

OBSAH

Charakteristika a význam analýzy kategoriálnych údajov

Vzorové údaje 4.1

Praktické použitie programu SAS (SAS Enterprise Guide)

Príklad 4.1 (SAS)

Vzorové údaje 4.2

Príklad 4.2 (SAS)

Manuálny výpočet

Príklad 4.3

Príklad 4.4

Príklad 4.5

Praktické použitie programu R (R Studio)

Príklad 4.1 (Program R)

Príklad 4.2 (Program R)

Príklad 4.4 (Program R)

Ďalšie príklady použitia analýzy kategoriálnych údajov

Zdroje a zoznam použitej literatúry

Charakteristika a význam analýzy kategoriálnych údajov

Analýza kategoriálnych údajov sa zaoberá kategoriálnymi premennými bez ohľadu na to, či sú vysvetľujúce premenné kategoriálne alebo spojité. Na analýzu kategorických údajov existuje množstvo štatistických metód. Ak chceme vybrať vhodnú metódu, musíte určiť rozsah merania pre hodnotenú premennú. Rozsah merania takýchto premenných je dôležitým faktorom pri rozhodovaní sa o vhodnej štatistickej metóde, ktorú chcete použiť. Po výbere vhodnej mierky merania môžeme charakterizovať vzťah medzi kategoriálnymi premennými pomocou grafov a pomocou jednoduchých frekvenčných tabuliek.

Pri kategoriálnych premenných možno jednotlivé javy zatriediť do kategórií, tried, skupín, úrovní, typov, stavov.

Kategoriálne premenné môžu byť nominálne a ordinálne. Nominálne premenné majú hodnoty bez logického usporiadania. Ordinálne premenné majú hodnoty s logickým poradím, ale vyjadrenie relatívnej vzdialenosti medzi takýmito hodnotami sa určuje dosť problematicky.

Najjednoduchšia kategoriálna premenná má len dva stavy (dve úrovne) a nazývame ju dichotómna premenná. Príkladom je označenie pohlavia (muž, žena), možná odpoveď na otázku (áno, nie), alebo farba (biela, čierna).

Ak máme dve nominálne premenné, je vhodné použitie Pearsonovej Chí-kvadrát štatistiky. Sila asociácie sa dá merať Cramerovým V. Pearsonova Chí-kvadrát štatistika vyžaduje dostatočne veľkú veľkosť vzorky. Ak máme ale malú veľkosť vzorky, mali by sme skôr použiť presnejší Fisherov F- test.

Ak máme dve ordinálne premenné, na správne zistenie asociácie by sme mali použiť Mantel-Haenszelovu chí-kvadrát štatistiku. Silu asociácie možno v tomto prípade merať Spearmanovou korelačnou štatistikou.

Chí-kvadrát testy všeobecne slúžia na skúmanie asociácie medzi kategorickými premennými. Úrovne jednotlivých kategórií môžu byť dve alebo viac.

Najčastejšie riešime uvedené úlohy:

- Sú rozdielne dva (alebo viaceré) podiely v jednej kategorickej premennej od predpokladaných hodnôt v rámci populácie ?
- Existuje asociácia alebo závislosť medzi dvoma kategorickými premennými ?
- Líšia sa dva (alebo viaceré) podiely medzi sebou ?
- Existuje rozdiel v pravdepodobnosti udalosti medzi dvoma skupinami ?

V genetických experimentoch, pri ktorých zisťujeme početnosť rôznych tried, štiepne pomery, frekvencie genotypov, je následne potrebné vyhodnotiť, či tieto údaje zodpovedajú teoreticky očakávaným hodnotám. Na overenie experimentálne zistených a teoreticky očakávaných údajov sa používa χ^2 test (Chí - kvadrát test). Pri tomto štatistickom teste overujeme pravdivosť nulovej hypotézy, čo je predpoklad, že neexistuje rozdiel medzi očakávanými a teoretickými výsledkami. Na základe výsledkov χ^2 testu nulovú hypotézu buď prijímame (rozdiely medzi experimentálnymi a očakávanými hodnotami vznikli len vplyvom náhody) alebo zamietame (rozdiely sú štatisticky významné).

Počítaná hodnota χ^2 je sumou podielov štvorcov diferencií medzi experimentálnou (e) a teoretickou hodnotou (t):

$$\chi^2 = \sum_{i=1}^n \frac{(e - t)^2}{t}$$

n = celkový počet fenotypových tried,

e = experimentálna (pozorovaná) frekvencia i-tej triedy,

t = teoretická (očakávaná) frekvencia i-tej triedy

Často používaná forma rovnakého vzorca pre výpočet χ^2 je nasledovná:

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

O = pozorovaná (Observed) frekvencia danej triedy,

E = teoretická (Expected) frekvencia i-tej triedy

Pomocou Tabuľky kritických hodnôt χ^2 rozdelenia (Tab. 7) na základe vypočítanej χ^2 hodnoty zistíme, či je odchýlka náhodná, alebo nenáhodná. Hladina významnosti predstavuje najnižšiu kritickú hranicu pravdepodobnosti, s akou zamietame nulovú hypotézu. To znamená, že ak vypočítaná χ^2 -hodnota je rovná alebo vyššia ako kritická χ^2 - hodnota v tabuľke, nulovú hypotézu zamietame, ak je nižšia, nulovú hypotézu prijímame. P je pravdepodobnosť, že odchýlka pozorovaných hodnôt od očakávaných vznikla v dôsledku náhody a nie je štatisticky významná. χ^2 test využívame napr. na overenie štiepných pomerov pri mendelistickom type dedičnosti, pri génových interakciách, overovaní nezávislej kombinácie génov, alebo v populačnej genetike pre potvrdenie rovnováhy sledovanej populácie.

Pri výpočte Chí-kvadrátu testu je potrebné urobiť kontrolu očakávaných početností E pre jednotlivé bunky zostavenej tabuľky. Test nie je preukazný, ak očakávané frekvencie v niektorých poliach sú menšie alebo rovné hodnote 5. Pri tabuľkách 2x2 by nemala existovať ani jedna bunka s hodnotou očakávanej frekvencie menšou, alebo rovnou hodnote 5. Pri väčších kontingenčných tabuľkách by počet takýchto buniek nemal prekročiť 20 %.

Z dôvodu presnejšieho výpočtu je možné použiť tzv. Yatesovu korekciu. Chí-kvadrátu test je s použitím tejto korekcie potom presnejší. Yatesova korekcia ponúkajú štandardne viaceré štatistické programy, ale dá sa použiť aj pri manuálnom výpočte.

Vzorec výpočtu Chí-kvadrátu testu s Yatesova korekciou 0,5:

$$\chi^2 = \sum \frac{(O-E-0,5)^2}{E} \quad \text{ak } (O-E) \text{ pre danú bunku je kladná hodnota, alebo}$$

$$\chi^2 = \sum \frac{(E-O-0,5)^2}{E} \quad \text{ak hodnota } (O-E) \text{ pre danú bunku je záporné číslo.}$$

Všetky hodnoty Chí-kvadrát testov pre príslušné bunky spočítame.

Zostavovanie frekvenčných tabuliek

Frekvencia je počet výskytov danej hodnoty. Napríklad, ak 24 zvierat má v lineárnom hodnotení exteriéru 5 bodov za telesnú stavbu, potom skóre 5 má frekvenciu 24. Frekvencia je počet výskytov hodnôt v množine údajov. Kumulatívna frekvencia sa používa na určenie počtu pozorovaní, ktoré ležia pod konkrétnou hodnotou v súbore údajov. Kumulatívna frekvencia sa vypočíta spočítaním každej frekvencie z frekvenčnej tabuľky k súčtu jej predchádzajúcich hodnôt. Posledná hodnota sa vždy bude rovnať súčtu všetkých údajov. Relatívna frekvencia je frekvencia vydelená počtom všetkých hodnôt. Relatívne frekvencie možno písať ako zlomky, percentá alebo desatinné miesta. Kumulatívna relatívna frekvencia je akumulácia predchádzajúcich relatívnych frekvencií. Posledná hodnota sa vždy bude rovnať 100 resp.1.

Vzorové údaje 4.1 (zdroj: Databáza projektu KEGA)

Vzorové údaje 4.1 tvorí 2778 záznamov lineárneho hodnotenia exteriéru kráv holštajnského plemena uskutočneného v období máj 2021 - november 2021.

Súbor údajov **Datanew.dbf** obsahuje nasledovné premenné: štruktúra súboru.

Príklad 4.1 (SAS, Describe - One-Way Frequencies)

Uskutočnite frekvenčnú analýzu bodového hodnotenia telesnej kráv hoštajnského plemena a zistite jej frekvencie výskytu. Pomocou Chí-kvadrát testu overte, či jednotlivé frekvencie výskytu

bodových hodnotení sú rovnaké a tiež ich porovnajte s celopopulačnými frekvenciami v Slovenskej republike.

Výsledky

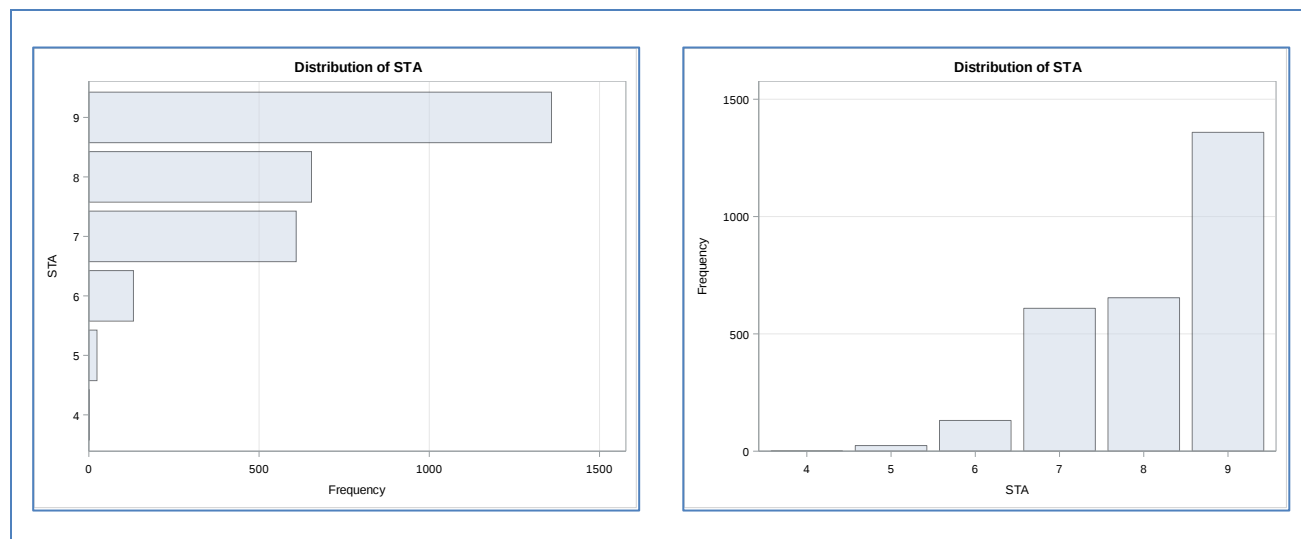
V tabuľke 4.1 uvádzame výsledky frekvenčnej analýzy lineárneho hodnotenia exteriéru kráv holštajnského plemena. Triediacim znakom je bodové hodnotenie telesnej stavby zvierat. Celkový počet hodnotených zvierat je 2778.

Tab. 4.1 Frekvenčná tabuľka - Telesná stavba

Telesná stavba	Frekvencia	Relatívna frekvencia	Kumulatívna frekvencia	Kumulatívna relatívna frekvencia
4	1	0.04	1	0.04
5	24	0.86	25	0.90
6	131	4.72	156	5.62
7	609	21.92	765	27.54
8	654	23.54	1419	51.08
9	1359	48.92	2778	100.00

Frekvenčné tabuľky je vhodné doplniť grafickými znázorneniami výskytu (distribúciou) jednotlivých frekvencií.

Tab. 4.2 Početnosti zvierat podľa hodnotenia telesnej stavby



Základným testom posúdenia rovnosti jednotlivých frekvencií je Chí-kvadrát test, ktorý v prípade jednoduchšej frekvenčnej analýzy testuje rovnaké pomery v rámci jednotlivých frekvencií výskytu hodnoteného znaku.

Tab. 4.3 Výsledky Chí-kvadrát testu

Chi-Square Test for Equal Proportions	
Chi-Square	2974.0864
DF	5
Pr > ChiSq	<.0001

Na základe uvedenej analýzy sa nepotvrdila stanovená nulová hypotéza o rovnosti jednotlivých frekvencií, čo je ale z hľadiska biologického správne, pretože inak by to znamenalo, že v skupine hodnotených zvierat by bola rovnosť čo nie je žiadúce (existovala by minimálna variabilita hodnoteného kvalitatívneho znaku).

Chí-kvadrát test môže byť použitý aj na testovanie nami zistených frekvencií oproti populačným frekvenciám, alebo priemerným frekvenciám rovnakého znaku v inej skupine pozorovaní. Uvádzame príklad použitia Chí-kvadrát testu pri známych celopopulačných frekvenciách hodnoteného znaku. V tabuľke 4.3 uvádzame populačné frekvencie telesnej stavby kráv holštajnskej populácie v Slovenskej republike (Candrák, Lichanec, 2021).

Tab. 4.4 Populačná frekvenčná tabuľka - Telesná stavba

Telesná stavba	Frekvencia	Relatívna frekvencia	Kumulatívna frekvencia	Kumulatívna relatívna frekvencia
4	8005	6.17	8005	6.17
5	14323	11.04	22328	17.21
6	27273	21.02	49601	38.22
7	37490	28.89	87091	67.11
8	22679	17.48	109770	84.59
9	19995	15.41	129765	100.00

Príklad nastavenia parametrov, procedúry FREQ a Chí-kvadrát testu v programe SAS (použitie populačných frekvencií kráv holštajnského plemena v Slovenskej republike):

```
PROC FREQ DATA=sasuser.data;
  TABLES STA / CHISQ testp=(6.17 11.04 21.02 28.89 17.48 15.41);
RUN;
```

Tab. 4.5 Výsledky Chí-kvadrát testu (populačné frekvencie)

Chi-Square Test for Specified Proportions	
Chi-Square	2910.7168
DF	5
Pr > ChiSq	<.0001

Výsledok Chí-kvadrát testu potvrdil, že relatívne frekvencie telesnej stavby hodnotenej skupiny zvierat sú štatisticky rozdielne oproti populačným frekvenciám. Nulovú hypotézu o rovnosti frekvencií musíme preto zamietnuť.

Vzorové údaje 4.2 (zdroj: Databáza údajov projektu KEGA)

Súbor údajov **MLIEKO.xlsx** obsahuje nasledovné premenné:

CISLO	číslo zvieraťa
PL	poradie laktácie
OTEC	lína-register otca zvieraťa
PLEM	plemenný typ
KODV	kód vyradenia zvieraťa
ROK	rok otelenia
LDNI	laktačné dni
MLIEKO	produkcia mlieka (kg)
TUK	produkcia tuku (kg)
TUKP	obsah tuku (%)
BIELK	produkcia bielkovín (kg)
BIELKP	obsah bielkovín (%)
LAKT	produkcia laktózy (kg)
LAKTP	obsah laktózy (%)

Vzor súboru údajov (10 záznamov)

CISLO	PL	OTEC	PLEM	KODV	ROK	LDNI	MLIEKO	TUK	TUKP	BIELK	BIELKP	LAKT	LAKTP
000003950	02	KF081	S0	54	2000	305	4393	152	3.46	140	3.19	210	4.78
000005026	02	TB001	S2	00	2003	305	4000	119	2.98	126	3.15	194	4.85
000005856	02	HX032	S0	54	2000	305	4972	239	4.81	160	3.22	222	4.47
000005866	03	DSO01	S0	00	2003	305	4454	232	5.21	154	3.46	210	4.71
000008866	02	RBB001	S0	00	2003	305	4793	199	4.15	154	3.21	226	4.72
000009950	03	KF081	SC	54	2003	305	4147	161	3.88	138	3.33	194	4.68
000013950	02	NEZ000	S2	54	2000	269	4368	137	3.14	139	3.18	213	4.88
000016950	03	STA002	SC	00	2001	305	4408	139	3.15	135	3.06	201	4.56
000027866	01	RBB001	S2	00	2002	305	4445	218	4.90	132	2.97	212	4.77

Príklad 4.2 (program SAS, Describe – Table Analysis)

Uskutočnite frekvenčnú a asociačnú analýzu skupiny 147 kráv slovenského strakatého plemena. Pomocou Chí-kvadrát testu overte, či skutočné frekvencie výskytu zvierat v rámci skupín plemenného typu a poradia laktácie sa rovnajú očakávaným frekvenciám.

Výsledky (program SAS)

Na základe frekvenčnej analýzy sme zistili nasledovné počty zvierat podľa plemenného typu a podľa poradia laktácie.

Tab. 4.6 Frekvenčná tabuľka – plemenný typ

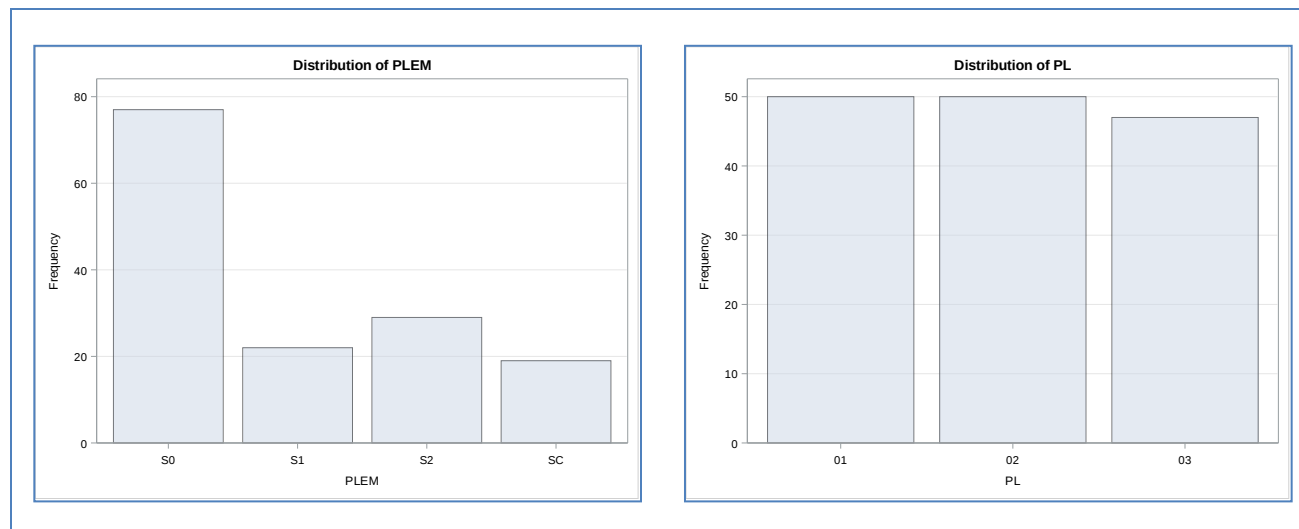
PLEM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
S0	77	52.38	77	52.38
S1	22	14.97	99	67.35
S2	29	19.73	128	87.07
SC	19	12.93	147	100.00

Tab. 4.7 Frekvenčná tabuľka – poradie laktácie

PL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01	50	34.01	50	34.01
02	50	34.01	100	68.03
03	47	31.97	147	100.00

Grafické vyjadrenie počtov kráv slovenského strakatého plemena uvádzame v stĺpcových grafoch (tabuľka 4.8).

Tab. 4.8 Počet kráv podľa plemenného typu a poradia laktácie



Tab. 4.9 Pozorované a očakávané frekvencie skupín zvierat

Table of PL by PLEM					
PL	PLEM				
Frequency Expected	S0	S1	S2	SC	Total
01	28 26.19	8 7.483	7 9.8639	7 6.4626	50
02	26 26.19	8 7.483	9 9.8639	7 6.4626	50
03	23 24.619	6 7.034	13 9.2721	5 6.0748	47
Total	77	22	29	19	147

Tab. 4.10 Výsledky Chí-kvadrát testu

Statistic	DF	Value	Prob
Chi-Square	6	3.1419	0.7908
Likelihood Ratio Chi-Square	6	3.0867	0.7979
Mantel-Haenszel Chi-Square	1	0.3836	0.5357
Phi Coefficient		0.1462	
Contingency Coefficient		0.1447	
Cramer's V		0.1034	

Tab. 4.11 Výsledky F- testu

Fisher's Exact Test	
Table Probability (P)	<.0001
Pr <= P	0.8081

Výsledky jednotlivých testov potvrdili platnosť nulovej hypotézy, že skutočne zistené a očakávané frekvencie kráv slovenského strakatého plemena podľa plemenného typu a poradia laktácie sú v zhode (sú približne rovnaké). Neexistuje štatistická preukaznosť rozdielného počtu zvierat v rámci jednotlivých kategórií.

Manuálny výpočet

Príklad 4.3 (Overovanie štiepných pomerov)

Krížením krátkosrstých morčiat s dlhosrstými sme v F_1 generácii získali všetky morčatá krátkosrsté (F_1 generácia je uniformná) a v F_2 generácii 49 krátkosrstých a 14 dlhosrstých (celkový počet jedincov $N = 63$).

Výsledky

Predpokladajme, že krátkosrstosť je podmienená dominantnou vlohou S a dlhosrstosť recesívnou vlohou v homozygotnom stave (ss). Na základe tohto predpokladu očakávame v F_2 generácii fenotypový štiepny pomer 3 : 1 (3/4 krátkosrstých a 1/4 dlhosrstých). Pomocou χ^2 testu overíme, do akej miery sa nami získaný štiepny pomer zhoduje s teoretickým. Teoretické zastúpenie jednotlivých fenotypových kategórií vypočítam nasledovne:

$$\text{Krátkosrstí} = \left(\frac{N}{4}\right) * 3 = \left(\frac{63}{4}\right) * 3 = 47,25$$

$$\text{Dlhosrstí} = \left(\frac{N}{4}\right) * 1 = \left(\frac{63}{4}\right) * 1 = 15,75$$

Tab. 4.12 Výpočet χ^2 hodnoty pre overenie štiepných pomerov

Trieda	Ideálny štiepny pomer	Fenotyp	e	t	$(e - t)^2$	$(e - t)^2/t$
	3	S_{-}	49	47,25	3,0625	0,0648
2	1	ss	14	15,75	3,0625	0,1944
n = 2	4		63	63		$\sum x^2 = 0,2592$

Vypočítanú hodnotu χ^2 porovnáme s tabuľkovou hodnotou (Tab. 4.13), ktorú nájdeme na priesečníku riadku so stupňami voľnosti (df) a stĺpca s pravdepodobnosťou (p). Stupne voľnosti predstavujú počet kategórií zmenšený o číslo 1 ($df = n - 1$, pri dvoch fenotypových triedach stupeň voľnosti 1). Štandardne prípustná hladina významnosti je $p = 0,05$, resp. $p = 0,01$.

V našom prípade je hodnota χ^2 testu (0,2592) nižšia ako kritická hodnota (3,851) v tabuľke na hladine významnosti $p = 0,05$, preto nulovú hypotézu prijímame a konštatujeme, že pozorovaný štiepny pomer zodpovedá očakávanému štiepnemu pomeru s pravdepodobnosťou platnosti na úrovni 95 %.

Tab. 4.13 Kritická hodnota pre stupne voľnosti (df) = 1 a pravdepodobnosť (p) = 0,05

Stupne voľnosti df	Pravdepodobnosť (p)				
	0,95	0,90	0,10	0,05	0,01
1	0,004	0,016	2,705	3,851	6,635
2	0,103	0,211	4,605	5,991	9,210
3	0,352	0,584	6,251	7,815	11,345

Príklad 4.4 (Overenie genetickej rovnováhy populácie)

Vypočítajte frekvencie alel, keď v populácii ľudí malo 450 krvnú skupinu MM, 325 MN a 225 NN. Zistíte, či je populácia v genetickej rovnováhe.

Výsledky

Populáciu považujeme za rovnovážnu, ak sa frekvencie alel a genotypov nemenia medzi generáciami. Tento stav je možné dosiahnuť v prípade, ak na populáciu nepôsobia faktory prostredia, ktoré by rovnováhu mohli narušiť. Vzťah pre výpočet rovnováhy je nasledovný:

$$\frac{H}{(D * R)^2} = 2$$

kde H je pozorovaný počet (resp. frekvencia) heterozygotov, D je pozorovaný počet (resp. frekvencia) dominantných homozygotov a R je pozorovaný počet (resp. frekvencia) recesívnych homozygotov. Ak sa výsledok rovná číslu 2, populácia je v rovnováhe. Tvrdenie, že populácia je, resp. nie je v rovnováhe môžeme overiť χ^2 testom, ktorým hodnotíme významnosť rozdielov medzi experimentálnymi a teoretickými počtami jedincov s konkrétnym genotypom.

V riešenom prípade sa jedná o kodominantný typ dedičnosti, môžeme preto genotyp MM považovať za dominantný (D), genotyp MN za heterozygotný (H) a genotyp NN za recesívny (R). Dosadením hodnôt do vzorca pre výpočet rovnováhy zistíme, či je populácia v rovnováhe:

$$\frac{325}{(450 * 225)^2} = 1,02 \neq 2$$

V našom prípade sa výsledok rovnice nerovná číslu 2, nepovažujeme preto populáciu za rovnovážnu. Následne toto tvrdenie overíme pomocou χ^2 testu. Pre výpočet χ^2 hodnoty musíme najprv zistiť očakávané frekvencie, resp. početnosť jednotlivých genotypových kategórií. Vychádzame z frekvencií alel, ktoré môžeme vypočítať dvomi spôsobmi:

1. na základe absolútnych genotypových frekvencií:

$$p_A = \frac{2D + H}{2N} = \frac{2 * 450 + 325}{2 * 1000} = 0,6 - \text{frekvencia dominantnej alely}$$

$$q_a = \frac{2R + H}{2N} = \frac{2 * 225 + 325}{2 * 1000} = 0,4 - \text{frekvencia recesívnej alely}$$

2. alebo na základe relatívnych genotypových frekvencií

$$d = \frac{450}{1000} = 0,45$$

$$h = \frac{325}{1000} = 0,325$$

$$r = \frac{225}{1000} = 0,225$$

$$p_A = d + \frac{1}{2}h = 0,45 + 0,16 = 0,6$$

$$q_a = r + \frac{1}{2}h = 0,225 + 0,16 = 0,4$$

Dosadením frekvencií alel do Hardy-Weinbergovej rovnovážnej rovnice zistíme očakávané frekvencie genotypov:

$$(p_A + q_a)^2 = p_{AA}^2 + 2pq_{Aa} + q_{aa}^2$$

$$(0,6 + 0,4)^2 = 0,36 + 0,48 + 0,16$$

kde p_{AA}^2 je frekvencia dominantných homozygotov, $2pq_{Aa}$ je frekvencia heterozygotov a q_{aa}^2 je frekvencia recesívnych homozygotov. Absolútne očakávané počty jedincov v jednotlivých genotypových kategóriách získame nasledovným výpočtom:

$$D = 0,36 * 1000 = 360$$

$$H = 0,48 * 1000 = 480$$

$$R = 0,16 * 1000 = 160$$

Experimentálne a teoretické hodnoty porovnáme χ^2 testom vyššie uvedeným spôsobom.

Tab.4.14 Výpočet χ^2 hodnoty pre overenie genetickej rovnováhy

Trieda	Fenotyp	e	t	$(e - t)^2$	$(e - t)^2/t$
1	M	450	360	8100	22,5
2	MN	325	480	24025	50,05
3	N	225	160	4225	26,4
n = 3		1000	1000		$\sum x^2 = 98,95$

Vypočítaná hodnota χ^2 (Tab. 4.14) je pri dvoch stupňoch voľnosti vyššia než kritická hodnota v tabuľke kritických hodnôt na hladine významnosti 0,05 (Tab. 4.15), preto nulovú hypotézu zamietame a môžeme konštatovať, že rozdiely medzi pozorovanými a očakávanými hodnotami sú štatisticky významné a populácia nie je v rovnováhe.

Tab.4.15 Kritická hodnota pre stupne voľnosti (df) = 2 a pravdepodobnosť p = 0,05

Stupne voľnosti (df)	Pravdepodobnosť (p)				
	0,95	0,90	0,10	0,05	0,01
1	0,004	0,016	2,705	3,851	6,635
2	0,103	0,211	4,605	5,991	9,210
3	0,352	0,584	6,251	7,815	11,345

Príklad 4.5 (Väzba génov)

Škvrnité sfarbenie králikov je recesívne oproti normálnemu (vloha D), dlhá sršť je recesívna oproti krátkej (vloha L). Po krížení normálne sfarbených krátkosrstých samíc so škvrnitým dlhosrstým samcom sme v B1 generácii získali 90 normálne sfarbených jedincov s krátkou sršťou, 20 škvrnitých s krátkou sršťou, 110 škvrnitých s dlhou sršťou a 10 normálne sfarbených s dlhou sršťou. Pomocou χ^2 -testu overte, či sú gény vo väzbe a o akú väzbovú fázu sa jedná.

P: ♀DL/dl x ♂dl/dl, Fenotyp: normálna, krátkosrstá x škvrnitý, dlhosrstý

Výsledky

Tab. 4.16 Praktické výsledky kríženia

	Fenotypová kategória	Fenotyp jedincov	Počet jedincov	Genotyp gamét heterozygotného rodiča
Parentálny genotyp	1	Normálne, krátka sršť	90	DL
	2	Škvrnité, dlhá sršť	110	dl
Rekombinovaný genotyp	3	Normálne, dlhá sršť	10	DI
	4	Škvrnité, krátka sršť	20	dL

Ak by sme uvažovali o klasickej mendelistickej dedičnosti týchto znakov, očakávali by sme na základe nezávislej kombinácie génov, že všetky typy rodičovských gamét budú vznikať s rovnakou pravdepodobnosťou. Na základe spätného testovacieho kríženia overíme, či sú gény podmieňujúce sledované znaky vo väzbe alebo nie a tiež to, v akej väzbovej fáze sa nachádzajú. Keďže sila väzby génov je daná frekvenciou crossing-overov, ktoré nastanú medzi sledovanými génmi, zaujíma nás, či je pomer parentálnych a rekombinovaných gamét 1:1, alebo je zmenený v dôsledku väzby génov. V našom prípade sme v B1 generácii pozorovali 230 jedincov, preto očakávaný počet rodičovských aj rekombinovaných genotypov pri uvedenom štiepnom pomere je 115. Významnosť rozdielov medzi pozorovaným a očakávaným pomerom overíme χ^2 testom.

Tab. 4.17 Výpočet χ^2 hodnoty pre voľnú kombinovateľnosť génov

Trieda	Ideálny štiepny pomer	Fenotyp	e		t	(e - t) ²	(e - t) ² /t
1	1	DL	90	200	115	7225	62,8
		dl	110				
2	1	DI	10	30	115	7225	62,8
		dL	20				
n = 2	2		230				$\sum x^2 = 125,6$

Vypočítaná χ^2 hodnota 125,6 (Tab. 4.17) je vyššia než kritická tabuľková hodnota v priesečníku stupňov voľnosti (df) = 1 a pravdepodobnosť (p) = 0,05, resp. p = 0,01 (Tab. 4.18), preto zamietame nulovú hypotézu a konštatujeme, že rozdiely medzi experimentálnymi a očakávanými hodnotami sú štatisticky signifikantné. Odchýlka od očakávaných hodnôt je spôsobená väzbou génov, ktoré ležia na jednom chromozóme a nedochádza k nezávislej kombinácii alel a nevzniká rovnaký počet rodičovských a rekombinovaných gamét. Rodičia mali gény D a L vo väzbovej fáze CIS, čo vieme určiť na základe počtu jedincov v jednotlivých fenotypových kategóriách (vznikali prevažne jedince s kombináciou DL a dl).

Tab. 4.18 Kritická hodnota pre stupne voľnosti df = 1 a pravdepodobnosť (p) = 0,05 a p = 0,01

Stupne voľnosti (df)	Pravdepodobnosť (p)				
	0,95	0,90	0,10	0,05	0,01
1	0,004	0,016	2,705	3,851	6,635
2	0,103	0,211	4,605	5,991	9,210
3	0,352	0,584	6,251	7,815	11,345

Praktické použitie programu R (R Studio)**Príklad 4.1** (program R)

Uskutočnite frekvenčnú analýzu bodového hodnotenia telesnej kráv hoštajnského plemena a zistite jej frekvencie výskytu. Pomocou Chí-kvadrát testu overte, či jednotlivé frekvencie výskytu bodových hodnotení sú rovnaké a tiež ich porovnajte s celopopulačnými frekvenciami v Slovenskej republike.

Zadanie analýzy

```
# Import a zobrazenie údajov
file_path <- "http://e-biostat.uniag.sk/wp-content/uploads/2022/01/datanew.txt"
Datanew <- read.delim(file_path)
View(Datanew)

# Frekvenčná analýza (Telesná stavba)
counts1 <- table(Datanew$STA)
View(counts1)
barplot(counts1, main="Telesná stavba",
  xlab="Počet zvierat")

# 2-Way Frequency Table
attach(Datanew)
mytable <- table(Datanew$STA)
mytable # print table
margin.table(mytable, 1)
prop.table(mytable) # cell percentages

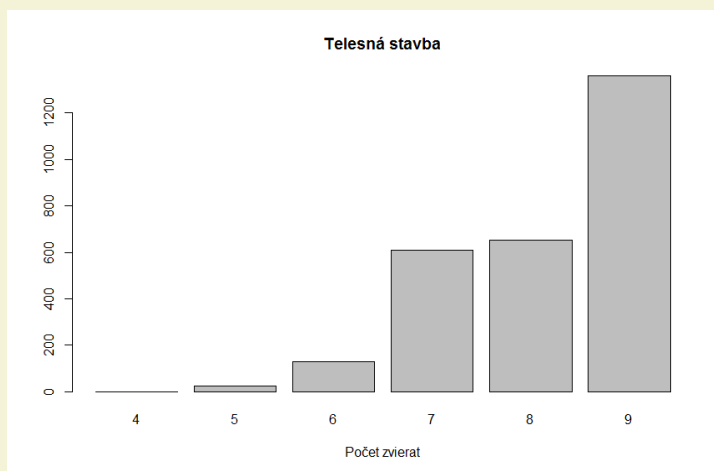
# Chí-kvadrát test
chisq.test(mytable)

# 2-Way Cross Tabulation
library(gmodels)
CrossTable (Datanew$STA)
```

Výsledky analýzy (program R)

Telesná stavba

	Var1	Freq
1	4	1
2	5	24
3	6	131
4	7	609
5	8	654
6	9	1359



Pearson's Chi-squared test

X-squared = 2974.1, df = 5, p-value < 2.2e-16

Cell Contents

					N
					N / Table Total

Total observations in Table: 2778

	4	5	6	7	8
1	1	24	131	609	654
	0.000	0.009	0.047	0.219	0.235
9					
					1359
					0.489

Príklad 4.2 (program R)

Uskutočnite frekvenčnú a asociačnú analýzu skupiny 147 kráv slovenského strakatého plemena. Pomocou Chí-kvadrát testu overte, či skutočné frekvencie výskytu zvierat v rámci skupín plemenného typu a poradia laktácie sa rovnajú očakávaným frekvenciám.

Zadanie analýzy

```
# Import a zobrazenie údajov
file_path <- "http://e-biostat.uniag.sk/wp-content/uploads/2022/01/Mlieko.txt"
Mlieko_R <- read.delim(file_path)
View(Mlieko_R)

# Frekvenčná analýza (jeden znak, Plemenné typy)
counts1 <- table(Mlieko_R$PLEM)
View(counts1)
barplot(counts1, main="Plemenné typy",
  xlab="Počet zvierat")

# Frekvenčná analýza (jeden znak, Poradie laktácie)
counts2 <- table(Mlieko_R$PL)
View(counts2)
barplot(counts2, main="Poradie laktácie",
  xlab="Počet zvierat")

# Frekvenčná analýza (kontingenčná tabuľka, 2 znaky)
attach(Mlieko_R)
mytable <- table(Mlieko_R$PL, Mlieko_R$PLEM)
mytable # print table
margin.table(mytable, 2)
margin.table(mytable, 1)
prop.table(mytable) # cell percentages
prop.table(mytable, 1) # row percentages
prop.table(mytable, 2) # column percentages

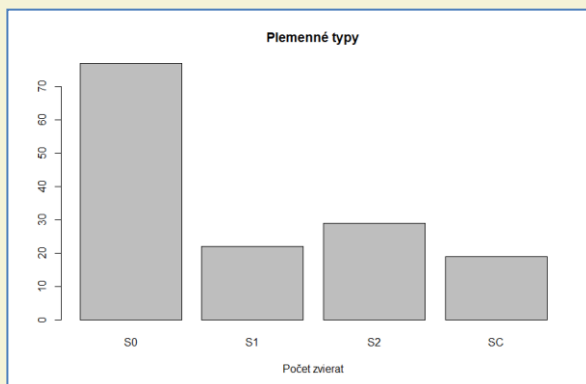
# Chí-kvadrát test
chisq.test(mytable)

# Kontingenčná tabuľka (2 znaky)
library(gmodels)
CrossTable (Mlieko_R$PLEM, Mlieko_R$PL)
```

Výsledky analýzy (program R)

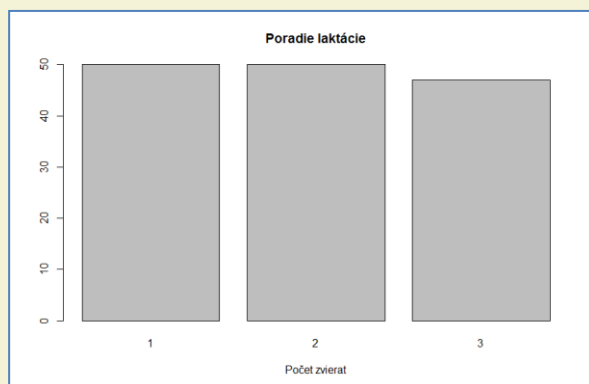
Plemenné typy

	Var1	Freq
1	S0	77
2	S1	22
3	S2	29
4	SC	19



Poradie laktácie

	Var1	Freq
1	1	50
2	2	50
3	3	47



Pearson's Chi-squared test

X-squared = 3.1419, df = 6, p-value = 0.7908

Cell Contents				
	N			
	Chi-square contribution			
	N / Row Total			
	N / Col Total			
	N / Table Total			

Total observations in Table: 147

Mlieko_R\$PLEM	Mlieko_R\$PL	1	2	3	Row Total
S0		28	26	23	77
		0.125	0.001	0.106	
		0.364	0.338	0.299	0.524
		0.560	0.520	0.489	
		0.190	0.177	0.156	
S1		8	8	6	22
		0.036	0.036	0.152	
		0.364	0.364	0.273	0.150
		0.160	0.160	0.128	
		0.054	0.054	0.041	
S2		7	9	13	29
		0.832	0.076	1.499	
		0.241	0.310	0.448	0.197
		0.140	0.180	0.277	
		0.048	0.061	0.088	
SC		7	7	5	19
		0.045	0.045	0.190	
		0.368	0.368	0.263	0.129
		0.140	0.140	0.106	
		0.048	0.048	0.034	
Column Total		50	50	47	147
		0.340	0.340	0.320	

Príklad 4.4 (Overenie genetickej rovnováhy populácie)

Vypočítajte frekvencie alel, keď v populácii ľudí malo 450 krvnú skupinu MM, 325 MN a 225 NN. Zistite, či je populácia v genetickej rovnováhe.

Zadanie analýzy**# Import a zobrazenie údajov**

Pre analýzy tohto typu nemusíme nevyhnutne importovať do programu R údaje. Pokiaľ máme uvedené početnosti jednotlivých analyzovaných frekvencií môžeme ich priamo zadať do jednotlivých príkazov súvisiacich s Hardy - Weinbergovou rovnováhou (v prípade individuálnych záznamov, ale samozrejme musíme uskutočniť ich import do systému).

Hardy-Weinberova rovnováha

```
# Použijeme špeciálnu knižnicu programu R - install.packages("HardyWeinberg")  
library("HardyWeinberg")
```

Zadanie jednotlivých frekvencií (frekvencie sa uložia do poľa x)

```
x <- c(MM = 450, MN = 325, NN = 225)
```

```
# Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
```

```
HW.test <- HWChisq(x, verbose = TRUE)
```

```
HW.test <- HWChisq(x,cc=0,verbose=TRUE)
```

```
# Likelihood ratio test for Hardy-Weinberg equilibrium
```

```
HW.lrtest <- HWLratio(x, verbose = TRUE)
```

```
# Haldane Exact test for Hardy-Weinberg equilibrium (autosomal) using SELOME p-value
```

```
HW.exacttest <- HWExact(x, verbose = TRUE)
```

Komplexné testy

```
HW.results <- HWAlltests(x, verbose = TRUE, include.permutation.test = TRUE)
```

Výpočet sily testu

```
n <- sum(x)
```

```
nM <- mac(x)
```

```
pw4 <- HWPowder(n, nM, alpha = 0.05, test = "exact", theta = 4,
```

```
pvalue.type = "selome")
```

```
print(pw4)
```


Tab. 4.19 Kritické hodnoty χ^2 – rozdelenia

Stupne voľnosti (df)	Pravdepodobnosť (p)				
	0,95	0,90	0,10	0,05	0,01
1	0,004	0,016	2,705	3,851	6,635
2	0,103	0,211	4,605	5,991	9,210
3	0,352	0,584	6,251	7,815	11,345
4	0,711	1,064	7,779	9,488	13,277
5	1,145	1,610	9,236	11,071	15,086
6	1,635	2,204	10,645	12,592	16,812
7	2,167	2,833	12,017	14,067	18,475
8	2,733	3,490	13,362	15,507	20,090
9	3,325	4,168	14,684	16,819	21,666
10	3,940	4,865	15,987	18,307	23,209
11	4,575	5,578	17,275	19,675	24,725
12	5,226	6,304	18,549	21,026	26,217
13	5,892	7,042	19,812	22,362	27,688
14	6,571	7,790	21,064	23,685	29,141
15	7,261	8,547	22,307	24,996	30,578
20	10,851	12,443	28,412	31,410	37,566
25	14,611	16,473	34,382	37,652	44,314
30	18,493	20,599	40,256	43,773	50,892
35	22,465	24,797	46,059	49,802	57,342
40	26,509	29,051	51,805	55,758	63,691
45	30,612	33,350	57,505	61,656	69,957
50	34,764	37,689	63,137	67,505	76,154
60	43,188	46,459	74,397	79,082	88,379
70	51,739	55,329	85,527	90,531	100,425
80	60,391	64,278	96,578	101,879	112,329
100	77,929	82,358	118,498	124,342	135,807
140	113,659	119,029	161,827	168,613	181,840
500	449,147	459,926	1057,724	1074,679	1106,969