**Užití Excelu ve statistice**

**(průvodce studiem)**

**Karel Pastor**

**Cíle:** Cílem studijního textu (respektive chystané přednášky) je ukázat některé základní možnosti využití Excelu při statistickém zpracování souboru dat. Po prostudování by se měl student dobře orientovat v základních pojmech *popisné statistiky* jako je například aritmetický průměr, odchylka, rozptyl, šikmost, špičatost, normální rozdělení a umět používat Excel k jejich výpočtu.

Druhým hlavním cílem je naučit se testovat *statistické hypotézy* a používat přitom Excel.

**Průvodce textem:** První část textu je věnována popisné statistice. Na příkladech budou vysvětleny jednotlivé charakteristiky, kterými můžeme soubor dat popsat. Ukážeme, pomocí kterých příkazů v Excelu je možné charakteristiky určit.

Ve druhé části texu se budeme věnovat statistickým hypotézám a jejich testování. Ukážeme si, které statistické testy je možné použít na typické situace, jež mohou v pedagogické praxi nastat. Pomáhat nám přitom bude opět Excel.

Budeme vycházet z posledních dvou verzí Excelu (2013 a 2016) a upozorňovat na případné rozdíly.

# Popisná statistika

Popisná statistika zkoumá *statistický soubor*, tj. množinu určitých dat, a stanovuje pro něj statistické znaky (charakteristiky) jako například aritmetický průměr, podle kterých se můžeme orientovat. Rovněž se zabývá tím, jak získaná data znázornit.





**Příklad 1:** Žáci jedné páté třídy psali test z matematiky. Nejvyšší počet bodů byl 100. Jednotlivým žákům se podařilo získat tyto body:

5, 6, 21, 39, 42, 15, 3, 38, 91, 26, 17, 4, 2, 13, 19, 24, 56, 78, 14, 25, 39, 64, 1, 16.

Na tomto příkladu si vysvětlíme jednotlivé pojmy popisné statistiky.

## Histogram a výsečové grafy

Často je vhodné a výhodné uspořádat získaná data do tříd (intervalů). K určení „optimálního“ počtu tříd můžeme využít například Sturgersovo pravidlo:

$$ZAOKROUHLIT(3,3\*LOG(24);0)+1,$$

kde 24je počet údajů v souboru a funkce (v Excelu 2013) $ ZAOKROUHLIT$ zaokrouhluje číslo $3,3\*LOG\left(24\right) $na 0 desetinných míst, tedy na celé číslo. Poznamenejme ještě, že funkce $LOG$ vrací dekadický logaritmus.

V našem příkladu dostaneme 6 tříd. Nyní můžeme v záložce DATA využít Analýzu dat (pokud chybí, můžeme ji přidat přes Možnosti a Doplňky) a vybrat Analytický nástroj *Histogram*. Jako Vstupní oblast zadáme buňky, do kterých jsme umístili body z testu (například $A$1:$F$4 znamená, že body jsou umístěny v buňkách A1-A4, B1-B4, C1-C4, D1-D4, E1-E4, F1-F4).

Dále potřebujeme doplnit na vstupu Hranice tříd. Zjistíme maximum bodů (příkazem *MAX*(A1:F4)) a minimum bodů (příkazem *MIN*(A1:F4)). Rozdíl těchto hodnot (91-1) vydělíme počtem tříd a máme tak šířku třídy 15. Můžeme nyní stanovit horní hranice tříd jako 15, 30, 45, 60, 75, 91 a poté již můžeme nechat vykreslit histogram.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| I:ikona6e |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Předchozí analytický nástroj *Histogram* nám umožňuje navíc zjistit četnosti bodů z testu zastoupené v jednotlivých třídách:

|  |  |
| --- | --- |
| *Třídy* | *Četnost* |
| 15 | 9 |
| 30 | 7 |
| 45 | 4 |
| 60 | 1 |
| 75 | 1 |
| 91 | 2 |

Získané četnosti můžeme použít při excelovské konstrukci výsečových (koláčových) graf, ať už dvourozměrného nebo prostorového:

## Charakteristiky polohy

### Aritmetický průměr $\overline{x}$:

### $$\overline{x}=\frac{1}{n}\sum\_{i=1}^{n}x\_{i}$$

V Excelu použijeme pro Příklad 1 příkaz *PRŮMĚR*(A1:F4) a dostaneme 27,41667.

* Modus $\hat{x}$ :

nejčastější hodnota.

V Excelu použijeme pro Příklad 1 příkaz *MODE.MULT*(A1:F4) a dostaneme 39, což je jediná bodová hodnota, kterou získali z testu dva žáci.

### Medián $\tilde{x}:$

při lichém počtu hodnot v souboru za mediánbereme prostřední hodnotu souboru;

při sudém počtu hodnot za medián bereme aritmetický průměr největší hodnoty dolní poloviny a nejmenší hodnoty horní poloviny.

V Excelu použijeme pro Příklad 1 příkaz *MEDIAN*(A1:F4) a vyjde nám 20.

### Dolní kvartil $Q\_{1}$:

Dolní kvartil odděluje dolní čtvrtinu naměřených hodnot (uspořádaných podle velikosti). Hodnota dolního kvartilu může (ale nemusí) odpovídat některé z naměřených hodnot.

V Excelu použijeme pro Příklad 1 příkaz *QUARTIL.EXC*(A1:F4;1), kde 1 udává první (dolní) kvartil a vyjde nám 7,75.

### Horní kvartil $Q\_{3}$:

Horní kvartil odděluje horní čtvrtinu naměřených hodnot (uspořádaných podle velikosti). Hodnota horního kvartilu může (ale nemusí) odpovídat některé z naměřených hodnot.

V Excelu použijeme pro Příklad 1 příkaz *QUARTIL.EXC*(A1:F4;3), kde 3 udává třetí (horní) kvartil a vyjde nám 39.

## Charakteristiky variability

## Rozptyl:

$$σ^{2}=\frac{1}{n}\sum\_{i=1}^{n}(x\_{i}-\overline{x})^{2}$$

V Excelu použijeme pro Příklad 1 příkaz *VAR.P*(A1:F4) a dostaneme hodnotu 571,3264.

## Směrodatná odchylka:

$$σ=\sqrt{σ^{2}}$$

 V Excelu použijeme pro Příklad 1 příkaz *SMODCH.P*(A1:F4) a dostaneme hodnotu 23,90243. Máme-li již k dispozici hodnotu rozptylu, můžeme také využít příkazu *ODMOCNINA*.

## Kvartilová odchylka:

$$Q=\frac{Q\_{3}-Q\_{1}}{2}$$

V případě symetrického rozdělení hodnot platí, že v intervalu $\left[\tilde{x}-Q,\tilde{x}+Q\right]$ leží přibližně 50 % všech hodnot.

V Excelu není příkaz pro kvartilovou odchylku, samozřejmě však můžeme použít Excel jako kalkulačku. V Příkladu 1 nám vyjde 15,625.

## Krabicový graf:

názorně zaznamenává pro daný statistický soubor minimální hodnotu, maximální hodnotu, medián, dolní a horní kvartil. Pro kvartilový graf se užívá také názvu krabicový (krabičkový) graf. Krabicový graf je možné vytvořit v Excelu 2016.



##  Šikmost:

charakteristika, která popisuje symetrii pravděpodobnostního rozdělení dat vzhledem k aritmetickému průměru.

Nulová šikmost značí, že hodnoty souboru jsou rovnoměrně rozděleny vlevo a vpravo od průměru.

Kladná šikmost značí, že vpravo (na číselné ose) od průměru se vyskytují odlehlejší hodnoty nežli vlevo a tím pádem se většina hodnot nachází vlevo od průměru.

U záporné šikmosti je tomu naopak.

V Excelu použijeme pro Příklad 1 příkaz *SKEW.P*(A1:F4) a dostaneme hodnotu 1,135002. Vyšlo nám kladné číslo, všimněme si na následujícím grafu, že skutečně většina hodnot se nachází vlevo od průměru (ten byl 27,41667) a vlevo na číselné ose znamená dole na svislé ose.

## Špičatost:

charakteristika, která porovnává dané rozdělení hodnot s normovaným normálním rozdělením, jehož grafem je tzv. Gaussova křivka:

Nulová špičatost značí normované normální rozdělení.

Kladná špičatost ukazuje na rozdělení špičatější než normální, tudíž má hodnoty více koncentrované k průměru.

Záporná špičatost je naopak plošší.

V Excelu použijeme pro Příklad 1 příkaz *KURT*(A1:F4) a dostaneme hodnotu 0,950127.

# Statistické hypotézy a testy

## Základní pojmy

*Statistické testy* jsou postupy sloužící k ověření hypotéz, zda mezi jevy je statisticky významný vztah.

*Nulová hypotéza* je statistické tvrzení, že mezi zkoumanými proměnnými není vztah.

Jestliže se při statistické analýze ukáže, že nulovou hypotézu je možné odmítnout, přijímáme tzv. *alternativní hypotézu*.

Riziko (pravděpodobnost), že neoprávněně odmítneme nulovou hypotézu (a tak nesprávně přijmeme alternativní), se nazývá *hladina významnosti*.

Pokud tedy budeme například něco tvrdit s 0% rizikem, pak se musí jednat o jistou věc. Pokud naopak budeme něco tvrdit se 100% rizikem, pak nám nemůže nikdo nic vytknout, ať už tvrdíme cokoliv.

## Test dobré shody chí-kvadrát

**Příklad 2:** 120 žáků jisté základní školy odpovídalo v dotazníku na otázku „Jakou čokoládu máš nejraději?“

1. Hořkou
2. Mléčnou
3. Bílou

Na základě testu dobré shody chí-kvadrát máme rozhodnout, zda mezi oblíbeností čokolád jsou statisticky významné rozdíly.

Odpovědi jsou patrné ve druhém sloupci následující tabulky.

|  |  |  |  |
| --- | --- | --- | --- |
| **Čokoláda** | **Zjištěné počty P** | **Očekávané počty O** | $$\frac{(P-O)^{2}}{O}$$ |
| HOŘKÁ | 42 | 40 | 0,1 |
| MLÉČNÁ | 50 | 40 | 2,5 |
| BÍLÁ | 28 | 40 | 3,6 |
|  | ∑120 | ∑120 | ∑6,2 |

Třetí sloupec obsahuje očekávané počty a konečně poslední sloupec udává postupně příspěvky k **testové hodnotě 6,2**.

Uvažujme nulovou a alternativní hypotézu:

$H\_{0}$ … počty žáků, kteří mají v oblibě jednotlivé čokolády, jsou statisticky stejné.



$H\_{1}$ … počty žáků, kteří mají v oblibě jednotlivé čokolády, jsou statisticky různé.

Nyní příkazem v Excelu *CHISQ.INV.RT*(0,05;2), kde 0,05 (=5%) je hladina významnosti (riziko) a 2 je počet stěžejních řádků v tabulce zmenšený o jedničku (tzv. stupně volnosti) dostaneme **kritickou hodnotu 5,99**.

Jelikož je tato hodnota menší než testová hodnota 6,2, můžeme s rizikem 5% odmítnout nulovou hypotézu a přijmout alternativní. Tedy rozdíl v oblibě čokolády je statisticky významný při hladině významnosti 5%.

## Test nezávislosti chí-kvadrát pro kontingenční tabulku

**Příklad 3:** 300 studentům pedagogické fakulty byl dán dotazník obsahující dvě otázky: „Chodíte pravidelně (tj. aspoň dvakrát za týden) běhat?“

1. ANO
2. NE

„Jaký byl Váš studijní průměr v posledním akademickém roce?“

1. lepší než 1,5
2. 1,5-2,2
3. horší než 2,2

Na základě testu dobré shody chí-kvadrát pro kontingenční tabulku máme rozhodnout, zda mezi běháním a studijními výsledky jsou statisticky významné rozdíly.

Uvažujme nulovou a alternativní hypotézu:

$H\_{0}$ … mezi odpověďmi na obě otázky není závislost.

$H\_{1}$ … mezi odpověďmi studentů je závislost.

Výsledky dotazníkového šetření zapíšeme do kontingenční tabulky. V posledním řádku a posledním sloupci jsou uvedeny kontrolní součty, tzv. marginální hodnoty. Čísla v závorce jsou očekávané hodnoty (násobíme odpovídající marginální hodnoty a dělíme celkovým počtem, například $ 22=\frac{55⋅120}{300}$).



|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **PRŮMĚRNÝ PROSPĚCH** |  |
|  |  | lepší než 1,5 | 1,5-2,2 | horší než 2,2 | ∑ |
| BĚHÁNÍ | ANO | 25 (22) | 75 (70) | 20 (28) | 120 |
| NE | 30 (33) | 100 (105) | 50 (42) | 180 |
|  | ∑ | 55 | 175 | 70 | 300 |

Testovou statistiku vypočítáme jako součet hodnot $\frac{(P-O)^{2}}{O}$, kde P jsou zjištěné (pozorované) počty (tedy postupně 25, 75, 20, 30, 100, 50) a O jsou očekávané hodnoty (tj. postupně 22, 70, 28, 33, 105, 42). Tudíž **testová statistika** je 0,41+0,35+2,29+0,27+0,24+1,52 = **5,08**.

Nyní příkazem v Excelu *CHISQ.INV.RT*(0,05;2), kde 0,05 (=5%) je hladina významnosti (riziko) a 2 (stupeň volnosti) je součin $(r-1)⋅(s-1)$, kde *r* je počet řádků v tabulce *s* je počet sloupců dostaneme **kritickou hodnotu 5,99**.

Jelikož 5,08 < 5,991, přijmeme nulovou hypotézu na hladině významnosti 0,05. Tedy můžeme říct, že mezi běháním a průměrným prospěchem není statisticky významný rozdíl při hladině významnosti 0,05.

## Další statistické testy

*Znaménkový* a *Wilcoxonův test* užíváme v případě dvou opakovaných měřeních na stejných objektech (lidech) při různých situacích. Pomocí znaménkového testu lze rozhodnout, zda mezi oběma opakovanými měřeními týchž objektů je nebo není významný rozdíl. Používáme při ordinálním měření, když je možné rozhodnout, která z opakovaně naměřených hodnot je větší.

*Studentův t-test* slouží k tomu, abychom mohli rozhodnout, zda dva soubory dat získané měřením ve dvou různých skupinách, mají statisticky stejný aritmetický průměr. Podrobnější informace o těchto i dalších testech je možné najít v následující literatuře.

**Literatura**: CHRÁSKA Miroslav: Metody pedagogického výzkumu, Grada Publishing, Havlíčkův Brod, 2016. KOPECKÝ MILAN: Úvod do počtu pravděpodobnosti a matematické statistiky, Univerzita Palackého, Olomouc, 2005. CALDA EMIL, DUPAČ VÁCLAV: Matematika pro gymnázia, Kombinatorika, pravděpodobnost, statistika, Prometheus, Praha 2003.

