Testování rozdílů mezi četnostmi

1. Analýza kategoriálních dat

Ve statistice se setkáme s různými typy statistických znaků, pro jejichž analýzu je nezbytné volit i relevantní statistické metody. Pro data s nejvyšším stupněm kvantifikace, kam patří kardinální znaky (např. koncentrace glukózy, teplota, hmotnost), používáme metody popsané v ostatních kapitolách těchto studijních materiálů. Pomocí těchto postupů nejčastěji testujeme rozdíly mezi středními hodnotami, rozptyly a dalšími parametry. V praxi se ovšem často setkáváme i s daty s nižším stupněm kvantifikace, označujeme je jako kvalitativní znaky. Kvalitativní znaky jsou vyjádřeny slovním popisem. Mohou nabývat dvou možných alternativ (tzv. alternativní znaky) nebo více alternativ (tzv. množný neboli kategoriální znak). Typickými příklady alternativních znaků jsou například výskyt onemocnění (vyskytuje/nevyskytuje), prodělání vakcinace (vakcinován/nevakcinován), zastoupení pohlaví v chovu (samec/samice) nebo hodnocení mortality v chovu (živá/mrtvá). Příkladem množných znaků je například různá barva srsti (hnědá x černá x bílá) nebo barva očí (modrá x hnědá x zelená). U těchto kvalitativních znaků hodnotíme zastoupení neboli četnost jednotlivých alternativ. Nástrojem pro statistické hodnocení kvalitativních dat jsou hojně využívající se kontingenční tabulky, která pro svůj výpočet využívají x^2 test. Tento test pracuje na principu porovnávání pozorovaných a očekávaných četností.

2. Popis a vizualizace kvalitativních dat

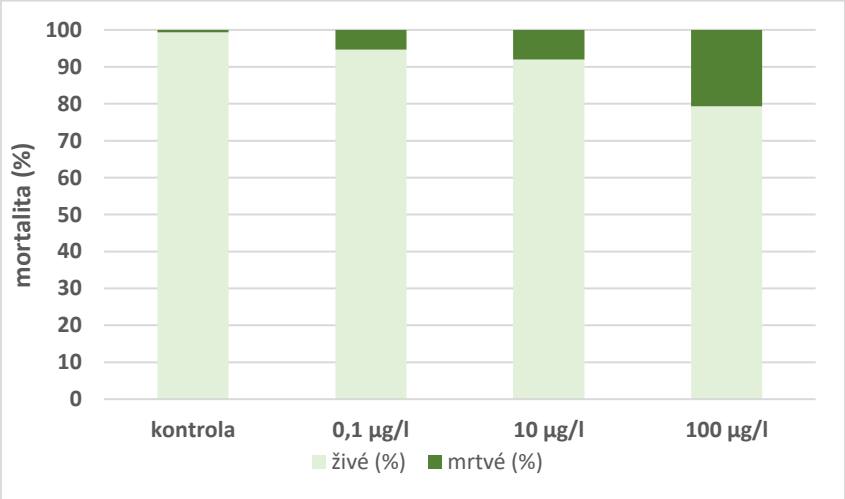
Četností rozumíme počet prvků se stejnou hodnotou statistického znaku, případně četností myslíme i počet prvků s hodnotami znaku, které patří do určité třídy/intervalu. Rozlišujeme tři základní druhy četností – absolutní, relativní a kumulativní četnost. Absolutní četnost je vyjádřena hodnotou četnosti zastoupených hodnot v daném statistickém souboru/intervalu. Součet četností všech možných hodnot znaku je roven počtu všech jednotek v souboru. Relativní četnost je dána podílem jednotlivých absolutních četností k rozsahu celého souboru. Vyjadřuje se buď desetinnými čísly, kdy součet relativních četností daného znaku je roven jedné nebo se vyjadřuje v procentech, kdy celkový součet všech relativních četností je roven 100 %. Kumulativní četnost lze vyjádřit opět jako absolutní nebo relativní hodnotu. Tento údaj představuje souhrnnou četnost statistických jednotek s hodnotami znaku menšími nebo rovnými hodnotě znaku. Při statistickém zpracování je třeba ale pracovat s absolutními

četnostmi. Relativní četnost se většinou využívá především pro prezentaci dat, a to ve formě tabulek nebo různých typů grafů (např. sloupcové či koláčové). V tabulce 1 je ukázka výpočtu absolutních a relativních četností pro reálný případ, na obrázku 1 je následně uvedena grafická prezentace výsledků relativní četnosti.

Tabulka 1: Ukázka výpočtu absolutních a relativních četností.
Zadání příkladu: V embryolarválním testu toxicity na kapru obecném byla v průběhu testování potenciálních negativních účinků antidepresiva fluoxetinu sledována mortalita. Do experimentu byla zařazena kontrolní skupina a tři různé testované koncentrace. V každé skupině bylo celkem 150 ks jiker. Hodnocení bylo provedeno v čase 96 hodin po oplození.

		kontrola	**0,1 µg/l**	**10 µg/l**	**100 µg/l**
absolutní četnosti	***živé (ks)***	149	142	138	119
	mrtvé (ks)	1	8	12	31
	celkem (ks)	150	150	150	150
relativní četnosti	***živé (%)***	99,3	94,7	92,0	79,3
	mrtvé (%)	0,7	5,3	8,0	20,7
	celkem (%)	100	100	100	100

Obrázek 1: Grafická prezentace relativních četností (viz zadání příkladu uvedené v tabulce1).



3. Kontingenční tabulky

Kontingenční tabulky se využívají pro sledování závislosti dvou nebo více kategoriálních proměnných. Kontingenční tabulky v sobě zahrnují rozložení výskytu jednotlivých kombinací sledovaných znaků. Kromě svého hojného využití při statistickém hodnocení kvalitativních dat, své uplatnění nachází i při analýze diskrétních kvantitativních dat a po rozdělení do kategorií lze kontingenční tabulky aplikovat i pro hodnocení spojitých kvantitativních dat. Před započetím statistického hodnocení si zvolíme nulovou (H_0) a

alternativní hypotézu (H_A). Nulová hypotéza uvádí, že závislost neexistuje. Naopak alternativní hypotéza potvrzuje přítomnost závislosti.

V praxi se nejčastěji vyskytuje závislost dvou znaků, kterou řešíme s využitím dvourozměrných (tzv. čtyřpolních) tabulek – 2 x 2. Tato tabulka je speciálním typem kontingenční tabulky, kdy hodnocená data nabývají pouze jedné ze dvou možných kategorií. Typickým příkladem je například výskyt konkrétního onemocní ve dvou chovech (vyskytuje/nevyskytuje), hodnocení zastoupení pohlaví ve dvou chovech (samice/samec) nebo hodnocení mortality ve dvou chovech (počet živých jedinců/počet mrtvých jedinců). Obecné schéma kontingenční tabulky formátu 2 x 2 je uvedeno na obrázku 2. Do kontingenční tabulky zapisujeme absolutní četnosti a je jedno, která proměnná tvoří sloupec a která tvoří řádky. Ukázka zpracování a vyhodnocení kontingenční tabulky formátu 2 x 2 ve statistickém programu Unistat for Excel 6.5 je uvedena na obrázku 4.

Obrázek 2: Obecné schéma kontingenční tabulky 2 x 2

		faktor 2		
		kategorie 1	**kategorie 2**	***celkem***
faktor 1	**kategorie 1**	**a**	**b**	a + b
	kategorie 2	**c**	**d**	c + d
	celkem	a + c	b + d	a + b + c +d

celková absolutní četnost

V některých případech je třeba studovat závislost více kategoriálních proměnných, kdy tedy pracujeme s kontingenční tabulkou formátu k x m (počet sloupců – k, počet řádků – m). Příkladem může být sledování mortality v chovu v průběhu tří let, kdy budeme hodnotit, zda frekvence nálezů se mění s lety a je tedy zřejmá závislost. Ukázka kontingenční tabulky formátu 3 x 2 je uvedena na obrázku 3. Zpracování dat s využitím kontingenční tabulky formátu 3 x 2 ve statistickém programu Unistat for Excel 6.5 je uvedena na obrázku 5.

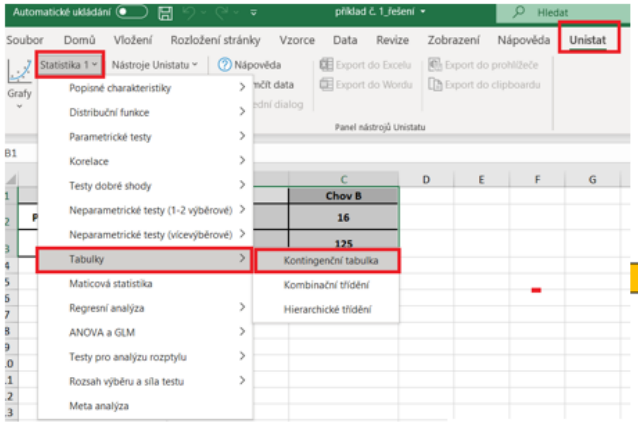
Obrázek 3: Kontingenční tabulka 3 x 2 – sledování mortality v chovu v průběhu tří let

	rok 2018	**rok 2019**	**rok 2020**	***celkem***
živé	**59**	**68**	**98**	225
mrtvé	**1**	**2**	**0**	3
celkem	60	70	98	228

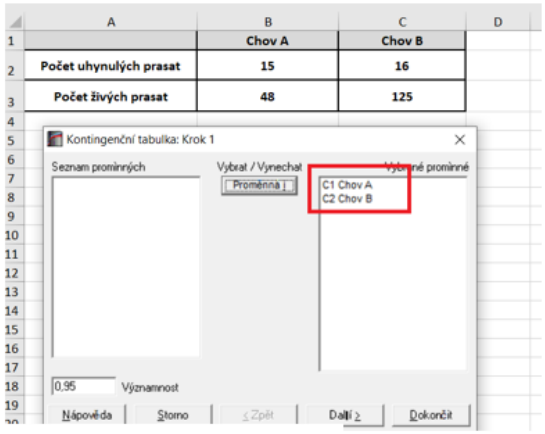
celková absolutní četnost

Obrázek 4. Kontingenční tabulka formátu 2 x 2. Řešení bylo provedeno v programu Unistat for Excel 6.5.

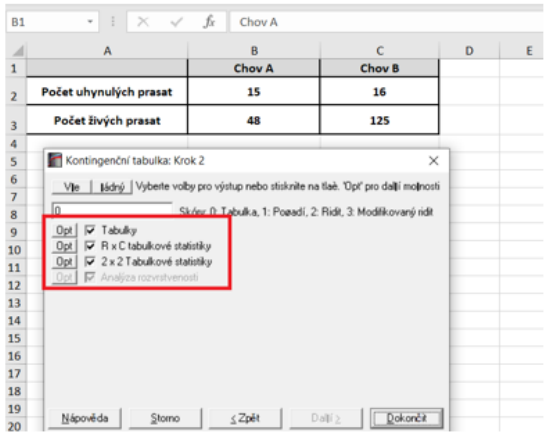
A. Označíme si zdrojová data a v hlavním menu si v nabídce Unistat zvolíme: **Statistika 1 → Tabulky → Kontingenční tabulka.**



B. V dalším okně si vybereme do položky proměnné oba porovnávané soubory a zvolíme „Další“.



C. Zaškrtneme si níže uvedené nabídky a potvrdíme tlačítkem „Dokončit“.



VYHODNOCENÍ

Chi-kvadrát testy

	Statistika Chi-kvadrát	Stupně volnosti	Pravostr. pravděp.
Pearson	5,2476	1	0,0220
Věrohodnostní poměr	4,9397	1	0,0262
Yatesova korekce	4,3251	1	0,0376
# Lineární /Lineární	5,2218	1	0,0223
~ McNemar-Bowker			

Tabulka skórů
~ Vykazováno pro 3 x 3 nebo větší čtvercové tabulky.
Buňky s očekávaným počtem < 5 = 0 (0,00%)
Minimální očekávaný počet = 9,5735

Fí = 0,1604
Koeficient kontingence = 0,1584
Cramerovo V = 0,1604

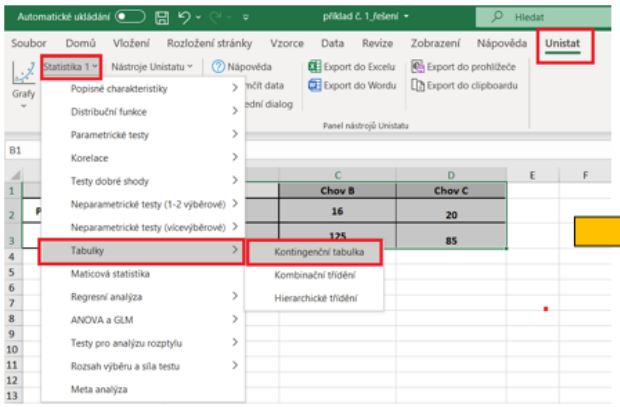
Fisherův přesný test

	Dvoustranná pravděpodobnost	Tabulka pravděpodobnosti
Fisherův přesný	0,0333	0,0135

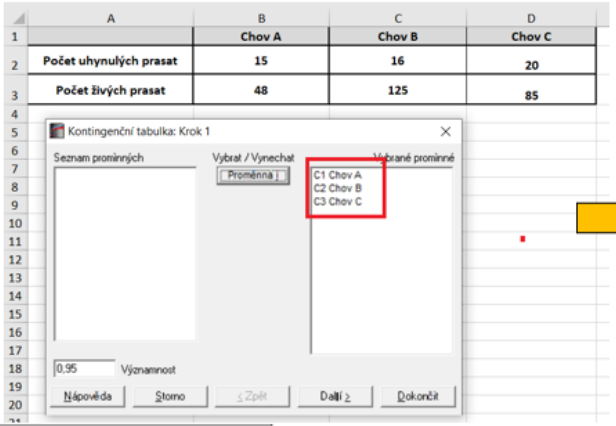
D. Dojde k vytvoření nového listu, kde nás pro vlastní vyhodnocení budou zajímat výsledky uvedené v tabulce Chí-kvadrát testy. Vzhledem k tomu, že provádíme testování tabulky 2x2, budeme využívat výsledek pravděpodobnosti Yatesovi korekce (v našem případě p = 0,0376). Na základě získaného výsledku můžeme tvrdit, že mezi porovnávanými chovy byl potvrzen rozdíl v mortalitě. Pokud by byla četnost menší než 5, brali bychom v potaz výsledek Fisherova přesného testu.

Obrázek 5. Kontingenční tabulka formátu 3 x 2. Řešení bylo provedeno v programu Unistat for Excel 6.5.

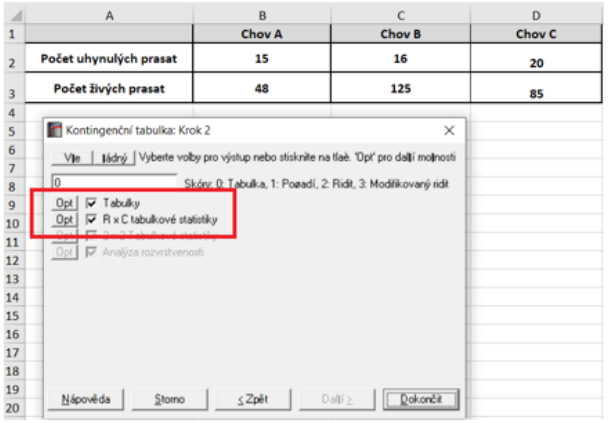
A. Označíme si zdrojová data (tabulka formátu 3 x 2) a v hlavním menu si v nabídce Unistat zvolíme: **Statistika 1 → Tabulky → Kontingenční tabulka.**



B. V dalším okně si vybereme do položky proměnné všechny porovnávané soubory a zvolíme „Další“.



C. Zaškrtneme si níže uvedené nabídky a potvrdíme tlačítkem „Dokončit“.



VYHODNOCENÍ

Chi-kvadrát testy

	Statistika Chí-kvadrát	Stupně volnosti	Pravostr. pravděp.
Pearson	5,6534	2	0,0592
Věrohodnostní poměr	5,6727	2	0,0586
+ Yatesova korekce			
# Lineární /Lineární	0,1664	1	0,6834
~ McNemar-Bowker			

\+ Vykazováno pro 2 x 2 tabulky.
\# Tabulka skórů
~ Vykazováno pro 3 x 3 nebo větší čtvercové tabulky.
Buňky s očekávaným počtem < 5 = 0 (0,00%)
Minimální očekávaný počet = 10,3981

Fí = 0,1353
Koeficient kontingence = 0,1340
Cramerova V = 0,1353

Fisherův přesný test

	Dvoustranná pravděpodobnost	Tabulka pravděpodobnosti
Fisherův přesný	0,0588	0,0012

D. Dojde k vytvoření nového listu, kde nás pro vlastní vyhodnocení budou zajímat výsledky uvedené v tabulce Chí-kvadrát testy. Vzhledem k tomu, že provádíme testování tabulky 3x2, budeme využívat výsledek pravděpodobnosti v řádku s označením Pearson (v našem případě p = 0,0592). Na základě získaného výsledku můžeme tvrdit, že mezi porovnávanými chovy nebyl potvrzen rozdíl v mortalitě. Pokud by byla hodnota pravděpodobnosti menší než 0,05, prováděli bychom následně dílčí testování kontingenčních tabulek ve formátu 2 x 2 pro odhalení potenciálního rozdílu.

Zdroje:

Hendl, J. Přehled statistických metod – analýza a metaanalýza dat. 2015. Vydavatelství Portál Praha, 736 s.

Lepš, J. Biostatistika. 1996. Jihočeská univerzita v Českých Budějovicích, 166 s.

Meloun, M., Militký, J. 2012. Kompendium statistického zpracování dat. Nakladatelství Karolinum, Univerzita Karlova v Praze. 982 s.