3 Testy shody, závislost dvou kategoriálních veličin

3.1 Testy dobré shody

Testy dobré shody slouží k ověření shody mezi empirickým a teoretickým rozdělením. Nulová hypotéza v těchto testech říká vždy "Je shoda mezi předpokladem a pozorovanými daty", alternativní hypotéza zní "Není shoda". V každém testu je však shoda ověřována jinak.

Nulová hypotéza v testech shody: je shoda mezi předpokladem a pozorovanými daty

Mnoho statistických metod předpokládá určitý typ rozdělení. Velmi často se vyskytuje požadavek normality dat, tedy, že data pochází ze základního souboru s normálním rozdělením. Budeme rozlišovat dva případy

- Testujeme typ rozdělení s předem známými parametry (například normální rozdělení s danou střední hodnotou a rozptylem). V takovém případě říkáme, že model je plně specifikován.
- 2) Model není plně specifikován a jeho parametry (například u normálního rozdělení střední hodnotu a rozptyl) je nutno odhadnout z výběrových dat.

3.1.1 Pearsonův chí-kvadrát test

Tímto testem můžeme testovat tři různé hypotézy:

- Test dobré shody ověřuje, zda má veličina rozdělení pravděpodobnosti určitého typu.
- Test nezávislosti posuzuje závislost dvou veličin měřených na jednotkách z jednoho výběru.
- Test homogenity slouží k porovnání rozložení veličin v alespoň dvou populacích.

Základní myšlenka testu chí-kvadrát: porovnáváme pozorované a očekávané četnosti. Pozorované četnosti jsou známy z výběrového souboru. Očekávané (teoretické) četnosti musíme vypočítat.

Předpoklady testu: Očekávané (teoretické) četnosti musí splňovat podmínku, že alespoň v 80 % musí být větší než 5 a všechny musí být větší než 1. Pokud tomu tak není, musíme sousední kategorie slučovat (pokud je to možné).

Dále se podrobněji budeme věnovat testu Chí kvadrát test dobré shody. Zjišťování závislosti dvou kategoriálních veličin odložíme do kapitoly 3.2

Ověřování shody testem chí kvadrát budeme demonstrovat na příkladech.

Příklad 1

Firma chce uvést na trh nový výrobek ve čtyřech různých provedeních designu a předpokládá, že zájem o jednotlivé druhy designu (označme je A, B, C, D) bude následující:

- A: 35%
- B: 10%
- C: 5%
- D: 50%

Pro potvrzení svého předpokladu provedla firma průzkum, ze kterého vyplynulo, že z 300 potencionálních zájemců o tento výrobek by zájem o design A projevilo 110 zájemců, o design B 20 zájemců, o design C 10 zájemců a o design D 160 zájemců. Ověřte na 5% hladině významnosti, zda tyto zjištěné výsledky potvrzují předpoklad firmy.

Máme dány čtyři teoretické pravděpodobnosti $\pi_1=0,35$; $\pi_2=0,1$; $\pi_3=0,05$; $\pi_4=0,5$. Očekávané četnosti spočteme jako počet všech zájemců krát teoretická pravděpodobnost, tedy $n\pi_j$.

0		າ -	(H-) =		-		
	Dor	nů	Vložen	í Rozlo	žení stránky	Vzorce	Data
Vic	ožit v v v v v v v v v v v v v v v v v v v		ΙŪ	• 11 •			Zar
	CHITE	ST	-	(• × ✓	<i>f</i> _{sc} =300	*0,35	
	А		В	С	D	E	F
1							
2					pozor.	oček.	
3					110	=300*0,35	
4					20		
5					10		
6					160		
7				součet	300		

		17 - (1 -)	Ŧ	-				_	_	Se
C	Do	mů Vlože	ní R	ozložení stránky	Vzorce	Data	Revize	Zobrazení	Acroba	ıt
ľ	3 🔏	Calibri	- 11	· A A	= = =	; »»~	Zalamov	at text		Obecný
VI	ožit √	BIU	· [] ·				🔤 Sloučit a	zarovnat na	střed 🔻 🚦	9 - 9
Sch	ránka 🖻		Písmo	Gi.		Zā	arovnání		R.	č
	CHIT	EST	- (• X	✓ f ∗ =CH	TEST(C2:C5	;D2:D5)				
	Α	В	C	D	E	F	G	Н	1	J
1			pozor	oček.						
2				110 105	5					
3				20 30						
4				10 15						
5				160 150						
0		soucet		300 300)					
2		-CHITEST	102.05	02.05)						
9		-criftEst	(02.03,	52.057						
10	Argun	nenty funkce							? X	
11		FOT								
12	CHI	IESI								
13		Akt	uální (2:C5		= {	{110 20 10 16	50}		
14		0čeká	vané D	2:D5		E = {	(105 30 15 15	50}		
15						= (),116336824]		
16	Vrátí	test nezávislosti	: hodnota	ze statistického r	ozdělení chí-k	vadrát a přís	lušné stupně v	/olnosti.		
17	_		00	ekávané je ob	last dat obsah	ující podíl sou	učinu součtů řa	ádků a sloupců	a celkového	
18	_			SOUC	.u.					
19										-
20	/ýsle	dek = 0,11633	6824							
21	Nápo	věda k této funk	ci					ок	Storno	
22										
25										

MS Excel: V souboru DataExcelchishoda.xls máme zadán v jednom sloupci pozorované četnosti, ve druhém máme vypočítány očekávané četnosti pomocí příkazu = pravděpodobnost krát rozsah výběru. Zvolíme statistické funkce, dále vybereme funkci CHITEST. Do okna Aktuální označíme oblast pozorovaných četností, do okna Očekávané označíme oblast očekávaných četností. Výsledná p-hodnota se objeví už zadávácím okně. Klikneme-li na OK, objeví se nám p-hodnota daného testu v příslušné buňce listu. V našem případě p=0,11633824.

Výsledná p-hodnota je větší než běžné hladiny významnosti 0,01; 0,05 i než 0,1 a proto nelze zamítnout hypotézu o shodě rozdělení. Předpoklad firmy není v rozporu se zjištěnou strukturou zájmu o výrobek z průzkumu.

SW Statistica: V souboru data shoda.sta máme uložená data. Proměnná 5 jsou pozorované četnosti, proměnná 6 očekávané četnosti. Vybereme záložku Statistiky a dále Neparametrické statistiky. V základním výběru zvolíme záložku Pozorované vs. očekávané X2. V dialogovém okně Proměnné vybereme z proměnných pozorované (v našem případě proměnná 5) a očekávané četnosti (v našem případě proměnná 6) a klikneme na záložku Výpočet: pozor. vs. oček. četnosti. Objeví se následující tabulka, ve které vidíme hodnotu testového kritéria Chi-Kvadr. (5,904762), stupně volnosti (3) a hlavně p-hodnotu (0,116339), kterou porovnáme s běžnými hladinami významnosti.

P-hodnota je větší než 0,01; 0,05 i než 0,1 a proto nelze zamítnout hypotézu o shodě rozdělení na jednoprocentní, pětiprocentní ani deseti procentní hladině významnosti. Předpoklad firmy není v rozporu se zjištěnou strukturou zájmu o výrobek z průzkumu.

		Pozorované Chi-Kvadr. =	vs. očekáva = 5,904762	ané četnosti sv = 3 p = ,1	(Tabulka2) 16339
		pozorov.	očekáv.	P - O	(P-O)^2
Příp	ad	Prom5	Prom6		/O
C:	1	110,0000	105,000(5,0000	0,238095
C:	2	20,0000	30,0000	-10,0000	3,333333
C:	3	10,0000	15,0000	-5,0000	1,666667
C:	4	160,0000	150,000(10,0000	0,666667
Sčt		300,0000	300,000(0,0000	5,904762

Příklad 2

Na úřadu byl sledován počet občanů přicházejících s žádostmi v průběhu rozšířených úředních hodin pro veřejnost (Od 9 do 19 hodin). Pro zjištění rovnoměrnosti využití těchto hodin pro veřejnost byly během jednoho úředního dne zjištěny údaje uvedené v souboru DataExcelchishoha.xls. Lze na základě těchto dat učinit závěr, že zákazníci přicházejí v průběhu dne (9 hod. - 19 hod.) v rámci dvouhodinových intervalů na úřad rovnoměrně?

V tomto případě si stačí uvědomit, že rovnoměrnost příchodu v daných pěti časových intervalech (9-11;11-13;13-15;15-17;17-19) znamená, že teoretické pravděpodobnosti π_j budou ve všech pěti kategoriích stejné, tedy $\pi_j = 20$ %. Tudíž i teoretické četnosti $n\pi_j$ budou stejné. V této souvislosti dodejme, že teoretické četnosti nemusí být celočíselné hodnoty.

MS Excel: Data ze souboru DataExcelchishoda.xls ve tvaru časového údaje hodin a minut musíme nejprve uvést do tvaru, ze kterého jsme schopni sestavit tabulku rozdělení četností. Použijeme funkci CELÁ.ČÁST a tou převedeme časový údaj v hodinách a minutách o příchodu občana na úřad na hodiny. Přesné zadání funkce je: =(A1-CELÁ.ČÁST(A1))*24. Údaj v tomto tvaru již můžeme pomocí funkce ČETNOSTI zapsat do tabulky intervalového rozdělení četností. Protože chceme celkové rozmezí hodin pro veřejnost rozdělit na 5 dvouhodinových intervalů, do pomocného sloupce zapíšeme vždy horní mez intervalu. Dále vedle pomocného sloupce s horními mezemi časových intervalů označíme stejně velké pole a mezi statistickými funkcemi vybereme funkci ČETNOSTI. Objeví se dialogové okno, kam do okna data označíme sloupec se zadanými daty převedenými na hodiny a následně do okna

hodnoty označíme oblast se zadanými horními hodnotami časových intervalů. Na závěr vše potvrdíme trojklikem CTRL+SHIFT+ENTER.

(9	· (· ·) ·						-		DataExcelchish	noda - Micr	osoft Excel	
	9	Domů	Vložen	í Rozlo	ožení stránky	Vzorce	Data	Revize	Zobrazení	Acro	bat			
	Ê	× E	Calibri	* 11	· A A	= = =	\$**	Zalamova	at text		Obecný	-		
	Vložit	3	BIU		<u>> A</u> -			•a• Sloučit a	zarovnat na	střed 🔹	9 000	€,0 ,00 ,00 ⇒,0	Podmíněné formátování	Formátov jako tabul
-	Schrán	ka 🖻		Písmo	G.		Zai	rovnání		G.	Číslo	Fai		Styly
		CHITEST	r -	(• × •	∕ <i>f</i> ∗ =ČET	NOSTI(B2:	3187;E2:E6)						
		А	В	С	D	E	F	G	Н		J	К	L	М
	1 př	íchody	v hodinác	h		horní mez	emp. Čet.		teor. Čet.					
	2	9:03	9,05			10,99999	=ČETNOST		37,2					
	3	9:03:00	9,05			12,99999	40		37,2		rovnoměr	nost přích	odu na úřad	
	4	9:05:00	9,083333			14,99999	27		37,2					
	5	9:08:00	9,133333			16,99999	39		37,2					
	6	9:09:00	9,15			18,99999	44		37,2					
	7	9:11:00	9,183333						186					
	8	9:15:00	9,25											
	9	9:21:00	9,35				p=	0,357463		shoda s	s rovnoměrno	stí příchou	du nelze zan	nítnout
1	10	9:34:00	9,566667											
	6 7 8 9 10	9:09:00 9:11:00 9:15:00 9:21:00 9:34:00	9,133333 9,15 9,183333 9,25 9,35 9,566667			18,99999	p=	0,357463	37,2 37,2 186	shoda s	s rovnoměrno	stí příchou	du nelze zan	nítno

Tímto jsme získali empirické četnosti. Pokud nás zajímá, zda přicházejí občané na úřad dvouhodinových intervalů stejnoměrně, v tomto rozmezí pěti jsou teoretické pravděpodobnosti příchodu náhodného občana na úřad v některém z dvouhodinových intervalů stejné a tedy 0,2. Očekávané četnosti jsou dány součinem 186*0,2=37,2. Na úřad totiž přišlo celkem 186 občanů. Nyní máme data připravena k testování. Ze statistických funkcí v MS Excelu zvolíme funkci CHITEST. Stejně jako v předchozím příkladu získáme informaci o p-hodnotě. V tomto případě je její hodnota 0,357463. Protože tato hodnota je vyšší, než běžné hladiny významnosti, nelze zamítnout nulovou hypotézu. Test nepotvrdil, že by příchody občanů na úřad v rámci pěti dvouhodinových intervalů byly nerovnoměrné.

A 🔒	900	÷ 🗑 🔻			STATI	STICA	Cz - PS1* - [Pozo	prované vs.	očekáva	né četnosti (Tak	oulka2)] - [PS1* -	Pozorované v
Soubor	Domů	Upravit	Zobrazit	Formát	Statistiky		Data mining	Grafy	Nástroje	Data	Sešit	
Základní statistiky	Vícenásobná regrese		4 Neparametrické statistiky Základ	Prokládání rozdělení	Rozdělení a simulace	▲	Pokročilé modely √ícerozm. anal. ▼ Analýza síly testu Pokročilé/Víc	 Reuro PLS, P VEPAG VEPAG 	on. sítě CA, :	Diagramy ř	ízení kvality v 👔 e 🔐 Průmyslová statis	Analýza proc DOE Six Sigma - tika
🧀 PS1* ≟∵👍 Ne	eparametrick	á statistika	(Tabulka2)		<		Pozorované vs Chi-Kvadr. = 4	. očekávan ,376344 sv	é četno = 4 p =	sti (Tabulka2) = ,357465		
ė- 🛵	Dialog poz	orovaných	vs. očekávaný	ch četností			pozoro	JV.		осекау.	P-0	(P-O)^2
	Pozorov	vané vs. oč	éekávané četno	osti (Tabulka2	2) Příp	ad	pozorované	četnosti	teore	etické četnosti		/0
					C:	1		36,0000		37,200	.1,2000	0,038710
					C:	2		40,0000		37,200	2,8000	0,210753
					C:	3		27,0000		37,200	-10,2000	2,796774
					C:	4		39,0000		37,200	1,8000	0,087097
					C:	5		44,0000		37,200	6,8000	1,243011
					Sčt			186,0000		186,000	-0.0000	4.376344

SW Statistica: Stejně jako v předchozím příkladu výstupem v programu SW Statistica bude následující tabulka. Pozorované a teoretické četnosti máme uloženy v souboru data shoda.sta programu Statistica. Pod záložkou statistiky klikneme na záložku Neparametrické statistiky. V základní nabídce zvolíme záložku Pozorované vs. očekávané X².

V dialogovém okně Proměnné vybereme z proměnných proměnnou nazvanou pozorované četnosti a v dialogovém okně proměnné za očekávané četnosti vybereme proměnnou teoretické četnosti. Nakonec klikneme na záložku Výpočet: pozor. vs. oček. četnosti. Objeví se následující tabulka, ve které vidíme hodnotu testového kritéria Chi-Kvadr. (4,376344), stupně volnosti (4) a hlavně p-hodnotu (0,357465), kterou porovnáme s běžnými hladinami významnosti.

Můžeme učinit závěr, že zjištěná data neprokázala (na běžných hladinách významnosti) nerovnoměrnost příchodu občanů na úřad v průběhu úředních hodin pro veřejnost

Příklad 3

Bylo prozkoumáno 25 m²látky a byl zaznamenáván počet kazů vždy na ploše 1 m². Data jsou uvedena v souboru DataExcelchishoda.xls v záložce kazy na látce. Rozhodněte, zda je možno počet kazů na 1 m² látky považovat za náhodnou veličinu, která se řídí Poissonovým rozdělením.

Máme otestovat, zda data pochází ze základního souboru s Poissonovým rozdělením. Poissonovo rozdělení má jeden parametr λ , který je roven střední hodnotě. Tento parametr nemáme zadán, proto ho odhadneme aritmetickým průměrem, což je typický odhad střední hodnoty.

MS Excel: Pomocí funkce Průměr zjistíme, že hodnota aritmetického průměru je 2,52. Následně zapíšeme do sloupce kategorie počtu kazů na 1 m² látky a pomocí funkce četnosti zjistíme empirické četnosti. Podotkněme, že na poslední kategorii je nutno pohlížet jako na kategorii 6 a více kazů. To proto, že nemá smysl uvažovat kategorie pro 7, 8, atd. kazů, které se v našem souboru vůbec nevyskytly, i když víme, že data s Poissonovým rozdělením mají nenulovou pravděpodobnost pro spočetně mnoho kategoriálních nezáporných hodnot. Pro výpočet teoretických četností musíme nejprve vypočítat hodnotu pravděpodobnostní funkce Poissonova rozdělení. Tu vypočteme podle vzorce $P(x)=\lambda^x e^{-\lambda}/x!$ v MS Excelu zadáním =\$A\$22^C2*EXP(-\$A\$22)/FAKTORIÁL(C2), kde v buňce A22 je hodnota průměru a v buňce C2 je první hodnota kategorie. Po výpočtu tohoto příkazu tažením rozkopírujeme a dopočítáme tak pravděpodobnosti odpovídající ostatním kategoriím až na tu poslední. Pravděpodobnost odpovídající poslední kategorii (6 a více kazů) musíme dopočítat jako doplněk do jedničky, tedy od jedné odečteme součet ostatních pravděpodobností. V dalším sloupci vynásobíme všechny hodnoty 25 (počet kontrolovaných metrů látky) a tím vypočítáme teoretické četnosti. Vidíme, že čtyři teoretické četnosti jsou menší než pět a proto musíme sousední kategorie sloučit. Vytvoříme tak kategorie 0-1 kaz, 2 kazy, 3 kazy a 4 a více kazů. Sloučíme také odpovídající teoretické a empirické četnosti a můžeme v nabídce Statistické funkce zvolit funkci CHITEST. Po kliknutí na klávesu Enter se nám zobrazí p- hodnota 0,636. Její vysoká hodnota značí, že hypotézu o shodě s Poissonovým rozdělením nelze zamítnout. Lze konstatovat, že data nejsou v rozporu s předpokladem, že pocházejí ze souboru s Poissonovým rozdělením.

		-	-	-	-				ExcelCihlái	- Microso	ft Excel			
C	Domů Vložer	ní Rozlo	žení stránky	Vzorce	Data	Revize	Zobrazení	Acrol	bat					
VI	Calibri	• 11 •		= = <mark>-</mark> E = 3	· ≫- 1 := :=	Zalamova	it text zarovnat na	střed 🔻	Obecný ∰ - % 000	▼ 00, 00, 00, 00,	Podmíněné formátování	Formáto ▼ jako tabu	vat Styly Iku = buňky =	Vložit (
Sch	ránka 🖻	Písmo	G.		Zai	ovnání		Ga.	Číslo	G.		Styly		
	J12 •	• (0	f _x											
	A B	С	D	E	F	G	н	- I	J	К	L	М	N	0
1	kazy na látce	kategorie	em.četnos	ti	pravděp.	teor.četno	sloučené	lsloučen	é sloučené	emp.				
2	1	0	3		0,08046	2,01149		teor.						
3	0	1	3		0,202758	5,068955	0 až 1	7,08044	45 6					
4	4	2	6		0,255475	6,386884	2	6,3868	84 6					
5	2	3	8		0,214599	5,364982	3	5,36498	82 8					
6	3	4	2		0,135198	3,379939	4 a více	6,1676	89 5					
7	5	5	2		0,06814	1,703489								
8	1	6	1		0,04337	1,084261								
9	6							p =	0,63614	p-hodno	ta je vyšší n	ež běžné h	ladiny význ	amnosti,
10	2									proto ne	zamítíme h	ypotézu o	shodě dat	
11	3									s Poisson	iovým rozdě	élením.		
12	2													
13	3													
14	3													
15	5													
16	0													
17	2													
18	3													
19	2													
20	1													
21	3													
22	3													
23	2													
24	3													
25	4													

SW Statistica: Stejný příklad ukážeme řešený v SW Statistica. Data jsou uvedena v souboru data shoda.sta. Pod záložkou Statistiky zvolíme Prokládání rozdělení a zde zvolíme Poissonovo rozdělení. V záložce Možnosti zaškrtneme Test chí-kvadrát kombinovat kategorie. Pak již jen potvrdíme Výpočet a objeví se nám následující tabulka. V této tabulce jsou vypočítány empirické i teoretické četnosti, je zde uveden odhad parametru λ =2,52 a je zde uvedena hodnota testového kritéria i počet stupňů volnosti. Nakonec je uvedena p-hodnota, která je vyšší než běžné hladiny významnosti a to vede k závěru, že data nejsou v rozporu s předpokladem o tom, že výběr pochází ze základního souboru s Poissonovým rozdělením.

٨		9 0 0	🖶 🗄 🔻			STATISTICA Cz	- [PS1* - Prom	enná: počet kaz	ů, Rozdělení:P	oissonovo, Lan	nbda = 2,520 (Tabul
4	Soubor	Domů	Upravit Z	obrazit Formát	Statistiky	Data mining	Grafy Nás	troje Data	Sešit			
Z	Základní V tatistiky	Vícenásobn regrese	á ANOVA Nepara stat Záklac	metrické Prokládání istiky rozdělení	Rozdělení a simulace	nterne se	Neuron. : M PLS, PCA, UEPAC Prozměrné	sítě 🔜 Diagra 🔛 Multiv	my řízení kval ariate :ive Průmyslov	ity ▼ 📷 Analýza ∰ DOE 6na Six Sigr vá statistika	a procesu 🛄 ma 구	🛃 STA 🔠 Dáv
>		٦	Proměnná: poč Chí-kvadrát = 1	et kazů, Rozdělen 70357 sv = 2 (up	ií:Poissonovo, l rav.) p = 0.420	Lambda = 2,520 (T 665	abulka17)					
_	4					000						
I	Kategor	rie	Pozorovane Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumui. % Pozorované	Ocekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.	Pozorovan Očekáv.	ié -
	Kategor <= 0,00	rie)000	Pozorovane Četnosti	Kumulativ. Pozorované 3 3	Procent Pozorované 12,0000	Kumul. % Pozorované 0 12,0000	Ocekáv. Četnosti 2,011490	Kumulativ. Očekáv. 2,01149	Procent Očekáv. 8,04596	Kumul. % Očekáv. 8,0460	Pozorovan Očekáv. 0,9	ié - 8851
	Kategor <= 0,00 1,00000	rie 0000 0	Pozorovane Četnosti	Kumulativ. Pozorované 3 3 6	Procent Pozorované 12,0000 12,0000	Kumui. % Pozorované 0 12,0000 0 24,0000	Ocekáv. Četnosti 2,011490 5,068957	Kumulativ. Očekáv. 2,01149 7,08045	Procent Očekáv. 8,04596 20,27583	Kumul. % Očekáv. 8,0460 28,3218	Pozorovan Očekáv 0,9 -2,0	ié - 8851 6896
	Kategor <= 0,00 1,00000 2,00000	rie 0000 0	Pozorovane Četnosti	Kumulativ. Pozorované 3 3 6 12	Procent Pozorované 12,0000 12,0000 24,0000	Kumul. % Pozorované 0 12,0000 0 24,0000 0 48,0000	Ocekáv. Četnosti 2,011490 5,068957 6,386884	Kumulativ. Očekáv. 2,01149 7,08045 13,46733	Procent Očekáv. 8,04596 20,27583 25,54753	Kumul. % Očekáv. 8,0460 28,3218 53,8693	Pozorovan Očekáv. 0,9 -2,0 -0,3	ié - 8851 6896 8688
	Kategor <= 0,00 1,00000 2,00000 3,00000	rie 2000 2 2 2	Pozorovane Četnosti	Kumulativ. Pozorované 3 6 5 12 8 20	Procent Pozorované 12,0000 12,0000 24,0000 32,0000	Kumul. % Pozorované 0 12,0000 0 24,0000 0 48,0000 0 80,0000	Ocekáv. Četnosti 2,011490 5,068957 6,386884 5,364981	Kumulativ. Očekáv. 2,01149 7,08045 13,46733 18,83231	Procent Očekáv. 8,04596 20,27583 25,54753 21,45992	Kumul. % Očekáv. 8,0460 28,3218 53,8693 75,3292	Pozorovan Očekáv. 0,9 -2,0 -0,3 2,6	ié - 8851 6896 8688 3502
	Kategor <= 0,00 1,00000 2,00000 3,00000 4,00000	rie 0000 0 0 0 0	Pozorovane Četnosti	Kumulativ. Pozorované 3 3 6 12 3 20 2 22	Procent Pozorované 12,0000 12,0000 24,0000 32,0000 8,0000	Kumul. % Pozorované 0 12,0000 0 24,0000 0 48,0000 0 80,0000 0 88,0000	Ocekáv. Četnosti 2,011490 5,068957 6,386884 5,364981 3,379939	Kumulativ. Očekáv. 2,01149 7,08045 13,46733 18,83231 22,21225	Procent Očekáv. 8,04596 20,27583 25,54753 21,45992 13,51976	Kumul. % Očekáv. 8,0460 28,3218 53,8693 75,3292 88,8490	Pozorovan Očekáv. 0,9 -2,0 -0,3 2,6 -1,3	ié - 8851 6896 8688 3502 7994
	Kategor <= 0,00 1,00000 2,00000 3,00000 4,00000 5,00000	rie 0000 0 0 0 0 0	Pozorovane Četnosti	Kumulativ. Pozorované 3 3 6 12 2 2 2 2 2	Procent Pozorované 12,0000 12,0000 24,0000 32,0000 8,0000 8,0000	Kumul. % Pozorované 0 12,0000 0 24,0000 0 48,0000 0 80,0000 0 88,0000 0 96,0000	Ocekáv. Četnosti 2,011490 5,068957 6,386884 5,364981 3,379939 1,703489	Kumulativ. Očekáv. 2,01149 7,08045 13,46733 18,83231 22,21225 23,91574	Procent Očekáv. 8,04596 20,27583 25,54753 21,45992 13,51976 6,81396	Kumul. % Očekáv. 8,0460 28,3218 53,8693 75,3292 88,8490 95,6630	Pozorovan Očekáv. 0,9 -2,0 -0,3 2,6 -1,3 0,2	eé - 8851 6896 8688 3502 7994 9651

3.1.2 Ověřování normality dat

Jak jsme psali již v úvodu této kapitoly, častým předpokladem použití určité statistické metody je ověření, že data pocházejí ze základního souboru s normálním rozdělením. K tomuto ověření si ukážeme dva možné přístupy. Normalitu dat lze ověřovat jednak grafickými metodami, jednak statistickými testy.

Vše budeme ukazovat na datech v souboru Data_deti_min.sta. Máme k dispozici záznamy o celkem 267 dětech, u kterých jsme zaznamenali jejich třídu, výšku, hmotnost, BMI, známku z tělocviku, pohlaví, věk, jejich vlastní hodnocení oblíbenosti tělocviku, dovednosti v tělocviku a výsledků v určitých sportovních disciplínách.

3.1.2.1 Grafické metody

Nejpoužívanější grafické metody, které si představíme, jsou:

- Histogram
- Q-Q graf
- Pravděpodobnostní graf
- P-P graf

My pro ověření normality vybereme spojité veličiny VYSKA (výška) a HMOTN (hmotnost). Vše ukážeme v SW Statistica.

Histogram je graf, ve kterém na vodorovnou osu vynášíme setříděné hodnoty zkoumané veličiny rozdělené do intervalů a na osu y vynášíme hodnoty absolutních nebo relativních četností v daném intervalu. Pokud máme dostatečný počet hodnot a pokud data pocházejí z normálního rozdělení, histogram by měl kopírovat Gaussovu křivku, která je grafem hustoty normálního rozdělení.

V programu Statistica histogram najdeme v záložce Grafy mezi běžnými grafy. Do proměnné zvolíme proměnnou VYSKA. V záložce kategorie nastavíme počet intervalů, do kterých budou data rozdělena, a zaškrtneme typ rozdělení normální. V záložce detaily můžeme ještě zaškrtnout dva testy Shapiro-Wilkův test a Kolmogorov-Smirnovův test. O těch pohovoříme později. V záložce typ proložení také vidíme, že můžeme histogram nechat proložit i jiným typem grafu, které reprezentují další rozdělení např. Exponenciální, Gama, Beta, Geometrické a další. Stejným způsobem můžeme tedy testovat i jiná rozdělení než rozdělení normální.





Graf histogramu kopíruje Gaussovu křivku a to značí, že data pochází z normálního rozdělení. To potvrzují i výsledné p-hodnoty obou zvolených testů, které jsou zobrazeny v tabulce vlevo dole na grafu histogramu. Okomentujeme je později.

Ukažme si ještě jeden histogram veličiny HMOTN.



Graf histogramu veličiny hmotnost již Gaussovu křivku tak nekopíruje a zdá se, že tato veličina nemá normální rozdělení. To potvrzují i p-hodnoty testů uvedené vlevo dole.

Q-Q graf, neboli kvantil kvantilový graf umožňuje posoudit, zda data pochází ze známého rozdělení. Program SW Statistica umožňuje pomocí tohoto grafu posoudit 8 typů rozdělení. My vše ukážeme na posouzení normality dat. Tento graf na svislou osu vynáší uspořádané hodnoty sledované veličiny a na vodorovnou osu kvantily vybraného (pro nás normálního) rozdělení. Tyto body jsou pak proloženy regresní přímkou (O tomto pojmu se dovíte více v následující kapitole). Čím blíže jsou body o souřadnicích [teoretický kvantil; empirický kvantil] blíže této přímce, tím větší je shoda mezi empirickým a teoretickým rozdělením. V záložce 2D grafy vybereme grafy typu Q-Q. V záložce detaily do proměnné zvolíme VYSKA. V záložce rozdělení vybereme normální a potvrdíme OK.

	STATIST	ICA
Soubor Domů Upravit Zobrazit Vlož	žit Formát Statistiky Data mining Grafy	Se
Histogram Bodový Průměry Krabice Rozptyl Spojnice graf Běžný Grafy typu Q-Q	Image: Set of the set o	r :a
Základní Detaily Vzhled Kategorizovaný Možnos Proměnné: VYSKA	sti 1 Možnosti 2 OK Stomo	
Rozdělení: Normální A Pote Exponenciální A Extr. hodn. A Gama A Lognormální A Rayleighovo A Weibullovo	e Čára proložení V Zobr. lin. prol. Vastní proložení: Měřítko: 1 Práh: 0 Úprava Pořadí: .375 N: .25 N: .25]]]
Pravděp. měřítko: ,01,05,1,25,5,75,9,95,99 K vykreslení kvantilů jedné proměnné vůči kvantilům druhé p Rozložení grafu Více grafů v jednom obr.) proměnné použijte 2D bodové grafy.	



Vidíme, že body se od přímky vzdalují jen pro vysoké kvantily. Další body leží téměř na přímce. Můžeme tedy učinit závěr, že data pochází ze souboru, který má normální rozdělení. Stejný postup provedeme i s proměnnou HMOTN. Získáme následující graf. Vidíme, že v tomto případě je mnohem více bodů mimo přímku. Tento graf normalitu dat nepotvrzuje.



N-P graf umožňuje posoudit pouze, zda data pocházejí z normálního rozdělení. Na vodorovnou osu jsou vynesena vzestupně setříděná data a na svislou osu jsou vyneseny kvantily normovaného normálního rozdělení. Pokud data pocházejí z normálního rozdělení, budou ležet na přímce. Pokud rozdělení nebude symetrické, ale zešikmené na jednu či druhou stranu, data budou tvořit křivku konkávně či konvexně prohnutou.

V záložce Grafy-2D-normální pravděpodobnostní grafy zvolíme typ rozdělení normální a zaškrtneme Shapiro-Wilkův test a potvrdíme OK. V následujícím grafu vidíme, že proměnná VYSKA leží téměř všechna na přímce. Můžeme učinit závěr, že data pochází ze souboru s normálním rozdělením. P-hodnota Shapiro-Wilkova testu na jednoprocentní hladině významnosti tento závěr podporuje.

٨		3 9 9 C) 🖨 🗄 -	·	1.	-	_						STA
s	oub	or Domů	Upravi	t Zob	razit	Vlo	žit	Formát	S	itatistiky	/ Da	ita mining	Grafy
			TI	ф _а п _т	TT I		a 20	•		/latice		🖌 Grafy	bloku dat
ні		uram Bodowi			tyl Spy		😻 ЗІ	O Sekv. 🔻	: : i I	kony		🛱 Grafy	vstupních
	300	graf	riunicity kie	ibice Rozp	Jun Jpn	ojinee	i (o xyz ₹		ategoria	zované •	by Dávk.	(skupin) ai
_	_	_	Běžný								Více		
	² N	ormální p-gra	afy							in the second	P	L B	
Γ	Zá	dadní Vzhled	Kategorizo	ovaný Mo	žnosti 1	1 Mož	nosti 2					ок	
	Ţ	yp grafu:		Pro	měnné:							Stomo	
		Normální		VYSKA								Mažnosti	_
		🔄 Polo-norma 🖸 Bez trendu	Ini									MOZHOSU	
	5										by	Anal. ski	Jp.
		statistiky V Shapiro-Wilk	test	Více	nıgratu e orafů v	ı / iednor	n obrázk	au l			SELECT CRSES	Filtr přípa	adů
	Ľ				-	1					8	Váhy	
	1] Neurčovat prů	im. pozici svá	ázaných po	zorovár	ní						Galerie g	rafů
											Aktu	alizace: Au	to 👻
				N	lormálr	ní p-grat		ΚA					
				I	Data_d	eti_min	20v*26	7c					
	4		· · ·		• •					• •		<u> </u>]
	3											۰ ۲	
											68	5	
oty	2	-									100		
odno	1								000				-
ál. h	0							1000 0					
lorm	0				-	000	000]
ek. r	-1	-			0000	-							-
Õ	2			8000									1
	-2	0	0000										1

160

Pozorovaný kvantil

170

180

190

Pro data HMOTN vypadá graf následovně.

150

140

VYSKA: SW-W = 0,9889; p = 0,0380

-3

-4 130



Body již neleží jednoznačně na přímce a nízká p- hodnota testu normality v tabulce vlevo dole na grafu potvrzuje, že data nepochází z populace s normálním rozdělením.

P-P graf - v případě tohoto grafu na vodorovnou osu vynášíme hodnoty teoretické distribuční funkce a na svislou osu hodnoty empirické distribuční funkce. V grafu je pak vyznačena přímka se směrnicí 1. Čím blíže jsou body o souřadnicích [hodnota teoretické distr. fce; hodnota empirické distr. fce] blíže této přímce, tím větší je shoda mezi empirickým a teoretickým rozdělením. V záložce 2D grafy vybereme grafy typu P-P. V záložce detaily do proměnné zvolíme VYSKA. V záložce rozdělení vybereme normální a potvrdíme OK.



Z grafu vidíme, že většina bodů leží na přímce a můžeme usoudit, že data pochází z normálního rozdělení. Stejný graf sestrojíme pro veličinu HMOTN. Na následujícím grafu





Výhody a nevýhody grafických metod:

- Grafy nám umožní vybudovat intuici, jak data vypadají. Pomohou odhalit chyby / překlepy v zápisu dat (např. váha člověka 1000 kg).
- Grafy mohou naznačit jiný typ rozdělení, než normální. Oproti tomu statistické testy vyjdou jen statisticky nevýznamné, ale jiné rozdělení nenaznačí.
- Grafy mohou naznačit, jestli je normalita dat zamítnuta z důvodu několika extrémních hodnot, či zda se jedná o jiné, než normální rozdělení.
- Zkušenému uživateli tento graf také naznačí, jaký typ transformace původní veličiny by mohl vést k jejímu převodu na veličinu s normálním rozdělením.
- Nevýhodou těchto metod může být, že posouzení grafu není jednoznačné; do jisté míry závisí na zkušenostech statistika. Proto je lepší grafickou metodu doplnit statistickým testem.

3.1.2.2. Statistické testy pro ověření normality dat

Připomínáme, že u všech testů, kterými můžeme testovat normalitu dat (ale i jakékoli jiné rozdělení) mají nulovou (tedy testovanou) hypotézu ve tvaru: data pocházejí ze souboru s normálním (či jiným testovaným) rozdělením dat. Alternativní hypotéza v těchto případech tvrdí: není tomu tak. Data pochází ze souboru, jehož data nemají normální (jiné testované) rozdělení. Z předešlé kapitoly víme, že výsledky testů nám statistický software uvádí ve formě p-hodnoty. Pokud je tato hodnota menší než běžné hladiny významnosti (5 %, 1 %), zamítáme nulovou hypotézu. Pokud je p-hodnota vyšší než běžné hladiny významnosti, nulovou hypotézu nelze zamítnout. U těchto testů tedy nezamítáme hypotézu o tom, že data pochází ze základního souboru, který má normální (jiné testované) rozdělení.

Kolmogorovův-Smirnovův test normality dat

Tento test testuje, zda data pochází z normálního rozdělení se střední hodnotou μ a rozptylem σ^2 , tedy testujeme specifikovaný model. Známe hodnoty teoretické distribuční funkce tohoto rozdělení s těmito parametry a tu porovnáme s hodnotami empirické distribuční funkce. Testovaná hypotéza zní, že data (náhodný výběr) pocházejí z normálního rozložení s danou teoretickou distribuční funkcí. Testová statistika je založena na výpočtu absolutní odchylky empirické a teoretické distribuční funkce. Tento test je velmi vhodný v případě malého souboru dat.

Lillieforsův test normality dat

Jedná se o modifikací Kolmogorova –Smirnovova testu pro případ, že nemáme plně specifikovaný model.

Shapiro-Wilkův test normality dat

Tento test je nejobecněji použitelný test normality. Je vhodný jak pro velké, tak malé soubory dat. Čím více je testová statistika W blíže jedné, tím spíše normalita dat nebude zamítnuta. Její hodnota je v SW Statistica uvedena.

SW Statistica: V nabídce vybereme Statistiky - Základní statistiky-Popisné statistiky a potvrdíme OK. Do proměnné zvolíme testovaná data. Zvolíme záložku Normalita a zaškrtneme K-S & Lillieforsův test normality a Shapiro-Wilkův W test. Dále vybereme Tabulky rozdělení četností. Objeví se výsledná tabulka. Pro data VYSKA a HMOTN jsou uvedeny následující tabulky. Pro data VYSKA výsledky potvrzují normalitu dat. V případě Lillieforsova testu (nemáme specifikovaný model) je p-hodnota < 0,15. Vyšší hodnota než 0,05 nás vede k nezamítnutí nulové hypotézy, která tvrdila normalitu. P-hodnota Shapiro-Wilkova testu p=0,03797 je hraniční. Na 5% hladině významnosti bychom normalitu zamítali, na 1% hladině významnosti normalitu dat nezamítáme. U proměnné HMOTN je tomu naopak. P-hodnoty obou testů jsou nižší než běžné hladiny významnosti a proto oběma testy zamítáme normalitu dat HMOTN.

٨	. I 🛃 '	୨ ୯ 🕞	i 🗐 🔻							STATISTICA	Cz - [PS	3* - Ta
4	Soubor	Domů	Upravit	Zobrazit	Formát		Statistiky	D	ata mining	Grafy Ná	stroje	Data
2 6	Základní tatistiky	icenásobná regrese		Neparametrick statistiky Základ	ké Prokládání rozdělení	Roz sin	dělení a nulace	Po Wi Vi	okročilé modely icerozm. anal. ▼ nalýza síly testu Pokročilé/Víc	 Reuron. PLS, PCA VEPAC Verozměrné 	sítě 🔄	🚆 Diag 🎬 Mult 🕐 Predi
>			(Tabulka čet K-S d=,047 Shapiro-Wil	ností:VYSK/ 39, p> .20; L k W=,98887	A (D illief , p=	ata_deti_ fors p<,1 ,03797	_min) 5				
				Cetnost	Kumulativr	ni	Rel.če	etn.	Kumul. %	Rel.četn.	Kumu	I. %
	Katego	rie			četnost		(platný	ich)	(platných)	všech	všeo	ch
	130,000	00 <x<=140< td=""><td>,0000</td><td>5</td><td></td><td>5</td><td>1,8</td><td>37266</td><td>1,8727</td><td>1,87266</td><td>1</td><td>,8727</td></x<=140<>	,0000	5		5	1,8	37266	1,8727	1,87266	1	,8727
	140,000)0 <x<=150< td=""><td>,0000,</td><td>32</td><td></td><td>37</td><td>11,9</td><td>8502</td><td>13,8577</td><td>11,98502</td><td>13</td><td>,8577</td></x<=150<>	,0000,	32		37	11,9	8502	13,8577	11,98502	13	,8577
	150,000)0 <x<=160< td=""><td>,0000</td><td>83</td><td></td><td>120</td><td>31,0</td><td>8614</td><td>44,9438</td><td>31,08614</td><td>44</td><td>,9438</td></x<=160<>	,0000	83		120	31,0	8614	44,9438	31,08614	44	,9438
	160,000)0 <x<=170< td=""><td>,0000</td><td>94</td><td>:</td><td>214</td><td>35,2</td><td>20599</td><td>80,1498</td><td>35,20599</td><td>80</td><td>,1498</td></x<=170<>	,0000	94	:	214	35,2	20599	80,1498	35,20599	80	,1498
	170,000)0 <x<=180< td=""><td>,0000</td><td>49</td><td>:</td><td>263</td><td>18,3</td><td>35206</td><td>98,5019</td><td>18,35206</td><td>98</td><td>,5019</td></x<=180<>	,0000	49	:	263	18,3	35206	98,5019	18,35206	98	,5019
	180,000)0 <x<=190< td=""><td>,0000</td><td>4</td><td></td><td>267</td><td>1,4</td><td>9813</td><td>100,0000</td><td>1,49813</td><td>100</td><td>,0000</td></x<=190<>	,0000	4		267	1,4	9813	100,0000	1,49813	100	,0000
	ChD			0		267	0,0	00000		0,00000	100	,0000

٨		9 P 🕞 🛛	🖶 🖥 🔻							STATISTICA	Cz - [F	PS3* - Tab
4	Soubor	Domů	Upravit	Zobrazit	Formát		Statistiky	D	ata mining	Grafy Ná	stroje	Data
	(ákladní tatistiky	/ícenásobná regrese		Veparametrick statistiky Základ	ké Prokládání rozdělení	Roz	dělení a nulace	Po Wi Vi	okročilé modely ícerozm. anal. ▼ nalýza síly testu Pokročilé/Víc	 Reuron. PLS, PCA VEPAC verozměrné 	sítě	Diagi
>				Tabulka čet K-S d=,079 Shapiro-Wil	ností:HMOTI 93, p<,10 ; L k W=,97206,	V (D illief , p=	ata_deti fors p<,0 ,00004	_min) 1				
1				Cetnost	Kumulativi	11	Rel.ce	an.	Kumul. %	Rel.četn.	Kur	nul. %
	Katego	rie			četnost		(platný	ich)	(platných)	všech	Vŝ	sech
	20,000	00 <x<=30,0< td=""><td>00000</td><td>1</td><td></td><td>1</td><td>0,3</td><td>37453</td><td>0,3745</td><td>0,37453</td><td></td><td>0,3745</td></x<=30,0<>	00000	1		1	0,3	37453	0,3745	0,37453		0,3745
	30,000	00 <x<=40,0< th=""><th>00000</th><th>54</th><th></th><th>55</th><th>20,2</th><th>2472</th><th>20,5993</th><th>20,22472</th><th></th><th>20,5993</th></x<=40,0<>	00000	54		55	20,2	2472	20,5993	20,22472		20,5993
	40,000	00 <x<=50,0< th=""><th>00000</th><th>88</th><th></th><th>143</th><th>32,9</th><th>5880</th><th>53,5581</th><th>32,95880</th><th></th><th>53,5581</th></x<=50,0<>	00000	88		143	32,9	5880	53,5581	32,95880		53,5581
	50,000	00 <x<=60,0< th=""><th>0000</th><th>76</th><th>2</th><th>219</th><th>28,4</th><th>6442</th><th>82,0225</th><th>28,46442</th><th></th><th>82,0225</th></x<=60,0<>	0000	76	2	219	28,4	6442	82,0225	28,46442		82,0225
	60,000	00 <x<=70,0< th=""><th>0000</th><th>40</th><th>2</th><th>259</th><th>14,9</th><th>8127</th><th>97,0037</th><th>14,98127</th><th></th><th>97,0037</th></x<=70,0<>	0000	40	2	259	14,9	8127	97,0037	14,98127		97,0037
	70,000	00 <x<=80,0< th=""><th>0000</th><th>8</th><th>2</th><th>267</th><th>2,9</th><th>9625</th><th>100,0000</th><th>2,99625</th><th>1</th><th>00,0000</th></x<=80,0<>	0000	8	2	267	2,9	9625	100,0000	2,99625	1	00,0000
1	ChD			0	1	267	0,0	00000		0,00000	1	00,0000

Jak jsme viděli v případě grafických metod, SW Statistica umožňuje Shapiro-Wilkův test zvolit pro ujištění i v případě histogramu či pravděpodobnostního grafu.

Pokud máme dostatečně velký výběrový soubor (n>50), můžeme ověřit normalitu dat i testem chí kvadrát. V SW Statistica ho najdeme pod záložkou Prokládání rozdělení. Zde mezi spojitými rozděleními vybereme (dvojitým klikem) normální rozdělení. V záložce Možnosti u testu chí-kvadrát označíme kombinovat kategorie. Pokud by s daným počtem kategorií (nastavitelným v záložce Parametry) nastal problém s malou teoretickou četností dané kategorie (Připomínáme, že podmínka testu chí-kvadrát je, že v každé kategorii by měla být teoretická četnost větší než 5), sousední kategorie budou sloučeny. Nakonec jen potvrdíme stisknutím záložky Výpočet. Na následujících dvou obrázcích vidíme výsledky testu pro proměnné VYSKA a HMOTN. Na základě p-hodnot (VYSKA p=0,18126 tedy vyšší než běžné hladiny významnosti, HMOTN p=0,00001, tedy nižší než běžné hladiny významnosti.) můžeme konstatovat, že test potvrdil zamítnutí normality u proměnné HMOTN a normalitu nezamítáme u proměnné VYSKA.

AIL 9 C S	Upravit Zo	brazit Formát	Statistiky	STATIS Data mining	TICA Cz - [PS2* Grafy Nástr	- Proměnná: H oje Data	IMOTN, Rozd Sešit	ělení:Normální	(Data_deti_	.min)]
····	1			2 Pokročilé modely z	Neuron sit	ě Diagra	my řízení kvali	tv z 🛄 Analýza	Drocesu	
	$\geq \frac{2}{4}$						niy nzeni kvan		procesu	
Základní Vícenásobn	a ANOVA Neparan	netricke Prokládání	Rozdělení a	B. vicerozm. anal. *	BU PLS, PCA,	. Multiva	ariate	E DOE		Carl Davi
statistiky regrese	statis	stiky rozdělení	simulace 🏻 🖁	Analýza síly testu	UEPAC	🚉 Predict	ive	Did Six Sigr	na 🔻	🔤 Kalk
	Základ			Pokročilé/Více	rozměrné		Průmyslov	á statistika		
>	Proměnná: HMC)TN Rozdělení:N	ormální (Data, d	leti min)						
	Chí-kvadrát = 35	.95460. sv = 7 (u	prav.), p = 0.00	001						
Horní	Fuzurované	Kumulativ.	Procent	Kumul, %	Očekáv.	Kumulativ.	Procent	Kumul. %	Pozorov	ané -
hranice	Četnosti	Pozorované	Pozorované	Pozorované	Četnosti	Očekáv.	Očekáv.	Očekáv.	Oček	áv.
<= 30,00000	1	1	0,37453	3 0,3745	5,57861	5,5786	2,08937	2,0894		4,5786
35,00000	7	8	2,62172	2,9963	10,89332	16,4719	4,07990	6,1693		3,8933
40,00000	47	55	17,60300	20,5993	22,99746	39,4694	8,61328	14,7825	2	4,0025
45,00000	44	99	16,47940	37,0787	38,18857	77,6580	14,30284	29,0854		5,8114
50,00000	44	143	16,47940	53,5581	49,88241	127,5404	18,68255	47,7679		5,8824
55,00000	47	190	17,60300	71,1610	51,25496	178,7953	19,19661	66,9645	-	4,2550
60,00000	29	219	10,86142	82,0225	41,42860	220,2239	15,51633	82,4809	-1	2,4286
65,00000	27	246	10,11236	92,1348	26,34082	246,5648	9,86548	92,3464		0,6592
70,00000	13	259	4,86891	97,0037	13,17345	259,7382	4,93388	97,2802	-	0,1735
75,00000	5	264	1,87266	98,8764	5,18174	264,9199	1,94073	99,2210		-0,1817
80,00000	3	267	1,12360	100,0000	1,60292	266,5229	0,60034	99,8213		1,3971
< Nekonečno	0	267	0,00000	100,0000	0,47713	267,0000	0,17870	100,0000		-0,4771
AIL タマの Soubor Domů	Upravit Zo	brazit Eormát	Statistiky	STATI Data mining	STICA Cz - [PS2' Grafy Nástr	* - Proměnná: ' oje Data	VYSKA, Rozdě Sešit	ilení:Normální (Data_deti_r	nin)]
		\sim		🕺 Pokročilé modely 🕇	r 💐 Neuron. sít	tě 🚟 Diagra	my řízení kval	ity 🔻 📠 Analýza	a procesu	🔡 STAT
7	1 1			🕻 Vícerozm. anal. 🔻	🗭 PLS, PCA,	. Multiv	ariate			by Dáv
Základní Vícenásobn	iá ANOVA Neparan	netrické Prokládání	ozdělení a	& Analýza sílv testu	VEPAC	2 Predict	tive		na •	Kalk
statistiky regrese	Základ	stiky Tozdeleni	siniulace 🚽	Polyro čilá Vícy	rozmárná		Drůmuslov	CHAIC SIN SIG.		[<u></u>]
-	Zakiad			Pokrocile/vice	rozmerne		Prumysio	a statistika		
>	Proměnná: VYS Chí-kvadrát = 8,1	KA, Rozdělení:N 86598, sv = 6 (up	ormální (Data_d rav.) , p = 0,181	leti_min) 126						
Horní	Pozorovane	Kumulativ.	Flocent	Kumul. %	Očekáv.	Kumulativ.	Procent	Kumul. %	Pozorov	ané -
hranice	Cetnosti	Pozorované	Pozorované	Pozorované	Cetnosti	Očekáv.	Očekáv.	Očekáv.	Oček	áv.
<= 135,00000	1	1	0,37453	3 0,3745	0,82078	0,8208	0,30741	0,3074	0	0,17922
140,00000	4	5	1,49813	3 1,8727	2,58348	3,4043	0,96760	1,2750	1	,41652
145,00000	9	14	3,37079	5,2434	7,81511	11,2194	2,92701	4,2020	1	,18489
150,00000	23	37	8,61423	3 13,8577	18,39485	29,6142	6,88946	11,0915	4	,60515
155,00000	32	69	11,98502	25,8427	33,69278	63,3070	12,61902	23,7105	-1	1,69278
160,00000	51	120	19,10112	44,9438	48,02779	111,3348	17,98794	41,6984	2	2,97221
165,00000	57	177	21,34831	66,2921	53,28252	164,61/3	19,95600	61,6544	3	5,71748
170,00000	37	214	13,85768	80,1498	46,00653	210,6238	17,23091	78,8853	-9-	9,00653
175,00000	28	242	10,48689	90,6367	30,91633	241,5402	11,57915	90,4645	-2	2,91633
180,00000	21	263	/,8651/	98,5019	16,16838	257,7086	0,05557	96,5201	4	1,03162
165,00000	4	267	1,49813	100,0000	0,5/90/	264,2684	2,46437	98,9844	-2	2,5/98/
< ivekonech0	U	267	0,00000	100,0000	2,71156	207,0000	1,01557	100,0000	-2	2,71158

3.2 Vyšetřování závislosti dvou kategoriálních náhodných veličin

V této části budeme vyšetřovat závislost dvou kategoriálních veličin. Tyto veličiny mohou být nominální, ordinální i kvantitativní. Naučíme se měřit sílu závislosti, která může být symetrická, či asymetrická. K tomuto zjišťování bude třeba mít data uspořádána v tabulce rozdělení četností podle dvou znaků.

3.2.1 Kontingenční tabulky

Tak jako jsme zadaná data roztřídili do tabulky rozdělení četností podle určitého znaku, stejně je dokážeme roztřídit i do tabulky rozdělení četností podle dvou znaků. Vznikne nám dvojrozměrné rozdělení četností. Oba znaky mohou být kategoriální a pak vzniklé tabulce říkáme kontingenční tabulka. Pokud oba znaky jsou kvantitativní diskrétního typu, říkáme vzniklé tabulce korelační tabulka. Ukažme si vytvoření tabulek v MS Excel a SW Statistica a naučíme se rozumět hodnotám v nich. Pracovat budeme se souborem Data_deti_min.sta a

Data_deti.xls. V souboru bylo sledováno 267 dětí, u kterých jsme zjišťovali věk, navštěvovanou třídu, známku z tělocviku, pohlaví, jejich výšku, hmotnost, BMI, jejich výkony v různých disciplínách (skok daleký, lehsed, a další) a jejich bodové ohodnocení v těchto disciplínách.

MS Excel: Vytvoření kontingenční tabulky v MS Excel je velmi jednoduché a užitečné. Klikneme kurzorem na záložku Data a následně pod záložkou Vložit klikneme na Kontingenční tabulka. Objeví se nám dialogové okno, které stačí potvrdit.

	•	5	Data_	deti - Mic	rosoft Excel			Nástroje	e kontinger	nční tabulky	-							×
Domů Vlo:	ení Rozlož	ení stránky	Vzorce	Data	Revize	Zobrazení	Acrol	bat Možn	osti	Návrh							0	- = x
Název kontingenční tabu Kontingenční tabulka 4 Možnosti ~ Kontingenční tabulka	ky: Aktivní po	le:	Rozbalit cel Sbalit celé p pole	lé pole	 Výběr skup Oddělit Skupinové Skupina 	pole Se	AZA ZA Seřadit eřadit	Aktualizovat ZC Data	Změnit droj dat •	Vymazat Vyl	brat Přesunou Akce	t Konting	enční Vzoro f v Nástroje	Nástroje OLAP *	Seznam Ti polí Zobra	+ ačítka +/- zit či skrýt		
A3	- (9	f _x																*
A B Choice - i vytvöft si pole ze seza kontingenční Choete-i vytvöft si pole ze seza kontingenční 10 11 11 13 14 15 16 17 18 19 20 21 22 23 24 26 27 28	C abulka 4 stawu, zvolte mu polí abulky.	D	E	F	G	H			ĸ		M	N		P	Q =	Seznam poli konti Zvolte pole, které 4 sestavy: Več 220 mic 221 mic 2	rgenčni tabulky hote přidat do	x x x x x y sloupc0 x y x y y y sloupc0 x y y y y y y y y y y y y y
29 30 I4 4 ▶ ▶I List2 ∠List	1 / List3 / L	ist4 List5	Datad	eti_min 🔬	2				1						▶ []	Odložit aktualiz	aci rozlo Akt	tualizovat

Můžeme si tu vybrat oblast, kam se má kontingenční tabulka zobrazit (přednastaveno je na nový list) a jak vypadá oblast dat. Kliknutím na OK se nám objeví následující list s označením pole, do kterého se zapíše kontingenční tabulka, napravo se objeví seznam jednotlivých proměnných v souboru a dále jsou to čtyři pole. Do pole s názvem Popisky řádků stáhneme proměnnou POHL. Jednotlivé kategorie této proměnné budou ve vznikající kontingenční tabulce jednotlivými řádky. Do pole s názvem Popisky sloupců přetáhneme proměnnou ZNAM. Její kategorie budou tvořit jednotlivé sloupce v kontingenční tabulce. Nakonec jednu z těchto proměnných (u nás ZNAM) přetáhneme do spodního pole s názvem \sum hodnoty. Kliknutím na proměnnou se rozbalí nabídkové menu a z něj klikneme dole na Nastavení polí hodnot. Po kliknutí se rozbalí další nabídka. Zvolte typ kalkulace, a z něj vybereme Počet. Po odkliknutí se nám vlevo na listě objeví následující tabulka.

Počet z známka	známka			
				Celkový
pohl.	1	2	3	součet
0	59	32	5	96
1	135	30	6	171
Celkový součet	194	62	11	267

Pro znázornění, že je jedno, zda naše proměnné budou kvalitativní (slovní) či zapsané pomocí číselných hodnot, jsme proměnnou Pohlaví, ve které značila 0 ženu a 1 muže, přepsali do slovních hodnot a proměnnou známka jsme přepsali do proměnné známka z těl. slovně (1-výborně, 2-chvalitebně,3-dobře). Vidíme, že obě tabulky jsou totožné, jen kategorie sloupců a řádků jsou jinak zapsané. Nakonec pravým tlačítkem klikneme na tabulky a v záložce Možnosti kontingenční tabulky pod záložkou Zobrazit zaškrtneme klasické rozložení kontingenční tabulky a potvrdíme OK.

Počet z ZNÁMKA Z TĚL	ZNÁMKA Z TĚL			Celkový
POHLAVÍ	VÝBORNĚ	CHVALITEBNĚ	DOBŘE	součet
ŽENA	59	32	5	96
MUŽ	135	30	6	171
Celkový součet	194	62	11	267

Nyní se zamyslíme nad hodnotami v tabulce. Nejprve se soustředíme na šest hodnot uvnitř tabulky rozložených do dvou řádků po třech hodnotách. Každá tato hodnota se vztahuje zároveň k určité kategorii řádku (muž, žena), či sloupce (výborně, chvalitebně, dobře). Například se z tabulky dovídáme, že žen, které měly jedničku z tělocviku, bylo v souboru 59, nebo že v souboru bylo šest chlapců, kteří měli z tělocviku trojku. Tedy každá hodnota uvnitř tabulky nás informuje o počtu jednotek v souboru, které mají hodnotu jednoho znaku odpovídající danému řádku a druhou odpovídající součtu daného řádku či sloupce s názvem Celkový součet, je informace odpovídající součtu daného řádku či sloupce. Informuje nás vždy o počtu jednotek v daném souboru, které přísluší vždy do jedné kategorie znaku v řádcích nebo znaku ve sloupcích. Tyto hodnoty se tedy vztahují vždy jen k jedné proměnné (jednomu znaku). Z naší tabulky například vyčteme, že v našem souboru je 96 žen a 171 mužů a také, že je v našem souboru 194 jedničkářů, 62 dvojkařů a 11 trojkařů z tělocviku. Poslední číslo uvedené v tabulce vpravo dole je součet posledního řádku i posledního sloupce a musí být rovno rozsahu souboru, tedy v našem případě počtu sledovaných dětí.

Ukažme si ještě, že data v kontingenční tabulce můžeme mít zachyceny také pomocí relativních četností. Při vytváření kontingenční tabulky stejně jako v předchozích případech do pole s názvem Popisky řádků stáhneme proměnnou POHL. Do pole s názvem Popisky sloupců přetáhneme proměnnou ZNAM. Nakonec jednu z těchto proměnných (u nás ZNAM) přetáhneme do spodního pole s názvem Σ hodnoty. Kliknutím na proměnnou se rozbalí nabídkové menu a z něj klikneme dole na Nastavení polí hodnot. Nyní klikneme na záložku Zobrazit hodnoty jako a z nabídnutého seznamu vybereme % řádku. Vše vidíme na následujícím obrázku.

C .,		9 - C				Data	_deti - N
	Dom	nů V	ložení l	Rozložení s	tránky	Vzorce	Dat
ľ	₿		* 1	0 • A	Ă	= = =	
VIo	žit 🍼	BI	<u>u</u> - <u>m</u>	- 🗠	<u>A</u> -		
Schrä	ánka 🖻		Písmo		Fa.		
	C5		-	f _x	33,33	3333333333	333%
	/	Д	В	(C	D	
1 (Nastave	ní polí ho	odnot			? -	×
2	-	1	4.1				
3	Nazev z	droje: zr	lamka				
5	<u>V</u> lastní r	název: P	očet z známk	a			
6	Souhr	n Zobr	azit hodnoty	iako			
7	Joodin						
8	Zobra	azit hodr	ioty jako				
9	% řá	dku					-
10	Základ	Iní pole:		Základr	ní položi	ka:	
12	pohl.			*			~
13	znám	ka					
14							
15				-			-
16							
17							
18	Eormá	t čísla		(ОК	Storn	•
19							

Po odkliknutí se nám vlevo na listě objeví následující tabulka.

Počet z známka	známka			
pohl.	1	2	3	Celkový součet
0	61,46%	33,33%	5,21%	100,00%
1	78,95%	17,54%	3,51%	100,00%
Celkový součet	72,66%	23,22%	4,12%	100,00%

Z této tabulce vidíme, že mezi ženami bylo 61,46 % jedničkářek z tělocviku, 33,33 % dvojkařek z tělocviku a 5,21 % trojkařek z tělocviku. Mezi muži, bylo 78,95 % jedničkám z tělocviku, 17,54 % dvojkařů z tělocviku a 3,51 % trojkařů z tělocviku. Ještě lépe je procentuální rozložení známek u jednotlivých pohlaví lépe vidět na následujících koláčových grafech.



Nakonec ukažme vytvoření kontingenční tabulky v SW Statistica.

SW Statistica: Vytvoření kontingenční tabulky je v SW Statistica velmi jednoduché. Pod záložkou Statistiky zvolíme Základní statistiky. V nabídce zvolíme Kontingenční tabulky a potvrdíme OK. Kliknutím na záložku Specif.tabulky (vyberte proměnnou) se otevře dialogové okno s názvem Vyberte 2 seznamy proměnných (faktorů) do tabulky. Z prvního seznamu označíme z nabídky proměnných proměnnou POHLA a z druhého seznamu proměnnou ZNAM a potvrdíme OK. A následně znovu OK. Nakonec klikneme na Výpočet souhrn tabulek a objeví se následující tabulka. V této tabulce jsou červeně vyznačeny četnosti větší než deset. Tabulka je stejná, jako tabulka z MS Excel a samozřejmě také interpretace dat z tabulky je stejná.



٨		9 0	× (+				
2	Soubor	Do	omů (Upravit		Zobrazit	Formát
	ákladní tatistiky	Vícena reg	ásobná A irese		1 Nepar st	rametrické atistiky	Prokládání I rozdělení
>	Konting Četnos (Maroir	genčr st ozn nální s	ní tabulka ačených součtv né	a (Data buněl eisou	a_deti k > 10 označ	i_min) 0 čenv)	
I	POH	LA	ZNAM 1	ZN	AM 2	ZNAM 3	Řádk. součty
	0		- 59	9	32	5	96
	1		135	5	30	6	171
	Vš.sku	ID.	194	1	62	11	267

3.2.2 Testy nezávislosti

K vyšetřování vzájemné závislosti dvou kategoriálních znaků se používá často *test chíkvadrát nezávislosti*. V tomto testu vycházíme z kontingenční tabulky a testujeme odlišnost empirických (napozorovaných, aktuálních) a teoretických četností. Teoretické četnosti vycházejí z předpokladu nezávislosti. Pokud jsou dva zkoumané znaky nezávislé, pak by rozdělení v každém řádku (resp. sloupci) mělo být ve stejném poměru jako v součtovém řádku (resp. sloupci). Teoretické četnosti v každém vnitřním poli kontingenční tabulky tudíž vypočteme jako součin sloupcového a řádkového součtu, vydělený celkovým rozsahem výběrového souboru. Poznamenejme, že ať uvažujeme jakýkoli test nezávislosti, platí následující.

Nulová hypotéza v testech nezávislosti je: veličiny jsou nezávislé

Testové kritérium je založeno na porovnání teoretických četností (jaké by měly být četnosti, kdyby dva znaky byly nezávislé) a empirických četností, které známe z výběrového souboru. Testová statistika má rozdělení chí-kvadrát a odtud jméno testu. Připomeňme, že předpokladem testu jsou dostatečně velké teoretické četnosti (Alespoň v 80 % musí být větší než 5 a všechny musí být větší než 1.)

MS Excel: V tomto SW máme k dispozici mezi statistickými funkcemi funkci CHITEST, kterou již známe z testování shody rozdělení. Tato funkce nám poslouží i nyní. Nejprve však musíme připravit data. Budeme vycházet ze souboru Data deti.xls a vyšetříme, zda známka z tělocviku je ovlivněna pohlavím, neboli zda známka z tělocviku závisí na pohlaví. Pro tyto dva znaky jsme si již sestavili kontingenční tabulku a z ní vyjdeme. Do stejného listu, kde máme kontingenční tabulku, opíšeme četnosti včetně posledního sloupce a řádku. Toto budou naše aktuální (empirické četnosti). Pod tvto aktuální četnosti budeme psát teoretické, tedv očekávané četnosti. Jejich tabulka bude mít stejný rozměr. Ty spočteme z posledního řádku a posledního sloupce tabulky aktuálních četností. Zadáme = a označíme první buňku sloupce celkem tabulky aktuálních četností a zafixujeme ji stisknutím klávesy F4, následně stiskneme klávesu součinu a označíme první buňku posledního sloupce tabulky aktuálních četností a zmáčknutím klávesy lomeno (děleno) a označením poslední buňky posledního řádku a po jejím zafixování klávesou F4, můžeme stisknout klávesu ENTER. V zápise funkce máme =\$F\$12*C14/\$F\$14. Uchopením pravého dolního rohu buňky roztáhneme daný vzorec na celý řádek. Stejně budeme postupovat i pro druhý řádek. Takto jsme vypočítali očekávané četnosti. Nepřekvapí nás, že teoretické četnosti nemusí být přirozená čísla. Pro kontrolu můžeme, pomocí funkce SUMA zkontrolovat, zda součty každého řádku a sloupce teoretických četností jsou stejné jako součty odpovídajících řádků a sloupců u aktuálních četností. Zkontrolujeme ještě předpoklad testu. Jedna z četností vyšla menší než 5, ale je splněna podmínka, že z 80 % jsou teoretické četnosti větší než 5. Vše je zachyceno na následujícím obrázku. Pak již stačí do určité buňky zadat funkci CHITEST. Objeví se okno, kam musíme do části Aktuální označit pole našich aktuálních četností (bez řádku a sloupce celkem). Do okna Očekávané označíme pole teoretických četností. A potvrdíme OK. Objeví se číslo, které je p-hodnota. Je-li toto číslo menší než běžné hladiny významnosti (<0,1),

zamítáme nulovou hypotézu (nezávislost). Naše hodnota je 0,008147 a ta nás vede k závěru, že oba znaky se ovlivňují.

	a) 🖬 🤊 - (H -) =		-	_	Data	a_deti - Micr
C	Domů Vložení	Rozložení stránky	Vzorce Data	Revize Zobrazen	í Acrobat	
				≥	at surf	
	Arial 🔹	10 • A A =		⊒ ⁴ Zalamovat text	Obecny	
VI	ožit 🧹 🖪 I 🗓 🗸	🛛 • 🔕 • 🗛 •		🔤 Sloučit a zarovnat n	a střed 👻 🕎 👻 🧌	/0 000 500
Sch	ránka 🗟 Písmo	0 5	Zar	ovnání	G (Číslo
	l13 - (*	f_{x}				
	А	В	С	D	E	F
1						
2	- · · · · · · · · · · · · · · · · · · ·					
3	Počet z ZNAMKA Z TEL	ZNAMKA Z TEL 💌	CUMALITEDNĚ	DODOT	C	
4		VIBURNE	CHVALITEBNE 22	DOBRE	Celkovy soucet	
6	ZENA MUŽ	135	30	6	171	
7	Celkový součet	194	62	11	267	
8						
9						
10			aktuální (empirické)	četnosti		
11						celkem
12			59	32	5	96
13			135	30	6	171
14		Ceikem	194	62	11	267
16			očekávané/teoretick	é) četnosti		celkem
17			69.75280899	22.29213483	3,95505618	96
18			124,247191	39,70786517	7.04494382	171
19		celkem	194	62	. 11	267
20						
21						
22						
23			p=	0,008146694	zamítáme H ₀	

V předchozím, jsme si ukázali, že kontingenční tabulku, můžeme také zapsat pomocí relativních četností. Měli jsme v každém řádku tabulky vyjádřeno procentuální zastoupení jednotlivých známek z tělocviku u žen i u mužů. Situaci jsme si ukázali i graficky pomocí koláčových grafů. Stejně můžeme postupovat i s tabulkou očekávaných četností. Tabulka **očekávaných** četností v procentním vyjádření je následující.

	očekávané(teoretické) četnosti v %			
	1	2	3	celkem
ženy	72,65918	23,22097	4,11985	100
muži	72,65918	23,22097	4,11985	100

Tuto tabulku jsme vytvořili z tabulky očekávaných četností tak, že každou očekávanou četnost jsme podělili řádkovým součtem a vynásobili 100. Vidíme, že procentuální zastoupení jednotlivých známek v případě mužů a žen je stejné. Ještě názorněji to ukazují koláčové

grafy procentuálního zastoupení jednotlivých známek z tělocviku u mužů a žen. Tato tabulka očekávaných četností tedy opravdu zachycuje situaci, kdy známka z tělocviku nezávisí pohlaví.



Nyní si ukažme použití testu Chi-kvadrát nezávislosti na našich datech v MS Statistica.

SW Statistica: V záložce Statistiky vybereme Popisné statistiky->Kontingenční tabulky->Specific.tabulky (vyberte proměnné). Zde do jedné proměnné vložíme proměnnou POHL a do druhé dáme proměnnou ZNAM a potvrdíme OK a znovu OK.

A	or Domů	Vpravit	Zobra	zit For	rmát S	statistiky	STATISTICA C2 - Data_deti_min y Data mining Grafy Nästroje Data	R Feature Fin
Záklao statist	dní Vícenásob iky regrese	aná ANOVA	1 2 Neparametr statistik Základ	rické Proklá y rozdě	idání Rozd Elení sim	ð lélení a Julace	Mathematical Statistics Mathematical Statistics Mathematical Statistics Mathematical Statistics Mathematical Statistics Nástroje	
)ata: Data_de	ti_min (20s k	rát 267ř)					
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17	1 2 TRIDA VE 7 7 7 8 8 8 8 8 8 8 8 8 8 8 7 7 7 7 8 8 7 7 7 7 8 8 7 7 7 7 7 7 7 8 8 7 7 7 7 8 7 7 7 8 7 7 7 7 8 7	R POHLA 13 (1)	4 ZNAM V 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 2 2 0 2 2 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1	5 /YSKA H 160 165 172 169 159 162 156 158 159 174 165 166 160 163 164 155	6 MOTN O 52 57 60 55 51 44 45 3 49 49 67 51 55 55 50 48 48 46 47	7 0 0 0 0 0 0 0 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1	8 9 10 11 12 13 14 15 16 17 18 19 20 DOVER Výsledky kontingenčni tabulky: Data deti_min 0 20,3 22,3 22,3 22,3 22,0,9 1 20,3 22,0,9 1 20,3 22,0,9 1 20,3 22,0,9 1 20,3 1 20,2 1 20,3 1 20,2 1 1 20,3 1 20,2 1 1 20,3 1 20,2 1 1 20,3 1 20,2 0 1 6 1 1 20,3 1 20,2 0 1 6 1 1 20,3 1 1 0 21,6 1 </td <td></td>	
18 19	7	13 (14 () 1	160 165	50 52	1	1 1 146 2 30 2 12,0 3 4,5 2 9 1 19,5 0 1 159 2 26 2 15,9 1 1,5 1 6 0 19,1	

V záložce Možnosti zaškrtneme Očekávané četnosti a Zvýraznit četnosti >5. A znovu potvrdíme Výpočet. Výsledkem bude tabulka očekávaných četností, v jejímž záhlaví je hodnota testové statistiky, počet stupňů volnosti (počet řádků-1) krát (počet sloupců-1) a výsledná p-hodnota. Její hodnota je 0,008147, menší než běžné hladiny významnosti a proto nulovou hypotézu o nezávislosti znaků pohlaví a známka zamítáme. Test prokázal závislost.

A 🔛	201	o 🖨 🗒	-	-	
Soubor	Domi	ů Upra	vit Zobra:	zit Formá	t Statistiky
\checkmark	/			\approx	
Základní statistiky	Vícenáso regres	bná ANOV je	A Neparametr statistiky	ické Prokládá / rozděler	ní Rozdělení a ní simulace
			Základ		
			Zakiau		
Souhrn Četnos Pearso	ná tab.: t označ nův chí-	Očekávar ených bun kv. : 9,620	ié četnosti (ľ ěk > 5 29, sv=2, p=	Data_deti_mi ;,008147	n)
Souhrn Četnos Pearso POH	ná tab.: t označ nův chí- LA	Očekávar ených bun kv. : 9,620 ZNAM	ié četnosti ([ěk > 5 29, sv=2, p= ZNAM	Data_deti_mi -,008147 ZNAM	n) Řádk.
Souhrn Četnos Pearso POH	ná tab.: t označ nův chí- LA	Očekávar ených bun kv. : 9,620 ZNAM 1	ié četnosti (ľ ěk > 5 29, sv=2, p= ZNAM 2	Data_deti_mi ,008147 ZNAM 3	n) Řádk. součty
Souhrn Četnos Pearso POH	ná tab.: t označ nův chí- LA	Očekávar ených bun kv. : 9,620 ZNAM 1 69,7528	té četnosti (ľ ěk > 5 29, sv=2, p= ZNAM 2 22,29213	Data_deti_m -,008147 ZNAM 3 3,95506	n) Řádk. součty 96,0000
Souhrm Četnos Pearso POH	ná tab.: t označ nův chí- LA	Očekávar ených bun kv. : 9,620 ZNAM 1 69,7528 124,2472	té četnosti (ľ ěk > 5 29, sv=2, p= ZNAM 2 22,29213 39,70787	Data_deti_mi .008147 ZNAM 3 3,95506 7,04494	ň) Řádk. součty 96,0000 171,0000

Pokud prokážeme závislost testem chí-kvadrát, má smysl se ptát po síle závislosti. Závislost může být symetrická či asymetrická (ptáme se, jak silně se proměnné ovlivňují vzájemně, či jak silně závisí jedna proměnná na druhé). Pokud budeme mít obě proměnné nominální, můžeme sílu závislosti měřit různými koeficienty, jejichž základem je statistika chí-kvadrát. Tyto míry závislosti vždy nabývají hodnoty z nějakého intervalu, jejichž spodní hranice je nula, kterou tyto koeficienty nabývají v případě nezávislosti. Čím více se hodnota koeficientu blíží jeho horní hranici, tím silnější závislost je. Jednou z takových měr je například

Pearsonův kontingenční koeficient. Nabývá hodnot z intervalu $\left\langle 0, \sqrt{\frac{(q-1)}{q}} \right\rangle$, kde

 $q = min \{r, s\}, r$ je počet řádků kontingenční tabulky a *s* je počet sloupců kontingenční tabulky. Další symetrickou mírou závislosti jsou koeficient *Fí* nebo koeficient *Cramerovo V*, které nabývají hodnot z intervalu $\langle 0,1 \rangle$. Pokud počet řádků nebo sloupců kontingenční tabulky je roven dvěma, koeficient *Cramerovo V* je roven koeficientu *Fí*.

V SW Statistica najdeme tyto míry v záložce Detailní výsledky. Celá cesta je následující: Statistiky -> Základní statistiky -> Kontingenční tabulky -> OK -> Specif. tabulky (vyberte proměn.) -> OK -> OK -> na záložce Možnosti vybereme například Pearsonův & M-V chíkvadrát nebo Fí(2x2)& Cramerovo V & C -> na záložce Detailní výsledky kliknout na Detailní 2-rozm. Tabulky. Pro naše proměnné POHLA a ZNAM, dostáváme následující výsledky.

	1 19 (P G	ê 🖥 🔻				-	-		-		ST	TATISTI	ica cz - I	Data_det	i_min						
Soub	or D	omů	Upravit	Zob	razit	Formát	Statistiky	Da	ta mining	Grafy	Nást	troje	Dat	ta								
Záklac statisti	lní Vícer iky re	lásobná grese		1 2 Aeparamo statist Základ	etrické Pro iky ro	okládání R zdělení	ozdělení a simulace	Pol	kročilé mo erozm. ani alýza síly t Pokročile	dely ▼ al. ▼ 🔞ů estu 💷 é/Vícerozm	Neuron. s PLS, PCA, VEPAC ěrné	itě [Diag	gramy říz Itivariate dictive Pr	ení kvalit ůmyslová	y ▼ 🔛 Ar ∰ D 6 g Si i statistika	nalýza proce DE x Sigma v	esu	BTATIST Dávk. a Kalkulá	iCA VB nalýza (c tory * Ná	ile ski	upin) e
	ata: Dat	a_deti_r	min (20s kr	át 267ř)																		
			·		·	·	· · · ·				,					-						
	1 TRIDA	2 VEK	3 POHLA 0	4 <u>ZNAM</u> 1	5 VYSKA 160	6 HMOTN 52	7 OBLIB 1		9 🔀 Výsle	10 edky; kont	11 ingenční t	1 tabulk	2 y: Data	13 _deti_mi	14 n	15	16	17	18 X		9 <u>ON</u> 0	20 BMI 20.3
23		7 13 3 13	0	1	165 172	57 60	1	1	Základ	lní výsledky	Detailní	í výsled	lky Ma	ožnosti					Výpočet		2	20,9 20,3
4 5		3 13 3 13	0	1	169 159	55 51	1	1	- Výpo	očet tabulek	. 5		St	atistiky de	tailních 2	-rozměrnýc chí Jovadra	ch tab.	5	itomo		1 1	19,3 20,2
6 7		7 13 3 13	0	1	162 156	44 53	1	1	▼ 2)čekávané	četnosti			Fisher ex	kakt., Yate	es, McNen	nar (2 x 2)		ložnosti▼		0	16,8 21,8
89		3 13 7 13	0	1	158 159	49	1	- 1	E F	Reziduální č	etnosti Ikového po	čtu		Fí (tabul	ky 2x2) &	Cramérovo & tau-c	V&C	Výpoče	vnal.skup et max.		1	19,6 19,4
10		7 13 7 13	0	1	174	67 51	1	1	P	rocenta z p	očtu v řád	ku		Goodma	novo-Kru	skalovo ga	ma	věroho analýza ví cero:	d. chí-kv. a a změrných		1	22,1 18,7
12 13		3 13 7 13	0	2	166 160	55	1	Ċ	P	rocenta z p	očtu ve slo	oupci		Spearma	anova kor vo d	elace		tabulek Log-lin	: četností vi eární modul	2	1	20,0 21,5
14 15		7 13 3 13	0	1	163 164	50 48	1	1						Koeficier	nty nejisto	ty					1	18,8 17,8
16 17		7 13 7 13	0	1	156 155	46	1	0		_	_	_	_			_	_		_		1	18,9 19,6
18	1	7 13	0	1	160	50	1	1	1	146	6 2	2	30	2	12,0		3 4,5		2	9	1	19,5

٨		900	÷ 🗄 -				
4	Soubor	Domů	Upravit	Zobrazit	Formát	Statistiky	Dat
Z	(ákladní tatistiky	Vícenásobná regrese	á ANOVA N	aparametrické P statistiky áklad	rokládání rozdělení	Rozdělení a simulace	<mark>₩</mark> Pok ₩ Více Více
>				Statist. : PO	HLA(2) x	ZNAM(3) (Da	ata deti
>	Statist			Statist. : PO Chí-kvadr.	HLA(2) x sv	ZNAM(3) (Da p	ata_deti
>	Statist Pears	onův chí-kv	v.	Statist. : PO Chí-kvadr. 9,62028	HLA(2) x sv df=2	ZNAM(3) (Da p p=,00815	ta_deti
>	Statist Pearso M-V cł	onův chí-kv ní-kvadr.	v.	Statist. : PO Chí-kvadr. 9,62028 9,38771	HLA(2) x sv df=2 df=2	ZNAM(3) (Da p p=,00815 p=,00915	ta_deti
>	Statist Pearso M-V ch Fí	onův chí-kv ní-kvadr.	v.	Statist. : PO Chí-kvadr. 9,62028 9,38771 ,189818	HLA(2) x sv df=2 df=2 4	ZNAM(3) (Da p p=,00815 p=,00915	ata_deti
>	Statist Pearso M-V ch Fí Konting	onův chí-kv ní-kvadr. genční koef	v.	Statist. : PO Chí-kvadr. 9,62028 9,38771 ,189818 ,186488	HLA(2) x sv df=2 df=2 df=2	ZNAM(3) (Da p p=,00815 p=,00915	ata_deti
>	Statist Pearse M-V ch Fí Kontine Cramé	o nův chí-k v ní-kvadr. genční koef r. V	v. icient	Statist. : PO Chí-kvadr. 9,62028 9,38771 ,189818 ,186488 ,189818	HLA(2) x sv df=2 1 df=2 4 5 4	ZNAM(3) (Da p p=,00815 p=,00915	ta_deti

V tabulce vidíme znovu hodnotu statistiky chí-kvadrát a výslednou p-hodnotu testu chíkvadrát nezávislosti. Dále vidíme tři hodnoty měr síly závislosti. Protože proměnná POHL má pouze dvě kategorie, hodnota koeficientu *Fí a Cramerova V* je stejná. Hodnoty všech tří koeficientů jsou malé, můžeme proto konstatovat, že síla závislosti pohlaví a známky z tělocviku je slabá.

Pokud budeme mít obě proměnné ordinální, jsou symetrickými mírami závislosti koeficienty *Goodmanova-Kruskalova gama, Kendalovo tau-b* a *Kendalovo tau-c*. Všechny tyto míry nabývají hodnot z intervalu $\langle -1,1 \rangle$ a hodnota 0 znamená nezávislost. Hodnoty blízké 1 svědčí pro silnou přímou závislost, hodnoty blízké -1 svědčí pro silnou nepřímou závislost. Asymetrickou mírou závislosti pro dvě ordinální proměnné je *Somerovo d*. Jednou z nejpoužívanějších měr závislosti v případě dvou ordinálních veličin je *Spearmanův* *koeficient pořadové korelace*. Tento koeficient je založen na myšlence, že obě proměnné jsou seřazeny vzestupně a je jim přiřazeno jejich pořadí. Tento koeficient nabývá hodnot z intervalu $\langle -1,1 \rangle$. Pokud budou u obou proměnných přiřazena stejná pořadí, jde o silnou přímou závislost a koeficient nabývá hodnoty 1. Pokud jsou pořadí jedné proměnné přesně opačná pořadím druhé proměnné, jde o silnou nepřímou závislost a koeficient nabývá hodnoty -1. Hodnota 0 svědčí o lineární nezávislosti.

V SW Statistica zobrazíme tyto míry opět v záložce Detailní výsledky. Celá cesta je následující: Statistiky -> Základní statistiky -> Kontingenční tabulky -> OK -> Specif. tabulky (vyberte proměn.)(Tentokrát vybereme proměnné POHL a VYKON) -> OK -> OK -> na záložce Možnosti vybereme například Pearsonův & M-V chí-kvadrát a označíme Kendalovo tau-b & tau c,Goodmanovo-Kruskalovo gama, Spearmanova korelace a Somerovo d -> na záložce Detailní výsledky klikneme na Detailní 2-rozm. Tabulky.

_									5		5												
	Soubo	0 C	n C⊋ 18 mů	Upravit	Zob	razit I	Formát	Statistik	y Di	ata mii	ning G	rafy	Nástroj	STATI:	STICA Cz · Data	Data_de	ti_min		ţ,				
	172		/		1	ŀ	<u>مم</u>		~? Po	kročil	é modely *	😹 Neu	ron. sítě		Diagramy ř	izení kvali	tv - 🖼 A	nalýza pro	cesu	ST/		A VB	_
	\mathbf{r}	$\dot{\gamma}$		\sim	4		X	A			n anal v		PCA.		Aultivariat		- <u>-</u>	OF 1		bull ná	wk ana	lýza (dla s	kur
	Základi	ní Vícená	sobná	ANOVA I	Veparame	etrické Pro	kládání Ro	ozdělení a	120 VI	Lerozi	. anal. ·	Ren	FCR,	200 N	nunnvariat	-	<u>ω</u> Συ					ilyza (ule s	Kup
	statistik	y reg	rese		statist	iky ro:	zdělení s	simulace	San Ar	alýza	sily testu	U VEP/	AC	₩Y P	redictive		DHAIC S	ix Sigma *		Kal	Ikulato	ry •	
					Základ					Pok	ročilé/Vícer	ozměrné				Průmyslov	á statistik	a				Nástro	je
ľ	III n-	tar Data	dati m	ain (20c k	-4+ 267 P																		-
I		ita. Data	_ueu_n	IIII (205 KI	at 2071)																		
I		4	2	2	A	E	c	7	0		0 1/		11	10	12	14	15	10	4	7	10	10	
I						VAZKV	ниоты				9 10			12	15	14	15	10	-		10	19 19	B
I	1	7	13	0	1	160	52	1	DOVL		Výsledky; k	ontinge	nční tab	ulky: Da	ata_deti_n	nin				8	×	0	2
I	2	7	13	0	1	165	57	1				-		-					_			2	2
I	3	8	13	0	1	172	60	1		Z	ákladní výsl	adky D	etailní vý:	sledky	Možnosti					Výpo	čet	1	2
I	4	8	13	0	1	169	55	1			Výpočet tak	ulek			Statistiky o	letailn ích i	2-rozměrný	ch tab		Ctomo		1	1
I	5	8	13	0	1	159	51	1			Z Zuúrazni	t četo N	10		Paamo	níw & M-V	(chíkwad	4		Storrio		1	2
I	6	7	13	0	1	162	44	1			V Zvylazi	, ceut. >	10		I carso		CITIKVAU	a.		Možnos	sti▼	0	1
I	7	8	13	0	1	156	53	1			Očekáv	ané četno	osti		- Fisher	exakt., Yat	es, McNer	mar (2 x 2)				0	2
I	8	8	13	0	1	158	49	1			📃 Reziduá	ní četno	sti		📃 Fí (tab	ulky 2x2) 8	Cramérov	oV&C		A <u>n</u> al.sk	kup	1	1
I	9	7	13	0	1	159	49	1			Procenta	a celkové	bo počtu		V Kenda	lovo tau-b	& tau-c		Výpo	čet max	L.	2	1
I	10		13	0	1	1/4	67	1			Decemb				Goode	na sua Ka			véro	hod.chi- vza	-kv.a	1	2
I	11		13	0	1	165	51	1	_		Procenta	i z poctu	v radku		Goodin	anovo-Nu	iskalovo g	ama	více	rozměrný	ých	2	1
I	12	8	13	0	2	166	55	1			Procenta	z počtu	ve sloup	ci	Speam	nanova ko	relace		Log-	ek četno ineární r	osti viz modul	1	2
I	13	7	10	0	2	100	50	1	_					-1	V Somen	ovo d			1			1	4
I	14	2	12	0	1	164	50	1							Koefici	entv nejista	otv					1	1
I	16	7	13	0	1	156	40	1	-								.,					1	1
1	17	7	13	0	1	155	40	1	-				_	_	_	_	_						1
I	18	7	13	0	1	160	50	1	-	1	1	146	2	30	2	12.	0	3 4.	5	2	9	1	1

V našem souboru Data_deti_min.sta nás zajímalo, zda věk ovlivňuje výkon hodnocený na celočíselné stupnici 0-2. Student s počtem bodů 0-8 byl v základní skupině. Student byl zařazen do první výkonnostní třídy, pokud součet bodů studenta ve čtyřech disciplínách byl vyšší než 8 a pokud byl tento součet vyšší než 13, byl zařazen do druhé výkonnostní třídy. Výsledky těchto dalších měr závislosti pro naše data jsou zobrazeny v následující tabulce.

Z tabulky vyplývá, že proměnné VYKON a VEK nejsou závislé, tedy že výkon nezávisí na věku. P-hodnota testu chí-kvadrát je vyšší než běžné hladiny významnosti a také všechny koeficienty měřící sílu závislosti jsou velmi malé. U Spearmanova koeficientu to potvrzuje i test o nulové hodnotě tohoto koeficientu. P-hodnota tohoto testu je příliš vysoká a to vede k závěru, že data nejsou v rozporu s nulovou hodnotou Spearmanova koeficientu a tedy obě proměnné jsou nezávislé. Asymetrická míra závislosti Somerovo d je uvedena v obou možných podobách. Měří závislost věku na výkonu i výkonu na věku.

٨		560	🗘 🛅 🔻	100	-			_
4	Soubor	Domů	Upravit	Zobrazit	Formát	Sta	tistiky	Da
Z	(ákladní tatistiky	Vícenásobná regrese		1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	Prokládání R rozdělení	ozděl simul	ení a ace	Po ₩ Víc N An
				Základ				
>				Statist. : VEK	(4) x VYKO	N(3)	(Data_	_deti_n
>	Statist			Statist. : VEK Chí-kvadr.	(4) x VYKO sv	N(3)	(Data_	_deti_n
>	Statist. Pearso	onův chí-kv	1.	Statist. : VEK Chí-kvadr. 5,354532	(4) x VYKO sv	N(3) df=6	(Data_ F p=,	_deti_n 5 49921
>	Statist Pearso M-V ch	onův chí-k v ní-kvadr.	1.	Statist. : VEK Chí-kvadr. 5,354532 5,357775	(4) x VYKO sv	N(3) df=6 df=6	(Data_ p=, p=,	_deti_n 5 49921 49881
>	Statist Pearso M-V ch Kendal	o nův chí-kv ní-kvadr. I. tau b & c	<i>ı</i> .	Statist. : VEK Chí-kvadr. 5,354532 5,357775 b=,0209368	(4) x VYKO sv c=,0205	N(3) df=6 df=6 5212	(Data_ [p=, p=,	_deti_n 5 49921 49881
>	Statist Pearso M-V ch Kendal Somer	n ův chí-kv ní-kvadr. I. tau b & c s. D(X Y), D	/. (Y X)	Statist. : VEK Chi-kvadr. 5,354532 5,357775 b=,0209368 X Y=,02204	(4) x VYKO sv c=,0205 Y X=,01	N(3) df=6 df=6 5212 1988	(Data_ F p=, p=,	_deti_n 5 49921 49881
>	Statist Pearso M-V ch Kendal Somer Gama	n ův chí-kv ní-kvadr. I. tau b & c s. D(X Y), D	/. (Y X)	Statist. : VEK Chí-kvadr. 5,354532 5,357775 b=,0209368 X Y=,02204 ,0319430	(4) x VYKO sv c=,0205 Y X=,01	N(3) df=6 df=6 5212 1988	(Data_ 	_deti_n 5 49921 49881

Zvláštní pozornost budeme ještě věnovat případu, kdy budeme mít dvě *dichotomické proměnné. Dichotomická proměnná* je proměnná, která nabývá pouze dvou hodnot. Velmi často jsou to hodnoty 0 a 1, které mohou také vyjadřovat odpovědi typu ano/ne. V tomto případě se kontingenční tabulce může také říkat *čtyřpolní tabulka*, protože má jen čtyři pole četností.(Můžeme je postupně označit A, B, C, D). V případě čtyřpolní tabulky můžeme použít k zjištění vztahu závislosti mezi dvěma veličinami kromě testu chí kvadrát i tzv. *Yatesovu korekci*. V tomto případě je testová statistika chí-kvadrát ještě upravena. Od rozdílu pozorovaných četností je ještě odečtena hodnota 0,5, které se říká *korekce na spojitost*. Dostáváme tedy v jistém smyslu citlivější hodnotu testové statistiky. Tuto *Yatesovu korekci* zvolíme k vyšetření nezávislosti zvláště v případě, když nemáme dostatečně velké četnosti ve čtyřpolní tabulce.

<u> </u>	1 1 2 (0	i 🗉 🔻									STA	ATISTI	ICA Cz - I	Data_det	i_min						
Soubo	Do	mů	Upravit	Zob	razit	Formát	Statistiky	Da	ta mining	Grafy	Nástre	oje	Dat	ta								
Základi statistil	ní Vícená sy reg	isobná rese		Veparamo statist Základ	etrické Pro iky ro:	kládání Ro zdělení	ozdělení a simulace	<mark>,</mark> Pol ₩ Víc ₩ An	kročilé mo erozm. an alýza síly t Pokročil	dely ▼ 💐 al. ▼ 😡İ estu 💷 é/Vícerozm	Neuron. sít PLS, PCA, VEPAC ěrné	tě 🔤	🚆 Dia 🚆 Mu 🔐 Pres	gramy říz Itivariate dictive Pr	ení kvalit růmyslová	iy ▼ 🔛 Ar ∰2 DC 6 g Sib á statistika	nalýza proces DE x Sigma *	su [🛃 STATIS 📲 Dávk. a 📑 Kalkula	ПСА analý átory	VB za (dle sl Nástro	k upin) je
Da	ita: Data	_deti_n	nin (20s kr	át 267ř)																		
	1	2	3	4	5	6	7	8	9	10	11	12	2	13	14	15	16	17	18	_	19	20
1	TRIDA 7	13	PUHLA	ZINAM 1	160	HIVIOTIN 52		JOVEL	🧏 Výsle	edky; kont	ingenční ta	bulky	: Data	_deti_mi	n			-		- P		20.3
2	7	13	0	1	165	57	1	-			_			_				_			2	20.9
3	8	13	0	1	172	60	1	1	Základ	lní výsledky	Detailní v	výsledk	cy M	ožnosti					Výpočet		1	20,3
4	8	13	0	1	169	55	1	1	Vín	nčet tabulek			St	tatistiky de	tailních 2	-rozměrnýc	tab		Sterne		1	19,3
5	8	13	0	1	159	51	1	1	·)p	luímanit čel				Perman	3. 2 M.V	obílovadri	4	<u> </u>	SLOITIO	1	1	20,2
6	7	13	0	1	162	44	1	1	× 2		II. 2 IV					CITIKVOUIC			⁄ložnosti▼		0	16,8
7	8	13	0	1	156	53	1	0)čekávané	četnosti			Fisher e	kakt., Yate	es, McNem	nar (2 x 2)				0	21,8
8	8	13	0	1	158	49	1	1	🗖 🗖 F	Reziduální č	etnosti			Fí (tabu	ky 2x2) &	Cramérovo	V&C	<u>ا ﷺ</u> /	A <u>n</u> al.skup.	-	1	19,6
9	7	13	0	1	159	49	1	1		mcenta cel	kového poč	itu i		Kendallo	vo tau-b	& tau-c		Výpoč	et max.		2	19,4
10	7	13	0	1	174	67	1	1			x. x/ "							věroho	od. chí-kv. a a	a 📕	1	22,1
11		13	0	1	165	51	1	1		'rocenta z p	octu v radki	u		Goodma	novo-rvu	skalovo ga	ma	vícero	změrných		2	18,7
12	8	13	0	2	166	55	1	1	🗖 🗖 F	rocenta z p	očtu ve slou	ipci		Spearma	anova kor	elace		tabule	k četností v vežrní modu	(iz	1	20,0
13		13	0	2	160	55	0	9					E	Somerov	vo d			209-11			0	21,5
14	/	13	0	1	163	50	1							Koeficie	ntv neiista	tv					1	10,0
15	0	13	0	1	104	40	1								ny nojato	.,					1	12 0
17	7	13	0	1	150	40	1					_									1	19,5
18	7	13	0	1	160	50	1	1	1	146	2	_	30	2	12.0		3 4.5	-	2	9	1	19.5

V SW Statistica zobrazíme tuto korekci opět v záložce Detailní výsledky. Celá cesta je následující: Statistiky -> Základní statistiky -> Kontingenční tabulky -> OK -> Specif.

tabulky (vyberte proměn.)(Tentokrát vybereme proměnné OBLIB a DOVED) -> OK -> OK -> na záložce Možnosti vybereme například Pearsonův & M-V chí-kvadrát a označíme Fisher exact., Yates, McNerman (2x2) -> na záložce Detailní výsledky klikneme na Detailní 2-rozm. Tabulky. V našem souboru Data_deti.sta vyšetříme, zda oblíbenost předmětu tělocvik je ovlivněna dovedností. Obě proměnné nabývají pouze hodnot 0 a 1. Studenti u sebe sami hodnotili dovednost v předmětu (ano/ne) a oblíbenost předmětu (ano/ne). Výsledky jsou uvedeny v následující tabulce. Z ní plyne, že testové kritérium chí-kvadrát testu má jinou hodnotu než Yatesova korekce. Závěr testu to však nemění. Byla prokázána závislost mezi proměnnými oblíbenost předmětu a dovedností v tomto předmětu.

٨		960	🖶 🗄 🔻	-		_		_
	Soubor	Domů	Upravit	Zobrazit	Fo	rmát	Statistik	У
Z	Základní tatistiky	Vícenásobna regrese	ANOVA Ne	parametrické statistiky	Prokla rozd	ádání R ělení	ozdělení a simulace	 ₩2
			Za	iklad				
>				Statist. : C	BLIE	B(2) x D	OVED(2)	(Da
>	Statist.			Statist. : C Chí-kvad)BLIE Ir.	8(2) x D sv	OVED(2) p	(Da
>	Statist. Pearso	onův chí-k	v.	Statist. : C Chí-kvad 31,160)BLIE I r. 010	B(2) x D sv df=1	DOVED(2) p p=,000	(Da
>	Statist. Pearso M-V ch	onův chí-k í-kvadr.	v.	Statist. : 0 Chí-kvad 31,16 24,53	DBLIE Ir. 010	B(2) x D sv df=1 df=1	DOVED(2) p p=,000 p=,000	(Da
>	Statist. Pearso M-V ch Yatesů	o nův chí-k í-kvadr. v chí-kv.	v.	Statist. : 0 Chí-kvad 31,16 24,53 28,46	DBLIE Ir. 010 964 611	B(2) x C sv df=1 df=1 df=1	DOVED(2) p=,000 p=,000 p=,000	(Da
>	Statist. Pearso M-V ch Yatesů Fisherů	o nův chí-k í-kvadr. v chí-kv. w přesný, *	v. 1-str.	Statist. : 0 Chí-kvad 31,160 24,533 28,460	DBLIE Ir. 010 964 611	8(2) x D sv df=1 df=1 df=1	DOVED(2) p=,000 p=,000 p=,000 p=,000	(Da
>	Statist. Pearso M-V ch Yatesů Fisherů Fisherů	o nův chí-k í-kvadr. v chí-kv. w přesný, ź w přesný, ź	v. 1-str. 2-str.	Statist. : C Chí-kvad 31,160 24,533 28,460	DBLIE Ir. 010 964 611	8(2) x D sv df=1 df=1 df=1	DOVED(2) p=,000 p=,000 p=,000 p=,000 p=,000 p=,000	(Da
>	Statist. Pearso M-V ch Yatesů Fisherů Fisherů McNen	onův chí-k í-kvadr. v chí-kv. v přesný, ž v přesný, ž narův chí-kv	v. 1-str. 2-str. /. (A/D)	Statist. : C Chí-kvad 31,16 24,53 28,460 156,54	DBLIE 17. 010 964 611 430	B(2) x C sv df=1 df=1 df=1 df=1	DOVED(2) p=,000 p=,000 p=,000 p=,000 p=,000 p=0,00	(Da

V případě malého souboru dat, kdy nemáme splněny předpoklady testu chí-kvadrát, můžeme k testování nezávislosti ve čtyřpolní tabulce použít *Fisherův exaktní test*. Tento test vychází z předpokladu, že data pochází ze souboru s hypergeometrickým rozdělením. Nulová hypotéza testuje, zda relativní četnost levého horního pole je rovna součinu relativních četností za první sloupec a za první řádek, protože ostatní relativní četnosti v tabulce jsou pak již jednoznačně určeny. Testu se také někdy říká *faktoriálový test*, protože pro každou variantu četností se počítají pravděpodobnosti pomocí faktoriálů. Představme si následující situaci, kdy jsme u 11 lidí zjišťovali, zda jsou kuřáci. Data jsou uvedena v následující tabulce.

	2-rozměrn	á tabulka:k	Kouření Četnost označených buněk > 5
pohlaví	kouření	kouření	Řádkové
pomavi	0	1	součty
0	5	2	7
1	1	3	4
součet	6	5	11

Chceme zjistit, zda kouření závisí na pohlaví. Pro test chí-kvadrát nezávislosti máme málo dat. Proto použijeme Fisherův test nezávislosti. V SW Statistica najdeme Fisherův test následně: Statistiky -> Základní statistiky -> Kontingenční tabulky -> OK -> Specif. tabulky

(vyberte proměn.) -> OK -> OK -> na záložce Možnosti označíme Fisher exakt., Yates, McNemar (2 x 2) -> na záložce Detailní výsledky klikneme na Detailní 2-rozm.tabulky. Výsledky jsou uvedeny v následující tabulce.

٨		960	÷ 🗄 🔻					
2	Soubor	Domů	Upravit	Zobrazit	For	mát	Statistik	y
Z	ákladní tatistiky	Vícenásobn regrese	á ANOVA Ne	parametrické statistiky	Proklá rozdě	dání R Ílení	ozdělení a simulace	¶≾ 1 11
			- /					
_			Za	klad				
>			Za	Statist. : p	ohlavi	(2) x k	ouření(2)	(poi
>	Statist.		Za	Statist. : p	ohlavi r.	(2) x k sv	couření(2) p	(poi
>	Statist. Yatesů	v chí-kv.	Za	Statist. : p Chí-kvad ,7366	ohlavi r.)71	í(2) x k sv df=1	ouření(2) p p=,390	(poi 75
>	Statist. Yatesů Fisherů	v chí-kv. v přesný, [*]	Za 1-str.	Statist. : p Chí-kvad ,73660	ohlavi r.)71	(2) x k sv df=1	couření(2) p p=,390 p=,196	(poi 75 97
>	Statist. Yatesů Fisherů Fisherů	v chí-kv. v přesný, ž	Za 1-str. 2-str.	klad Statist. : p Chí-kvad ,73660	ohlavi r.)71	(2) x k sv df=1	couření(2) p p=,390 p=,196 p=,242	(poi 75 97 42
>	Statist. Yatesů Fisherů Fisherů McNen	v chí-kv. w přesný, w přesný, ź narův chí-ky	2a 1-str. 2-str. v. (A/D)	klad Statist. : p Chí-kvad ,73660 ,12500	ohlavi r. 071	(2) x k sv df=1 df=1	couření(2) p p=,390 p=,196 p=,242 p=,723	(poi 75 97 42 67

V tabulce jsou uvedeny p-hodnoty Fisherova jednostranného i oboustranného testu. V obou případech jsou p-hodnoty vyšší než běžné hladiny významnosti a proto můžeme konstatovat nezávislost proměnných kouření a pohlaví.

Posledním testem, který v souvislosti se čtyřpolní tabulkou uvedeme, je *McNemarův test,* případně *McNemarův test symetrie.* Tímto testem vyšetřujeme např., zda se shodují názory dotazovaných na otázky s odpověďmi ano/ne ve dvou různých obdobích. Jedná se vlastně o párový test. Označme jednotlivá pole čtyřpolní tabulky následujícím způsobem.

А	В
С	D

Nulová hypotéza tohoto testu je tvaru shody četností ležících v polích B a C čtyřpolní tabulky (Jiná podoba testuje shodu četností v polích A a D čtyřpolní tabulky.). Alternativní hypotéza má tvar oboustranné hypotézy. Ukažme si použití tohoto testu. 96 studentům ve věku 14-17 let jsme zadali úlohu, kterou měli řešit a u každého studenta jsme zaznamenali úspěšnost řešení této úlohy (vyřešil/nevyřešil). Následně se studenti zúčastnili přednášky, ve které byla, mimo jiné, vysvětlena metoda, kterou se dala úloha vyřešit. Po určité době studenti dostali znovu úlohu vyřešit a opět byla zaznamenána jejich úspěšnost. Ptáme se, zda vyslechnutí přednášky mělo vliv na úspěšnost řešené úlohy. Data jsou uvedena v následující tabulce.

٨		990	🖶 🛅 🔻						_		STATIS
S	oubor	Domů	Upravit	Zobrazit	Fo	rmát	Statist	tiky	D	ata mining	G
z st	ákladní tatistiky	Vícenásobna regrese	á ANOVA N	eparametrické statistiky	Prokla	ádání ělení	Rozdělení simulace	ia №	n ¶ ∦ ∧	okročilé mo ícerozm. ana nalýza síly te Pokročilé	dely ▼ al. ▼ estu
>	Konting Četnos (Margir	genční tabu st označený nální součty	ilka (List1 v ích buněk : / nejsou oz	/ Data_Piser > 10 mačeny)	nky_z	ari_2	015)				
I	Prete	st4uspesn	ost Pos	test4uspesn 0	ost	Pos	test4usp 1	esnos	t	Řádk. součty	
	0				65				29	94	
	1				2				0	2	
	Vš.sku	ip.			67				29	96	

Z tabulky vidíme, že při prvním řešení úlohu vyřešili jen dva studenti z 96. Po přednášce už to bylo 29 studentů. Po přednášce úlohu nevyřešilo 67 studentů. Dva studenti ji před přednáškou vyřešili, po přednášce však úlohu nevyřešili. Ověřme McNermanovým testem, zda nastává shoda četností v polích B a C čtyřpolní tabulky. Tedy, zda je stejně studentů, kteří úlohu nejprve nevyřešili a po přednášce ano a studentů, kteří ji nejprve vyřešili a po přednášce si s úlohou neporadili. V SW Statistica se k testu dostaneme následující cestou. Statistiky -> Základní statistiky/tabulky -> Kontingenční tabulky -> OK -> Specif. tabulky (vyberte proměn.) -> OK -> OK -> na záložce Možnosti označíme Fisher exact, Yates, McNemar (2 x 2) -> na záložce Detailní výsledky klikneme na Detailní 2-rozm. tabulky.

Výsledky jsou uvedeny v následující tabulce.

٨		960	🖶 🗑 🔻					_
2	Soubor	Domů	Upravit	Zobrazit	Fo	rmát	Statisti	cy 🛛
Z	Základní tatistiky	Vícenásobn regrese	á ANOVA Nej	arametrické statistiky	Prokl	ádání R ělení	ozdělení a simulace	
			Za	klad				
>			Za	Statist. : p	rete	st(2) x	posttest(2) (p
>	Statist.		Za	Statist. : p Chí-kvad	rete Ir.	st(2) x sv	posttest(p	2) (p
>	Statist. Yatesü	v chí-kv.	Za	Statist. : p Chí-kvad	orete Ir. 809	st(2) x sv df=1	posttest(p p=,87	2) (p 122
>	Statist. Yatesů Fisherů	v chí-kv. iv přesný,	La 1-str.	Statist. : p Chí-kvad	orete Ir. 809	st(2) x sv df=1	posttest(p p=,87 p=,484	2) (p 122 487
>	Statist. Yatesů Fisherů Fisherů	v chí-kv. iv přesný, iv přesný, ž	1-str. 2-str.	Statist. : p Chí-kvad ,0262	orete Ir. 809	st(2) x sv df=1	posttest(p p=,87 p=,484 p=1,0	2) (p 122 487 000
>	Statist Yatesů Fisherů Fisherů McNen	i v chí-kv. Iv přesný, Iv přesný, J narův chí-k	1-str. 2-str. v. (A/D)	Statist. : p Chí-kvad ,0262	orete Ir. 809	st(2) x sv df=1 df=1	posttest(p=,87 p=,484 p=1,00 p=,00	2) (p 122 487 000 000

Z výsledků v tabulce se zaměříme na poslední řádek - McNemarův chí. kvadr. (B/C). Tedy situace, kdy testujeme, že se četnosti na vedlejší diagonále (četnosti zachycující situace, kdy došlo k změně úspěšnosti řešení) jsou shodné, nebo se statisticky významně liší. Z velmi nízké p-hodnoty plyne, že zamítáme nulovou hypotézu. Můžeme konstatovat, že přednáška má pozitivní efekt na zvládnutí úlohy studenty.