Příloha I

Statistické zpracování dat – test nezávislosti chí-kvadrát (χ^2)

Běžným způsobem sběru dat o nejrůznějších, často složitých jevech, je dotazníkové šetření. Výsledkem je řada dat, z kterých se snažíme zjistit něco zajímavého a užitečného. Můžeme například zjistit, jestli spolu souvisí dvě kvalitativní veličiny nebo zda jsou na sobě nezávislé a případně jak silná je závislost. Při statistickém zpracování dat je často vhodné použít test nezávislosti chí-kvadrát. Pro výpočet testu lze použít program, který si vytvoříme např. v Excellu nebo můžeme použít již vytvořené programy. Velmi povedená je aplikace Milana Kábrta na http://www.milankabrt.cz/testNezavislosti/index.php

Použitím této aplikace získáme rychle a snadno výsledky, které vyhodnotí statistický soubor. Pro měření síly vztahu můžeme použít korelační koeficienty. Jednoduchým způsobem lze vypočítat např. korigovaný koeficient kontingence pomocí Pearsona nebo Cramerův koeficient. Oba korelační koeficienty jsou z intervalu (0, 1). Na základě vypočtené hodnoty můžeme určit korelaci mezi hodnotami. Pokud je hodnota koeficientu 0, není mezi hodnotami žádný vztah. Je-li hodnota rovna 1, je mezi hodnotami v kontingenční tabulce silná závislost.

Test nezávislosti – test chí-kvadrát (χ^2)

Test nezávislosti chí-kvadrát se používá, pokud chceme zjistit závislost dvou kvalitativních veličin, které zjišťujeme na prvcích téhož výběru. Máme náhodný výběr rozsahu *n* rozdělený do dvou znaků (znak 1, znak 2). Úkolem testu je rozhodnout, zda znaky jsou na sobě závislé nebo nezávislé (zda znak 1 má vliv na znak 2).

Test chí-kvadrát porovnává skutečné (naměřené) a očekávané četnosti. Skutečné (naměřené) četnosti zjišťujeme z kontingenční tabulky. V kontingenční tabulce jsou ve sloupcích vyjádřené hodnoty znaku 1, v řádcích hodnoty znaku 2. Očekávané četnosti vypočítáme. Při výpočtu předpokládáme, že platí nulová hypotéza. Nulová hypotéza předpokládá, že znaky jsou nezávislé. Velikost rozdílů mezi skutečnými (naměřenými) a očekávanými četnosti se posuzuje pomocí testové

statistiky chí-kvadrát. Porovnává se vypočtená hodnota s kritickou hodnotou chí-kvadrát na dané hladině významnosti. Hladina významnosti se volí obvykle 0,05 nebo 0,1. Hladina významnosti představuje pravděpodobnost chyby při zamítnutí nulové hypotézy. Je-li hladina významnosti 0,05 (0,1), je pravděpodobnost, že jsme se dopustili chyby 5 % (10 %). Kritickou hodnotu pro daný stupeň volnosti najdeme v tabulkách. Počet stupňů volnosti zjistíme podle vztahu $(a\text{-}1)*(b\text{-}1)$, kde a je počet řádků a b je počet sloupců. Je-li kritická hodnota menší než vypočtená hodnota testového kritéria, zamítáme nulovou hypotézu a na dané hladině významnosti a přijímáme hypotézu o závislosti. Je-li kritická hodnota větší než vypočtená hodnota testového kritéria, nezamítáme nulovou hypotézu na dané hladině významnosti a platí, že znaky jsou nezávislé.

Hypotézy

- Nulová hypotéza: znaky 1 a 2 jsou nezávislé
- Alternativní hypotéza: mezi znaky 1 a 2 existuje závislost

Postup výpočtu

- Sestaví se tabulka skutečných (naměřených) relativních četností
- Vypočte se tabulka očekávaných četností
- Provede se kontrola podmínek pro použití testu nezávislosti v kontingenční tabulce:
  - nejvíce 20 % očekávaných četností může být menších než 5
  - žádná očekávaná četnost nesmí být menší než 1

Pozn. Platí pro náhodný výběr $n > 40$. Pro tabulku 2x2 je nutná úprava testového kritéria, pokud $20 < n < 40$, provádí se pomocí Yatesovy korekce. Pokud $n < 20$, používá se Fisherův test.

- Vypočte se testové kritérium (dosazení do vzorce – výsledek hodnota)
- Testové kritérium se srovná s kritickou hodnotou (tabulková hodnota, je potřeba zohlednit počet stupňů volnosti)
- Je-li testové kritérium < kritická hodnota, potom nezamítáme hypotézu o nezávislosti a nezávislost lze předpokládat
- Je-li testové kritérium > kritická hodnota, potom zamítáme hypotézu o nezávislosti a lze předpokládat závislost

Korigovaný koeficient kontingence pomocí Pearsona

Korigovaný koeficient kontingence pomocí Pearsona udává sílu vztahu. Nabývá hodnot z intervalu (0,1). Hodnota 0 znamená, že mezi hodnotami v kontingenční tabulce není žádný vztah, hodnota 1 znamená silnou závislost.

Korigovaný koeficient kontingence pomocí Pearsona vypočteme podle vztahu

$$C_{kor} = \frac{\sqrt{\frac{\chi^2}{\chi^2 + n}}}{\sqrt{\frac{m-1}{m}}}$$

kde χ^2 je hodnota testového kritéria, *n* je rozsah souboru, *m* je počet řádků nebo počet sloupců v kontingenční tabulce (je-li větší počet řádků, je *m* počet řádků; je-li větší počet sloupců, je m počet sloupců).

Cramerův koeficient

Cramerův koeficient udává sílu vztahu. Nabývá hodnot z intervalu (0,1). Hodnota 0 znamená, že mezi hodnotami v kontingenční tabulce není žádný vztah, hodnota 1 znamená silnou závislost.

Cramerův koeficient *V* vypočteme podle vztahu

$$V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

kde χ^2 je hodnota testového kritéria, *n* je rozsah souboru, *m* je počet řádků nebo počet sloupců v kontingenční tabulce (je-li větší počet řádků, je *m* počet řádků; je-li větší počet sloupců, je m počet sloupců).

Pro lepší pochopení zpracování dat uvádíme dva příklady.

Příklad 1

Chceme zjistit, zda spolu souvisí péče o zrak a nejvyšší dosažené vzdělání. Máme k dispozici 660 dotazníků týkajících se vad a ochrany zraku (náhodný výběr o rozsahu n=660). Z dotazníku vybereme otázky týkající se péče o zrak a dosaženého vzdělání.

Vybrané otázky z dotazníku:

Nejvyšší dosažené vzdělání

a) základní
b) středoškolské
c) vyšší odborné
d) vysokoškolské

Domníváte se, že se dostatečně pečujete o svůj zrak?

a) ano
b) ne
c) někdy

Znak 1 – nejvyšší dosažené vzdělání

Znak 2 – péče o zrak

Úkol testu – rozhodnout, zda nejvyšší dosažené vzdělání má vliv na péči o zrak

Postup výpočtu

1. Sestavíme tabulku skutečných (naměřených) relativních četností

Tab. 1a Skutečné (relativní) četnosti

	základní	SŠ	VOŠ	VŠ	celkem
ano	33	132	28	69	262
někdy	6	74	0	70	150
ne	11	128	6	103	248
celkem	50	334	34	242	660

Ve sloupcích tabulky jsou vyjádřené hodnoty znaku 1 – nejvyšší dosažené vzdělání, v řádcích hodnoty znaku 2 – péče o zrak. V jednotlivých buňkách tabulky je zaznamenáno, jak odpovídali respondenti z dané skupiny. Např. 33 respondentů se základním vzděláním odpovědělo, že se domnívá, že se dostatečně pečuje o svůj zrak.

Pro výpočet použijeme následující odkaz:

http://www.milankabrt.cz/testNezavislosti/index.php

Spustíme aplikaci a podle pokynů aplikace doplníme počet skupin znaku 1 a počet skupin znaku 2. Znak 1 má celkem 4 skupiny (základní, SŠ, VOŠ, VŠ), znak 2 má celkem 3 skupiny (ano, někdy, ne). Dále musíme doplnit hladinu významnosti α. Obvykle se volí 0,1 nebo 0,05. Zvolíme 0,05 a dále stiskneme tlačítko pokračovat. V následujícím kroku zadáme do tabulky naměřené relativní četnosti a dále stiskneme tlačítko pokračovat. Zobrazí se nám výsledky testu – tabulka očekávaných četností, hodnota testového kritéria, kritická hodnota testového kritéria pro daný počet stupňů volnosti a rozhodnutí.

Zkontrolujeme podmínky pro použití testu nezávislosti v kontingenční tabulce:

- nejvíce 20 % očekávaných četností může být menších než 5
- žádná očekávaná četnost nesmí být menší než 1

Tab. 1b Očekávané četnosti

	základní	SŠ	VOŠ	VŠ	celkem
ano	19,85	132,59	13,5	96,07	262
někdy	11,36	75,91	7,73	55	150
ne	18,79	125,5	12,78	90,93	248
celkem	50	334	34	242	660

Podmínky pro použití testu jsou v našem případě splněny a můžeme použít test nezávislosti chí-kvadrát. Hodnota testového kritéria je 54,792. Počet stupňů

volnosti je 6 (počet řádků 3, počet sloupců 4, odtud (3-1) ∗(4-1)=2∗3=6). Kritická hodnota pro hladinu významnosti 0,05 a počet stupňů volnosti 6 je 12,592 (viz tabulka 3). Protože kritická hodnota je menší než vypočtená hodnota testového kritéria, zamítáme nulovou hypotézu a na dané hladině významnosti a přijímáme hypotézu o závislosti.

Závěr: Zjišťovali jsme, zda péče o zrak souvisí s nejvyšším dosaženým vzděláním. Pro testování jsme použili test nezávislosti chí-kvadrát. Při výpočtu jsme použili program pro statistiku test nezávislosti chí-kvadrát http://www.milankabrt.cz/testNezavislosti/index.php
Porovnali jsme skutečné (naměřené) a očekávané četnosti. Skutečné (naměřené) četnosti jsme zaznamenali do kontingenční tabulky. Očekávané četnosti jsme vypočítali. Při výpočtu jsme předpokládali, že platí nulová hypotéza.
Nulová hypotéza: Péče o zrak nesouvisí s nejvyšším dosaženým vzděláním.
Alternativní hypotéza: Péče o zrak souvisí s nejvyšším dosaženým vzděláním.
Velikost rozdílů mezi skutečnými (naměřenými) a očekávanými četnosti jsme posoudili pomocí testové statistiky chí-kvadrát. Porovnali jsme vypočtenou hodnotu s kritickou hodnotou chí-kvadrát na hladině významnosti 0,05. Hladina významnosti 5 % představuje pravděpodobnost chyby při zamítnutí nulové hypotézy. Počet stupňů volnosti je v našem případě 6, kritická hodnota pro 6 stupňů volnosti a hladinu významnosti 0,05 je 12,592. Vypočtená hodnota testového kritéria je 54,792. Kritická hodnota je v našem případě menší než vypočtená hodnota testového kritéria, zamítáme nulovou hypotézu a na hladině významnosti 0,1 (10 %) a přijímáme hypotézu, že mezi péčí o zrak a nejvyšším dosaženým vzděláním existuje určitá závislost.

Výpočet korigovaného koeficientu kontingence pomocí Pearsona

Dosadíme do vztahu pro korigovaný koeficient kontingence pomocí Pearsona a dostaneme: $C_{kor} = \frac{\sqrt{\frac{\chi^2}{\chi^2+n}}}{\sqrt{\frac{m-1}{m}}} = \frac{\sqrt{\frac{54,792}{54,792+660}}}{\sqrt{\frac{4-1}{4}}} = 0,320$

Výpočet Cramerova koeficientu

Dosadíme do vztahu pro Cramerův koeficient a dostaneme:

$$V = \sqrt{\frac{\chi^2}{n(m-1)}} = \sqrt{\frac{54{,}792}{660(4-1)}} = 0{,}166$$

Získané hodnoty koeficientů naznačují, že mezi hodnotami v kontingenční tabulce je jenom slabá závislost. Nejvyšší dosažené vzdělání má vliv na péči o zrak, ale tato závislost není silná.

Příklad 2

Chceme zjistit, zda spolu souvisí péče o zrak a pohlaví. Máme k dispozici 660 dotazníků týkajících se vad a ochrany zraku (náhodný výběr o rozsahu *n*=660). Z dotazníku vybereme otázky týkající se péče o zrak a pohlaví.
Vybrané otázky z dotazníku:
Pohlaví
a) žena
b) muž
Domníváte se, že se dostatečně pečujete o svůj zrak?
a) ano
b) ne
c) někdy

Znak 1 – pohlaví
Znak 2 – péče o zrak

Úkol testu – rozhodnout, zda pohlaví má vliv na péči o zrak

Postup výpočtu
1. Sestavíme tabulku skutečných (naměřených) relativních četností

Tab. 2a Skutečné (relativní) četnosti

	muž	žena	celkem
ano	58	204	262
někdy	40	110	150
ne	47	201	248
celkem	145	515	660

Ve sloupcích tabulky jsou vyjádřené hodnoty znaku 1 – pohlaví, v řádcích hodnoty znaku 2 – péče o zrak. V jednotlivých buňkách tabulky je zaznamenáno, jak odpovídali respondenti z dané skupiny. Např. 58 mužů odpovědělo, že se domnívá, že se dostatečně pečuje o svůj zrak.

Pro výpočet použijeme následující odkaz:

http://www.milankabrt.cz/testNezavislosti/index.php

Spustíme aplikaci a podle pokynů aplikace doplníme počet skupin znaku 1 a počet skupin znaku 2. Znak 1 má celkem 2 skupiny (muž, žena), znak 2 má celkem 3 skupiny (ano, někdy, ne). Dále musíme doplnit hladinu významnosti α. Obvykle se volí 0,1 nebo 0,05. Zvolíme 0,05 a dále stiskneme tlačítko pokračovat. V následujícím kroku zadáme do tabulky naměřené relativní četnosti a dále stiskneme tlačítko pokračovat. Zobrazí se nám výsledky testu – tabulka očekávaných četností, hodnota testového kritéria, kritická hodnota testového kritéria pro daný počet stupňů volnosti a rozhodnutí.

Zkontrolujeme podmínky pro použití testu nezávislosti v kontingenční tabulce:

- nejvíce 20 % očekávaných četností může být menších než 5
- žádná očekávaná četnost nesmí být menší než 1

Tab. 2b Očekávané četnosti

	muž	žena	celkem
ano	57,56	204,44	262
někdy	32,95	117,05	150
ne	54,48	193,52	248
celkem	145	515	660

Podmínky pro použití testu jsou v našem případě splněny a můžeme použít test nezávislosti chí-kvadrát. Hodnota testového kritéria je 3,253. Počet stupňů volnosti je 2 (počet řádků 3, počet sloupců 2, odtud (3-1) ∗(2-1)=2∗1=2). Kritická hodnota

pro hladinu významnosti 0,05 a počet stupňů volnosti 2 je 5,991 (viz tabulka 3). Kritická hodnota je větší než vypočtená hodnota testového kritéria, nezamítáme nulovou hypotézu a na dané hladině významnosti, platí, že znaky jsou nezávislé.
Závěr: Z našeho šetření vyplynulo, že péče o zrak a pohlaví na sobě nezávisí. Závislost jsme ověřili pomocí kontingenčních tabulek a použili jsme test chí kvadrát. Kritická mez pro hladinu významnosti byla zvolena 0,05. Vypočtená hodnota testového kritéria je 3,253. Počet stupňů volnosti je 2, kritická hodnota pro 2 stupně volnosti je 5,991. Protože kritická hodnota je větší než vypočtená hodnota, z provedeného testu vyplývá, že veličiny jsou na sobě nezávislé.

Výpočet korigovaného koeficientu kontingence pomocí Pearsona
Dosadíme do vztahu pro korigovaný koeficient kontingence pomocí Pearsona a dostaneme: $C_{kor} = \frac{\sqrt{\frac{\chi^2}{\chi^2+n}}}{\sqrt{\frac{m-1}{m}}} = \frac{\sqrt{\frac{3,253}{3,253+660}}}{\sqrt{\frac{3-1}{3}}} = 0,086$

Výpočet Cramerova koeficientu
Dosadíme do vztahu pro Cramerův koeficient a dostaneme:

$$V = \sqrt{\frac{\chi^2}{n(m-1)}} = \sqrt{\frac{3,253}{660(3-1)}} = 0,050$$

Získané hodnoty koeficientů ukazují, že mezi hodnotami v kontingenční tabulce není závislost. Pohlaví nemá vliv na péči o zrak.

Kritické hodnoty testového kritéria chí-kvadrát

Tab. 3 Kritické hodnoty testového kritéria chí-kvadrát pro hladinu významnosti 0,05 a 0,01

Stupně volnosti	Hladina významnosti	
	0,05	0,01
1	3,841	6,635
2	5,991	9,21
3	7,815	11,341
4	9,483	13,277
5	11,070	15,086
6	12,592	16,812
7	14,067	18,475
8	15,507	20,09
9	16,919	21,666
10	18,307	23,209
11	19,675	24,725
12	21,026	26,217
13	22,362	27,688
14	23,685	29,141
15	24,996	30,578
16	26,296	32
17	27,587	33,409
18	28,868	34,805
19	30,144	36,191
20	31,410	37,566

Zdroj: www.milankabrt.cz/testNezavislosti/index.php