Základní výpočty pro MPPZ

Teorie

Aritmetický průměr = součet hodnot znaku zjištěných u všech jednotek souboru, dělený počtem všech jednotek souboru
Modus = hodnota souboru s nejvyšší četností
Medián = prostřední hodnota souboru
Relativní četnost = procentuální podíl četnosti znaku na celkovém rozsahu souboru
Rozptyl = průměr druhých mocnin odchylek od aritmetického průměru
Směrodatná odchylka = druhá odmocnina z rozptylu
Variační koeficient = podíl směrodatné odchylky a aritmetického průměru, obvykle se vyjadřuje v procentech, má smysl tehdy, nabývá-li znak x jen nezáporných hodnot.
Koeficient korelace = vyjadřuje míru závislosti 2 znaků
Chí-kvadrát

Příklad 1:
V 9.B, kde je 20 žáků, se v hodině ČJ psal diktát. Výsledné známky byly tyto: 2, 2, 4, 3, 3, 1, 5, 5, 4, 2, 2, 2, 1, 3, 4, 3, 5, 3, 2, 1.
Vypočítejte: Aritmetický průměr (průměrnou známku), modus (nejčastější známku), medián, relativní četnost, rozptyl, směrodatnou odchylku, variační koeficient
Postup:
Krok 1 – Uděláme si první sloupce tabulky. Každý sloupec bude mít tolik řádek, kolik je v souboru různých hodnot (5) + 1 pro záhlaví. Zjistíme si absolutní četnosti hodnot, tzn. kolikrát je v daném souboru známka 1, kolikrát 2, atd...

Hodnota znaku (známka)	Četnost výskytu
1	3
2	6
3	5
4	3
5	3
	20

Kontrola: Sečteme-li sloupec „Četnost výskytu“ musí nám vyjít hodnota rozsahu souboru (našich 20 žáků)

Krok 2 – Vypočítáme si relativní četnost
- Jako vzor použijeme známku 3 – ta se ve známkování vyskytla 5x (absolutní počet). V kolika procentech případů dostal žák trojku? (relativní počet)

Počítáme trojčlenkou:

20 známek	je	100%
5 známek (trojek)	je	x %

x = (100 x 5) / 20 = 25 %

Výsledky promítneme do dalšího sloupce...

Pozn.: Ve správně napsané trojčlence vždy postupujeme takto: Hodnotu nad X vynásobíme tou nalevo od X a nakonec vydělíme tou poslední, která je křížem od X.

Hodnota znaku (známka)	Četnost výskytu	Relativní četnost (%)
1	3	15
2	6	30
3	5	25
4	3	15
5	3	15
		100

Kontrola: Sečteme-li sloupec „Relativní četnost“ musí nám vyjít 100% (přibližně – podle toho, jak drasticky jsme zaokrouhlovali)

Krok 3 – Vypočítáme aritmetický průměr

Zjednodušeně řečeno: sečteme všechny známky dohromady a pak vydělíme celkovým počtem známek – 20

- to lze u souborů s malým rozsahem hodnot, máme-li hodnot mraky (a nemáme-li excel :)) postupujeme takto (abychom minimalizovali možnost vzniku chyby při mačkání na kalkulačce):

Jak jinak, přidáme si další sloupec – Y (např. nevim, jak ho pojmenovat)

-Vypočítáme ho tak, že Hodnotu znaku (známku) vynásobíme Četností výskytu. Tedy: 1x3, 2x6, 3x5...

Hodnota znaku (známka)	Četnost výskytu	Relativní četnost (%)	Y
1	3	15	3
2	6	30	12
3	5	25	15
4	3	15	12
5	3	15	15
			57

Hodnoty ve sloupci Y sečteme a potom vydělíme celkovým počtem známek – 20, vyjde nám aritmetický průměr:

57:20 = 2,85

- Pro kontrolu použijme i ten jednoduchý způsob:

(1+1+1+2+2+3+3+3+3+3+4+4+4+5+5+5) : 20 = 2,85

Pro fajnšmekry – inteligentně napsaný vzorec aritmetického průměru je: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Krok 4 – Určíme modus

-modus je nejčastěji se vyskytující hodnota v souboru. Nahlédneme do sloupce Četnost výskytu a na první pohled vidíme, že nejčastější známkou byla dvojka (celkem 6)

-také může vyjít (což se v testu určitě nestane), že nejčastější hodnoty jsou dvě – (třeba by bylo šest dvojek i šest trojek), neva – tak jsou prostě modusy dva. Tak abyste se nevyděsili, až to někde uvidíte...

Krok 5 – Vypočteme medián

Medián je prostřední hodnota souboru.

Zjednodušeně řečeno seřadím-li všechny hodnoty souboru od nejmenší po největší, tak ta co bude uprostřed je medián.

Je-li soubor hodnot lichý např.: 1, 2, 3, 4, 5 – je to jednoduché, prostřední číslo je jasně trojka

Je-li soubor sudý např. 1, 2, 3, 4, 5, 6 – jsou uprostřed dvě čísla – totiž 3 a 4, absolutní střed tohoto souboru potom leží mezi těmito dvěma hodnotami – proto z nich udělám aritmetický průměr: (3+4) : 2 = 3,5. Takže mediánem bude číslo 3,5.

- Mediánem lichého souboru je vždy nějaká hodnota z tohoto souboru x mediánem sudého souboru může být některá z hodnot souboru (např. u hodnot 2, 2, 4, 4, 6, 6 je mediánem 4, protože prostředními čísly jsou dvě 4 a jejich aritmetický průměr je logicky 4), ale také nemusí (jak jsme si ukázali na příkladu 1, 2, 3, 4, 5, 6)

Postup, jak se mediánu dobrat vědecky (až budete mít 500 čísel):

U lichého souboru: Celkový počet hodnot vydělím 2 a výsledek zaokrouhlím nahoru. Získám tak pořadové číslo hledané hodnoty mediánu

př.: Mám už seřazené hodnoty (2, 4, 7, 8, 9, 10, 11, 11, 13, 14, 15, 16, 19), hodnot je dohromady 13; 13:2= 6,5 - zaokrouhleno 7. Najdu sedmé číslo v řadě – to je 11 a to je právě medián

U sudého souboru: Musím vlastně zjistit 2 prostřední čísla. První zjistím tak, že celkový počet hodnot vydělím 2. A druhé tak, že celkový počet hodnot vydělím dvěma a přičtu jedničku.

viz. náš příklad: 1. číslo 20:2 = 10

2. číslo (20:2) + 1 = 11

My ale nemáme hodnoty v řadě, máme je v tabulce. Zase vezmeme sloupec Četnost výskytu a jako řadu si ho představíme.

Vidíme, že známka 1 se vyskytla 3x, 2 – 6x, 3 – 5x ... – Když si to představíme jako řadu od nejmenšího po největší, tak to vypadá takhle

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	2	2	2	2	2	2	3	3	3	3	3	4	4	4	5	5	5

Jasně vidíme, že 10. a 11. číslem našeho souboru jsou trojky. Aritmetický průměr dvou trojek je 3. Medián je tedy 3.

Krok 6 – Vypočteme rozptyl
Rozptyl = průměr druhých mocnin odchylek od aritmetického průměru
Sice je to možný narvat do kalkulačky hned a najednou, ale pro názornost zpomaleně:
V tabulce budeme potřebovat 2 nové sloupce

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

Hodnota znaku (známka)	Četnost výskytu	Relativní četnost (%)	Y	Z	Q
1	3	15	3	3,4225	10, 2675
2	6	30	12	0,7225	4,335
3	5	25	15	0,0225	0,1125
4	3	15	12	1,3225	3,9675
5	3	15	15	4, 6225	13, 8675
					30,55

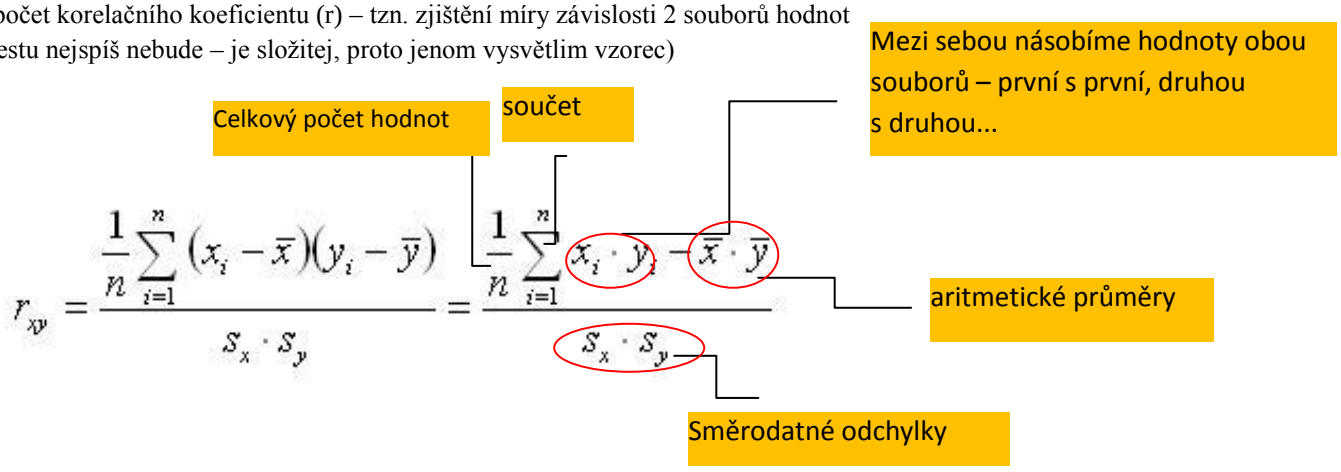
Sloupec Z = Hodnota znaku mínus aritmetický průměr, to celé na druhou - např. $(1 - 2{,}85)^2 = 3{,}4225$
Sloupec Q – Hodnoty ze sloupce Z vynásobíme Četností výskytu - např. 3,4225 x 3 = 10, 2675
Potom sečteme všechny hodnoty sloupce Q a vydělíme celkovým počtem hodnot: 30,55 : 20 = 1,6275 – to je rozptyl
Pro kontrolu: - rozptyl by měl vždy vycházet v rozmezí krajních hodnot (u nás něco mezi 1-5), sečteme-li rozptyl a aritmetický průměr, neměli bychom překročit maximální hodnotu (5), odečteme-li rozptyl od aritmetického průměru, neměli bychom spadnout pod minimální hodnotu (1)

Krok 7 – Vypočteme směrodatnou odchylku
Směrodatná odchylka je odmocnina z rozptylu: S = √1,6275 = 1,28

Krok 8 – Vypočteme variační koeficient
Variační koeficient nám ukazuje homogenitu souboru. Udává se v procentech – a vyjde-li víc jak 50%, říká nám to, že máme příliš rozházené hodnoty a jsou nám úplně na prd charakteristiky jako třeba aritmetický průměr. Náš soubor pěti známek je homogenní (hodnoty jsou si blízké). Ale proti tomu, výška měsíčních příjmů občanů ČR se pohybuje v obrovském rozpětí. Je tedy logické, že nemá smysl vůbec mluvit o průměrném platu...

Výpočet je jednoduchý – VK = (Směrodatná odchylka / Aritmetický průměr) x 100%
(1,28 : 2,85) x 100 = 44,9 %

Příklad 2:
Výpočet korelačního koeficientu (r) – tzn. zjištění míry závislosti 2 souborů hodnot (v testu nejspíš nebude – je složitej, proto jenom vysvětlim vzorec)



Postup:
1. Vynásobíme mezi sebou hodnoty obou souborů (x_1 krát y_1, x_2 krát y_2, atd...) a všechny tyto výsledky sečteme
2. Vydělíme výsledek celkovým počtem hodnot souboru
3. Odečteme od toho součin aritmetických průměrů
4. Vzniklý výsledek vydělíme součinem směrodatných odchylek obou souborů
5. Vyjde nám číslo od -1 do 1 a říká nám toto:

Koeficient korelace	Interpretace	Koeficient korelace	Interpretace
$r = 1$	naprostá závislost (funkční závislost)	$0{,}40 > r \geq 0{,}20$	nízká závislost
$1{,}00 > r \geq 0{,}90$	velmi vysoká závislost	$0{,}20 > r > 0{,}00$	velmi slabá závislost
$0{,}90 > r \geq 0{,}70$	vysoká závislost	$r = 0$	naprostá nezávislost
$0{,}70 > r \geq 0{,}40$	střední závislost		

Příklad 3:
Výpočet chí-kvadrátu
V určitém roce zemřelo v ČR na následky dopravních nehod 114 dětí. Je počet úmrtí ve všech měsících stejný?

Měsíc	Pozorovaná četnost *P*	$(P-O)^2$	$(P-O)^2/O$
leden	7	6,25	0,658
únor	7	6,25	0,658
březen	8	2,25	0,237
duben	10	0,25	0,026
květen	10	0,25	0,026
červen	9	0,25	0,026
červenec	14	20,25	2,132
srpen	12	6,25	0,658
září	10	0,25	0,026
říjen	12	6,25	0,658
listopad	7	6,25	0,658
prosinec	8	2,25	0,237
	114		6,000

H_0 = Počet úmrtí dětí je ve všech měsících v roce stejný

- hypotéza 0 vychází z otázky zadání

$$\chi^2 = \sum \frac{(P-O)^2}{O}$$

Postup:
1. Jako první si spočítáme aritmetický průměr (jenže u chí kvadrátu se mu říká Očekávaná četnost, tedy O)
- Tedy součet hodnot za všechny měsíce děleno 12; O = 114 : 12 = 9,5
2. Spočítáme si druhý sloupec – od Pozorované četnosti odečteme Očekávanou četnost a celé umocníme
3. Spočítáme třetí sloupec a to tak, že výsledky z druhého sloupce vydělíme Očekávanou četností (9,5)
4. Nakonec všechny hodnoty z třetího sloupce sečteme, vyjde nám 6
Výsledek nám říká, jaký je faktický rozdíl mezi očekávanou a pozorovanou četností. Jde nám o to, je-li nulová hypotéza platná, nebo jestli ji můžeme zamítnout.
Při rozhodování o platnosti nulové hypotézy porovnáváme vypočítanou hodnotu testového kritéria (6) s tzv. kritickou hodnotou. Příslušnou hodnotu hledáme v tabulkách
K tomu potřebujeme znát hladinu významnosti (P), ta je zadaná. Nejčastěji má hodnotu 0,05 (v testu bude určitě doslovně zadaná).
Dál je nutno znát počet stupňů volnosti tabulky. Ten se spočítá tak, že se vezme úplně původní tabulka, kde jsou jen zadané hodnoty. A pak počet řádek tabulky (těch s číselnými údaji) mínus 1 to celé krát počet sloupců mínus jedna, je-li sloupec (nebo řádek) jen jeden jako v našem případě, jednička se neodečítá
Tzn. Počet stupňů volnosti = (12-1) x 1 = 11
V další tabulce (která bude zadaná) si najdeme příslušnou hodnotu, to je pro P 0,05 a 11 stupňů volnosti číslo 19,68

Je-li náš výsledek (6) menší než hodnota z tabulky (19,68), znamená to, že hypotézu 0 nelze odmítnout.

Měsíc	Pozorovaná četnost *P*
leden	7
únor	7
březen	8
duben	10
květen	10
červen	9
červenec	14
srpen	12
září	10
říjen	12
listopad	7
prosinec	8

základní tabulka

stupně volnosti

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70

Příklad 4: Výpočet nezávislosti chí-kvadrát pro kontingenční tabulku
Počítáme, když chceme zjistit, jestli existuje souvislost mezi dvěma jevy, které byly zachyceny pomocí nominálního (i ordinálního) měření.

Skupině 400 vybraných studentů byl předložen dotazník, ve kterém byly mimo jiné dvě otázky:
Byl(a) jste ubytován(a) na kolejích? A Ano B Ne
Jaký byl Váš průměrný prospěch v loňském studijním roce?

- A lepší než 1,6
- B 1,6 – 2,1
- C horší než 2,1

Existuje vztah mezi tím, zda studenti bydlí na kolejích, a tím, jakých studijních výsledků dosahují?
H_0 = Mezi bydlením na koleji a studijními výsledky není závislost, zvolená hladina významnosti je 0,05

Postup:
1. Vypočítáme očekávanou četnost. U kontingenční tabulky ji počítáme pro každé políčko se základními údaji (39, 41, 107, 73...) zvlášť a to tak, že O = suma příslušného sloupce krát suma příslušného řádku, to celé lomeno celkovým počtem
př. pro 39: O = (80 x 239) / 400 = 47,8 nebo pro 47: O = (140 x 161) / 400 = 56,35
2. Jako další spočítáme pro každé políčko hodnotu chí-kvadrát podle známého vzorce $X = (P-O)^2 / O$, kde P jsou zadané údaje (39, 107, 73...)
př. pro 73 : $X = (73 – 72,45)^2 / 72,45 = 0,004$ nebo pro 41: $X = (41 – 32,2)^2 / 32,2 = 2,405$
3. Všechna X sečteme, vyjde nám: 6,628
4. Teď už potřebujeme určit jen počet stupňů volnosti (počítáme pro základní tabulku s údajema – označena modře):
PSV = (3 řádky – 1) x (2 sloupce – 1) = 2 x 1 = 2 stupně volnosti
5. V tabulce z příkladu 3 odečteme hodnotu pro 2 stupně volnosti a hladinu významnosti 0,05 a to je 5,99.
5,99 je menší než náš výsledek chí-kvadrát – 6,628, proto můžeme nulovou hypotézu směle zamítnout

	A **do 1,6**	**B** **1,6 - 2,1**	**C** **nad 2,1**	∑
bydlel(a) na koleji	39	107	93	**239**
	O = 47,8	O = 107,55	O = 83,65	
	X = 1,62	X = 0,003	X = 1,045	
nebydlel(a) na koleji	41	73	47	**161**
	O = 32,2	O = 72,45	O = 56,35	
	X = 2,405	X = 0,004	X = 1,551	

∑	**80**	**180**	**140**	**400**

Příklad 5:Test nezávislosti chí-kvadrát pro čtyřpolní tabulku (což je jen specifický druh kontingenční tabulky)
Kouří více studentů než studentek? Hladina významnosti je 0,01.
H_0 = Všichni kouří stejně.

	kouří	nekouří	∑
muži	11	1	**12**
ženy	15	21	**36**
∑	**26**	**22**	**48**

	kouří	nekouří	∑
muži	*a*	*b*	***a+b***
ženy	*c*	*d*	***c+d***
∑	***a+c***	***b+d***	***n***

Postup:
1. Nejdříve si tabulku přepíšeme podle ukázky, budeme to potřebovat pro následující vzorec:

2. Dosadit do něj to je „trivium Marie Terezie“...
$X = [48 \times (11 \times 21 - 1 \times 15)^2] / (12 \times 26 \times 22 \times 36) = 9{,}063$
3. Určíme stupeň volnosti: 1 x 1 je 1
4. Z tabulky příkladu 3 vidíme, že pro 1 stupeň volnosti a hladinu významnosti je hodnota 6,64
Náš výsledek 9,063 je vyšší, tudíž nulová hypotéza neplatí.