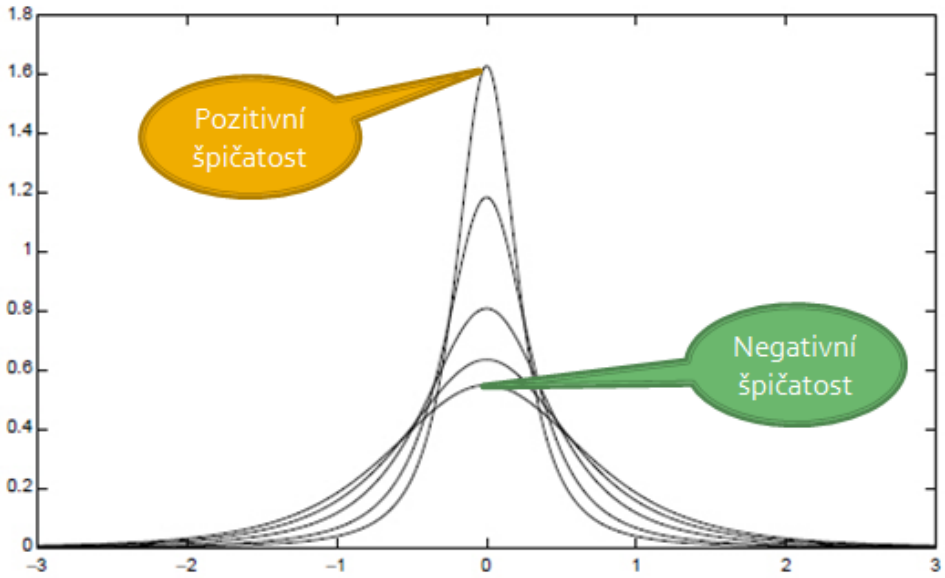
Spojitá data nabývají málokdy tzv. normálního rozložení. Při popisu dat si všímáme zešikmení a špičatosti dat.




Ad **šikmost**:

- **Symetrické (normální) rozložení** - aritmetický průměr, medián a modus mají stejné nebo velmi podobné hodnoty. (0)
- Pokud je aritmetický průměr větší než medián, který je zase větší než modus, znamená to, že je více případů menších než průměr a naše **rozložení je šikmé doprava**. (+)
- Třetí možností je, že je více případů větších než aritmetický průměr. Ten je pak menší než medián a ten je menší než modus. Naše **rozložení je šikmé doleva**. (-)

**Špičatost** zase udává, jak moc jsou data nakumulována v oblasti středních hodnot.



### *Šikmost a špičatost v Excelu*

- příkaz **SKEW** (šikmost)
- příkaz **KURT** (špičatost)

### *Šikmost a špičatost v SPSS*

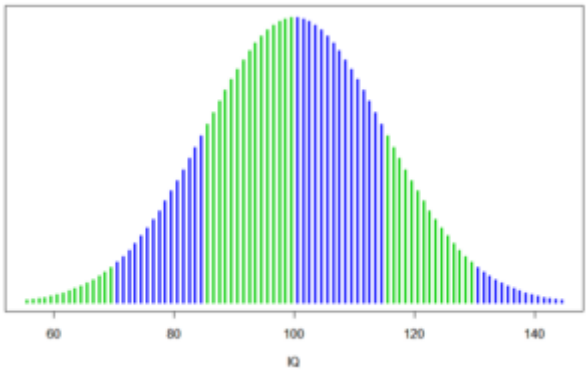
**Analyze – Frequencies – Statistics** (políčko Distribution).

## Zobrazování kardinálních dat

Pro zobrazování kardinálních dat se používá několik možných grafů

## *Histogram*

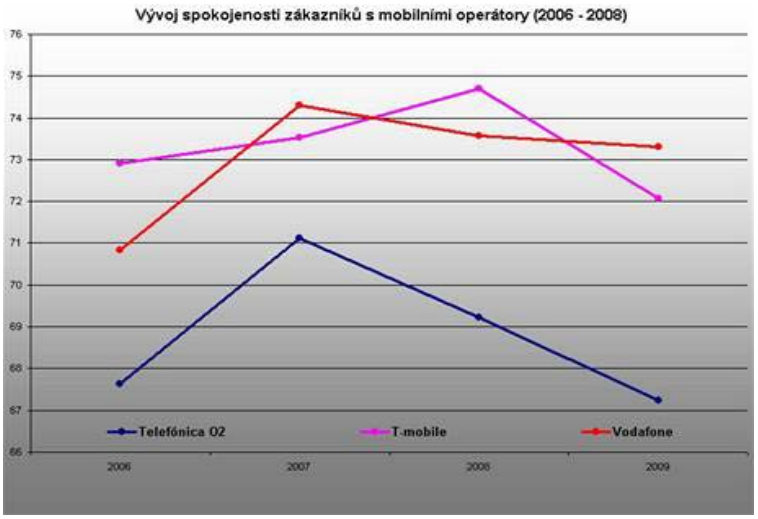
Histogram je podobný sloupcovému grafu, mezi jednotlivými sloupci ale nejsou mezery. Pracujete-li v Excelu, můžete využít klasický sloupcový graf.




*Příklad histogramu – distribuce IQ v populaci (zdroj: IQscope.com)*

## *Spojnicové grafy*

Chcete-li ukázat, jak se hodnoty proměnné měnily v čase, je vhodné použít spojnicový graf.




*Příklad využité spojnicového grafu – spokojenost s mobilními operátory 2006-2008*

## *Bodové grafy*

Bodové grafy zachycují jednotlivé hodnoty proměnných a využívají se v třídění druhého stupně jako zachycení toho, jak jedna proměnná ovlivňuje druhou (o tomto grafu více v dalších modulech).

## Literatura

Hendl, J. *Přehled statistických metod analýzy dat*. Praha : Portál 2009

Levine, D. M., & Stephan, D. (2010). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. Upper Saddle River, N.J: FT Press.