

## 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

- 1) 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.

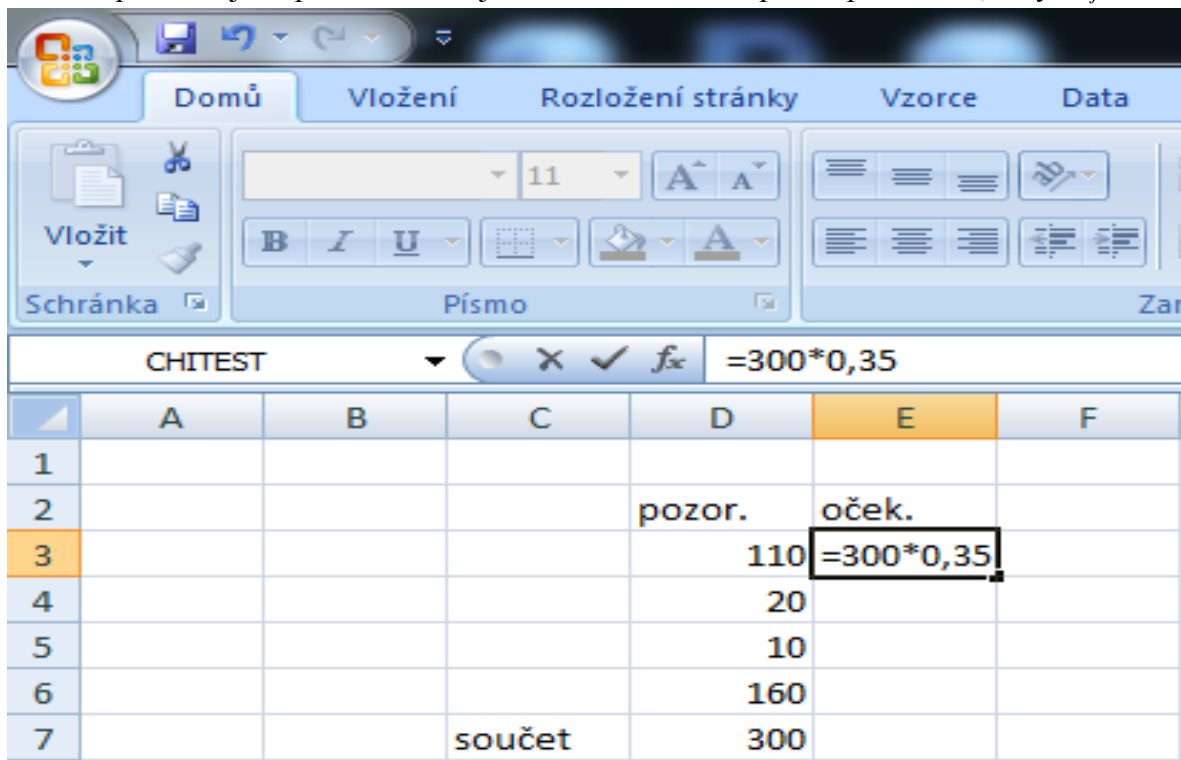
## Příklad 1

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 potenciálních 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  $n\pi_j$ .



	A	B	C	D	E	F
1						
2				pozor.	oček.	
3				110	=300*0,35	
4				20		
5				10		
6				160		
7			součet	300		

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J
1			pozor.	oček.						
2			110	105						
3			20	30						
4			10	15						
5			160	150						
6		součet	300	300						
7										
8		=CHITEST(C2:C5;D2:D5)								

The 'Argumenty funkce' dialog box shows the following details:

- Funkce:** CHITEST
- Aktuální:** C2:C5 = {110|20|10|160}
- Očekávané:** D2:D5 = {105|30|15|150}
- Výsledek:** = 0,116336824

Text in the dialog box: "Vrátí test nezávislosti: hodnota ze statistického rozdělení chí-kvadrát a příslušné stupně volnosti. Očekávané je oblast dat obsahující podíl součinu součtů řádků a sloupců a celkového součtu."

**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 zadávacím 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 jednoprocenní, pětiprocenní ani deseti procentní hladině významnosti. Předpoklad firmy není v rozporu se zjištěnou strukturou zájmu o výrobek z průzkumu.

Pozorované vs. očekávané četnosti (Tabulka2)				
Chi-Kvadr. = 5,904762 sv = 3 p = ,116339				
Případ	pozorov. Prom5	očekáv. Prom6	P - O	(P-O) <sup>2</sup> /O
C: 1	110,0000	105,0000	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	150,0000	10,0000	0,666667
Sčt	300,0000	300,0000	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  $\pi_j$  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.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	příchody	v hodinách			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					
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					
10	9:34:00	9,566667											

Tímto jsme získali empirické četnosti. Pokud nás zajímá, zda přicházejí občané na úřad v tomto rozmezí pěti dvouhodinových intervalů stejnoměrně, jsou teoretické pravděpodobnosti příchodu náhodného občana na úřad v některém z dvouhodinových intervalů stejné a tedy 0,2. Očekávané četnosti jsou dány součinem  $186 \cdot 0,2 = 37,2$ . 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.

The screenshot shows the SW Statistica interface with the following data table:

Případ	pozorov. pozorované četnosti	očekav. teoretické četnosti	P - O	(P-O) <sup>2</sup> / O
C: 1	36,0000	37,2000	-1,2000	0,038710
C: 2	40,0000	37,2000	2,8000	0,210753
C: 3	27,0000	37,2000	-10,2000	2,796774
C: 4	39,0000	37,2000	1,8000	0,087097
C: 5	44,0000	37,2000	6,8000	1,243011
Sčt	186,0000	186,0000	-0,0000	4,376344

Summary statistics from the screenshot:  
 Pozorované vs. očekávané četnosti (Tabulka2)  
 Chi-Kvadr. = 4,376344 sv = 4 p = ,357465

**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  $X^2$ .

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 m<sup>2</sup> látky a byl zaznamenáván počet kazů vždy na ploše 1 m<sup>2</sup>. Data jsou uvedena v souboru DataExcelchishoda.xls v záložce kazy na látce. Rozhodněte, zda je možno počet kazů na 1 m<sup>2</sup> 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  $\lambda$ , 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 запиšeme do sloupce kategorie počtu kazů na 1 m<sup>2</sup> 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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	kazy na látce		kategorie	em. četnosti		pravděp.	teor. četnosti	sloučené	sloučené	sloučené emp.					
2	1		0	3		0,08046	2,01149		teor.						
3	0		1	3		0,202758	5,068955	0 až 1	7,080445	6					
4	4		2	6		0,255475	6,386884	2	6,386884	6					
5	2		3	8		0,214599	5,364982	3	5,364982	8					
6	3		4	2		0,135198	3,379939	4 a více	6,167689	5					
7	5		5	2		0,06814	1,703489								
8	1		6	1		0,04337	1,084261								
9	6									p =	0,63614	p-hodnota je vyšší než běžné hladiny významnosti, proto nezamítáme hypotézu o shodě dat s Poissonovým rozdělením.			
10	2														
11	3														
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  $\lambda=2,52$  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.

Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekav. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.	Pozorované - Očekáv.
<= 0,00000	3	3	12,00000	12,0000	2,011490	2,01149	8,04596	8,0460	0,98851
1,00000	3	6	12,00000	24,0000	5,068957	7,08045	20,27583	28,3218	-2,06896
2,00000	6	12	24,00000	48,0000	6,386884	13,46733	25,54753	53,8693	-0,38688
3,00000	8	20	32,00000	80,0000	5,364981	18,83231	21,45992	75,3292	2,63502
4,00000	2	22	8,00000	88,0000	3,379939	22,21225	13,51976	88,8490	-1,37994
5,00000	2	24	8,00000	96,0000	1,703489	23,91574	6,81396	95,6630	0,29651
< Nekonečno	1	25	4,00000	100,0000	1,084261	25,00000	4,33704	100,0000	-0,08426

### 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:

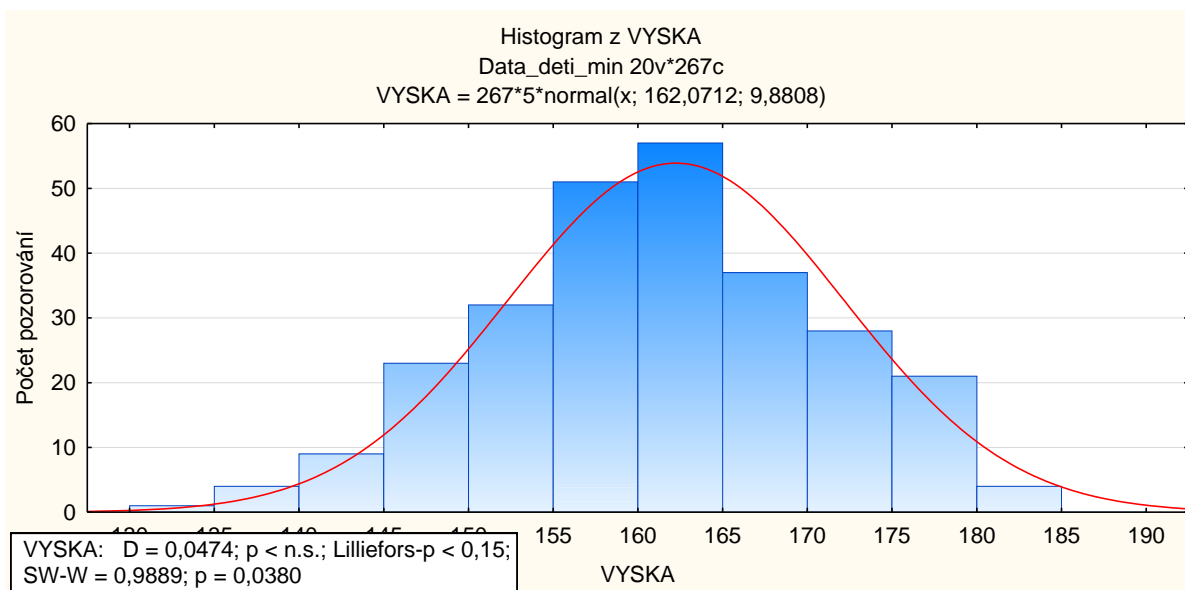
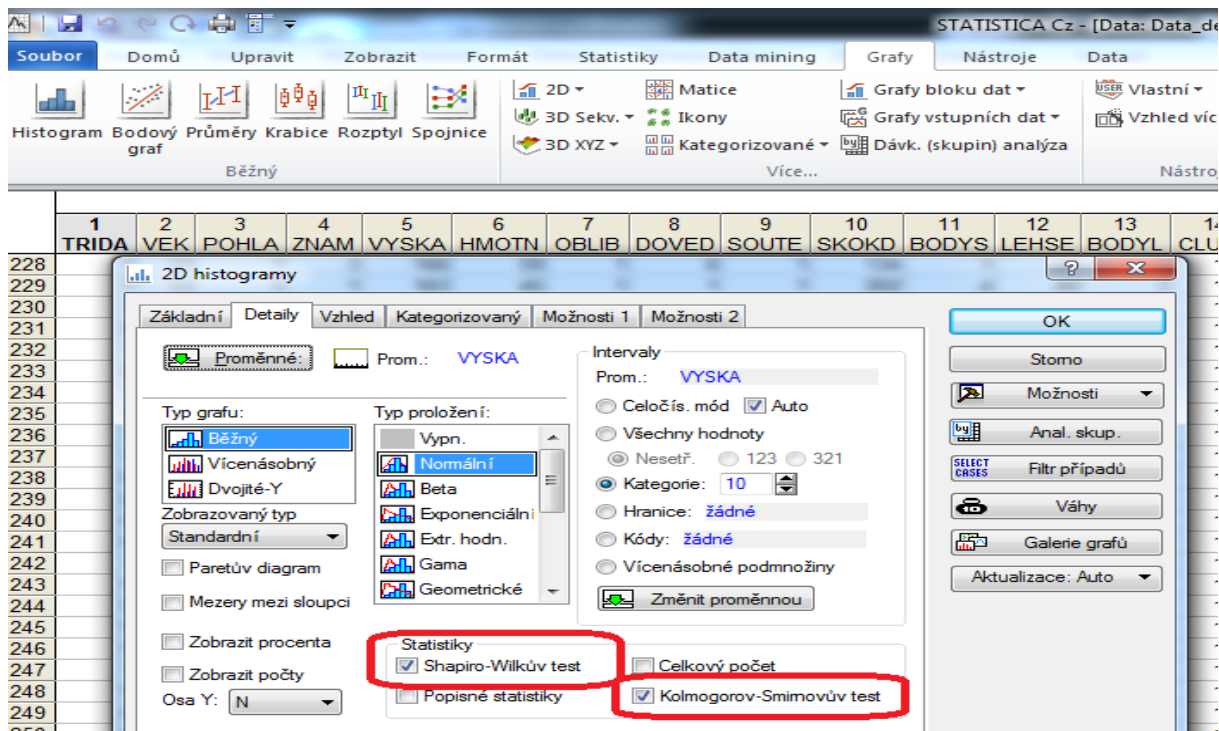
- Histogram
- Q-Q graf
- Pravděpodobnostní graf
- P-P graf

My pro ověření normality vybereme spojitě 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.

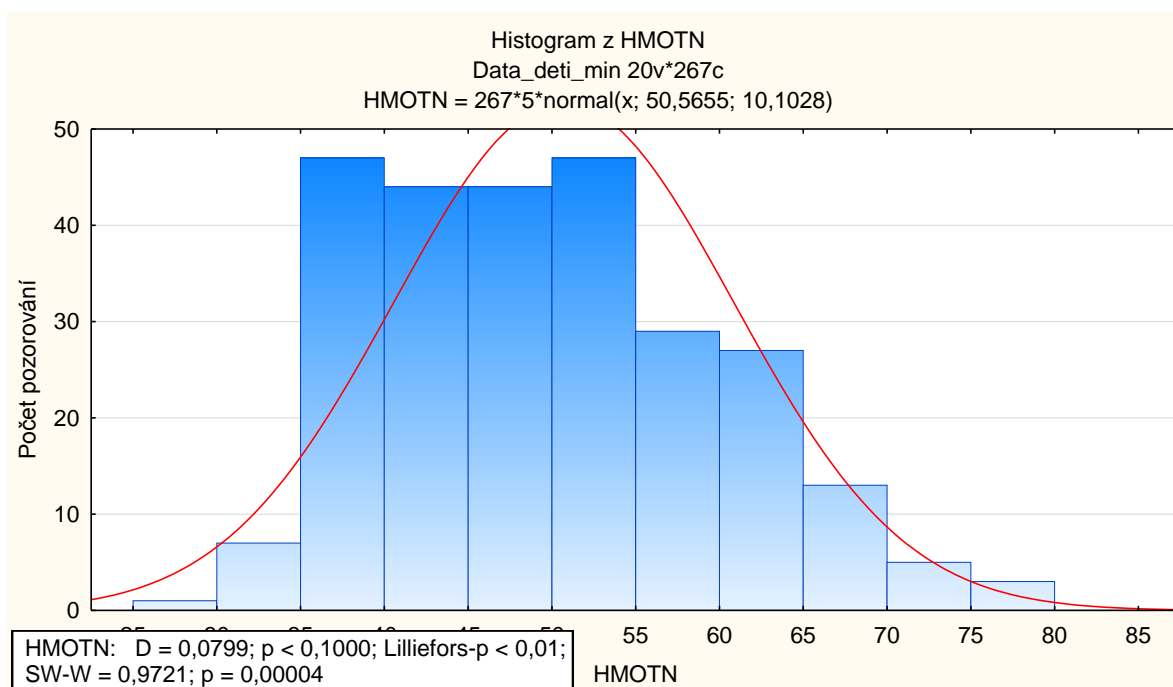
V programu Statistica histogram najdeme v záložce Grafy mezi běžnými grafy. Do proměnné zvolíme proměnnou VYSKA. V záložce kategorie nastavíme počet intervalů, do kterých budou data rozdělena, a zaškrtneme typ rozdělení normální. V záložce details můžeme ještě zaškrtnout dva testy Shapiro-Wilkův test 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.





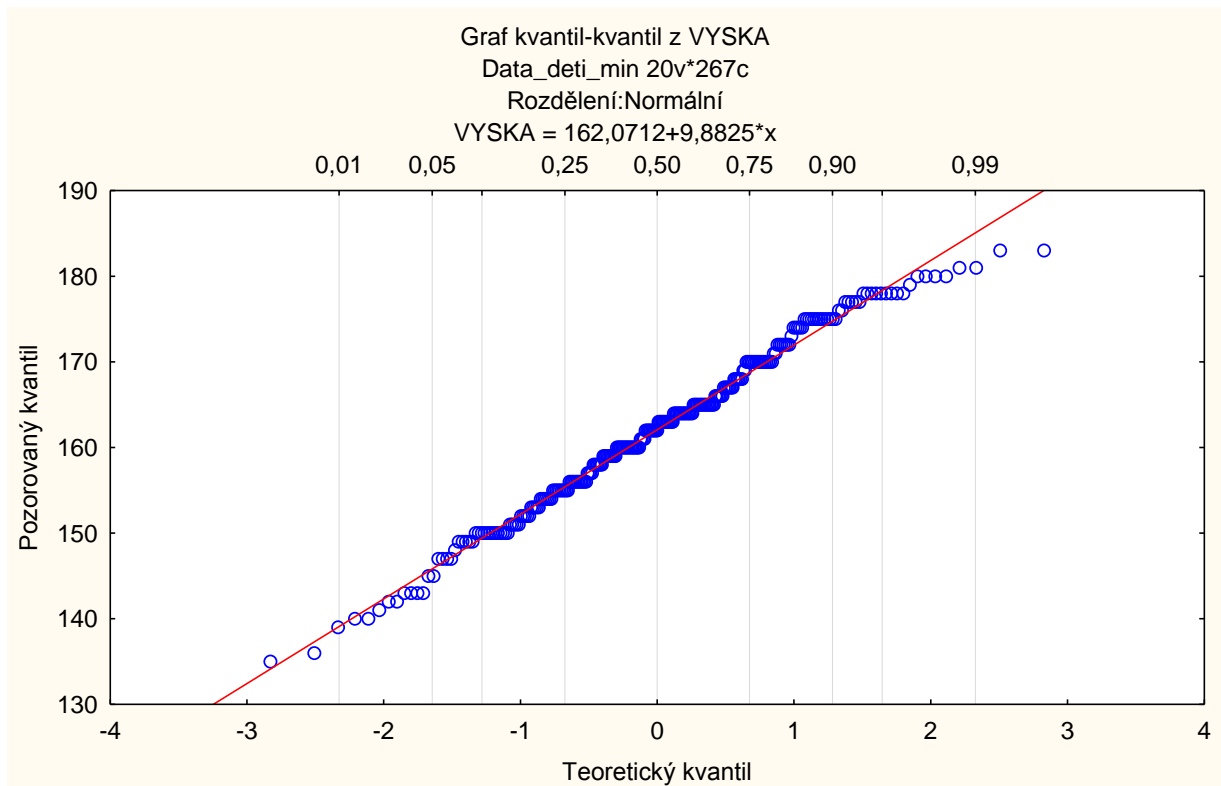
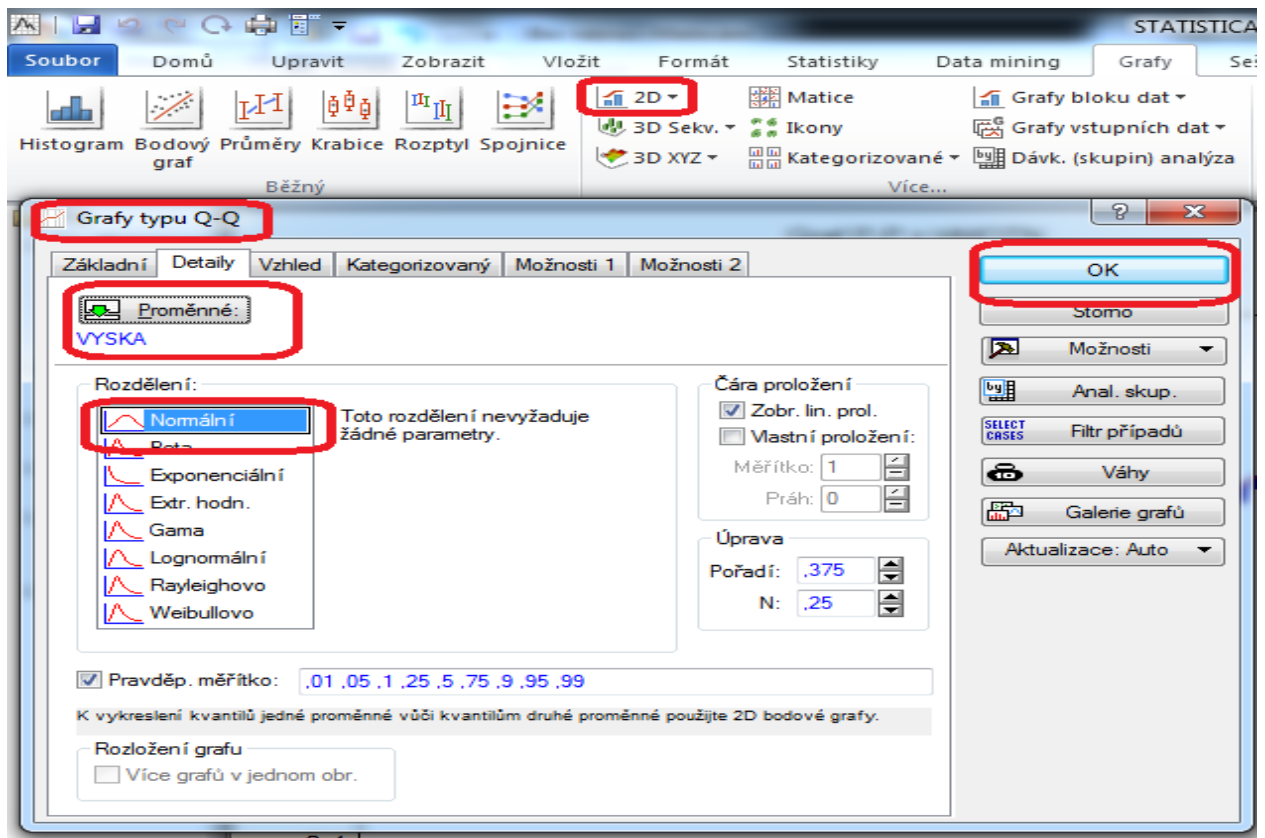
Graf histogramu kopíruje Gaussovu křivku a to značí, že data pochází z normálního rozdělení. To potvrzují i výsledné p-hodnoty obou zvolených testů, které jsou zobrazeny v tabulce vlevo dole na grafu histogramu. Okomentujeme je později.

Ukažme si ještě jeden histogram veličiny HMOTN.

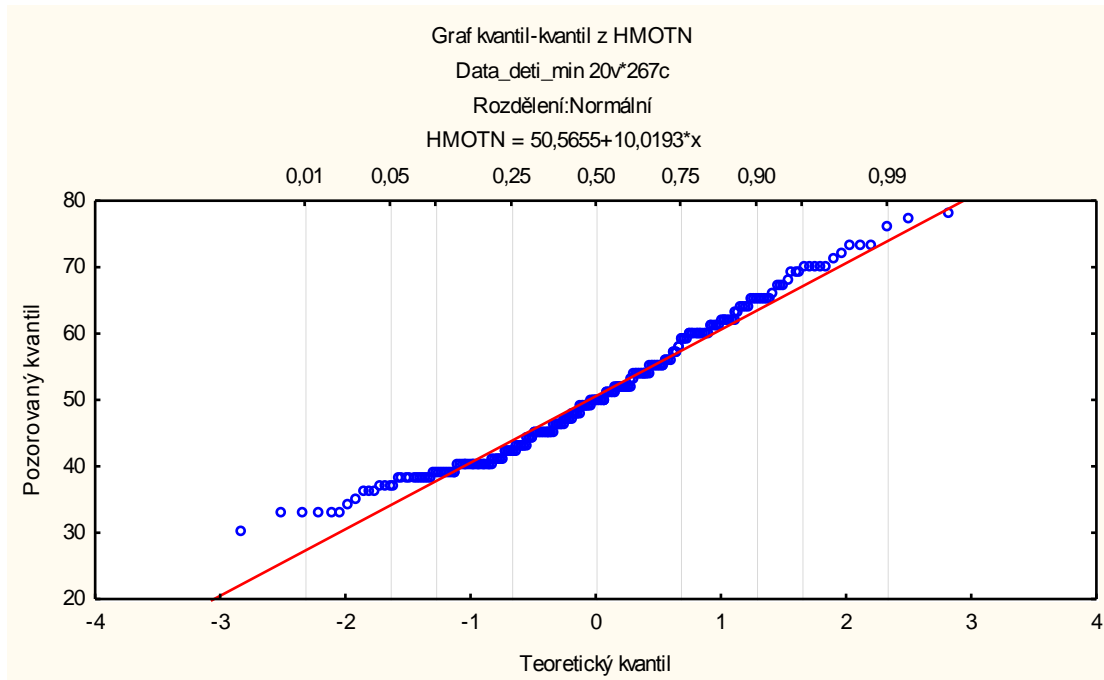


Graf histogramu veličiny hmotnost již Gaussovu křivku tak nekopíruje a zdá se, že tato veličina nemá normální rozdělení. To potvrzují i p-hodnoty testů uvedené vlevo dole.

*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.

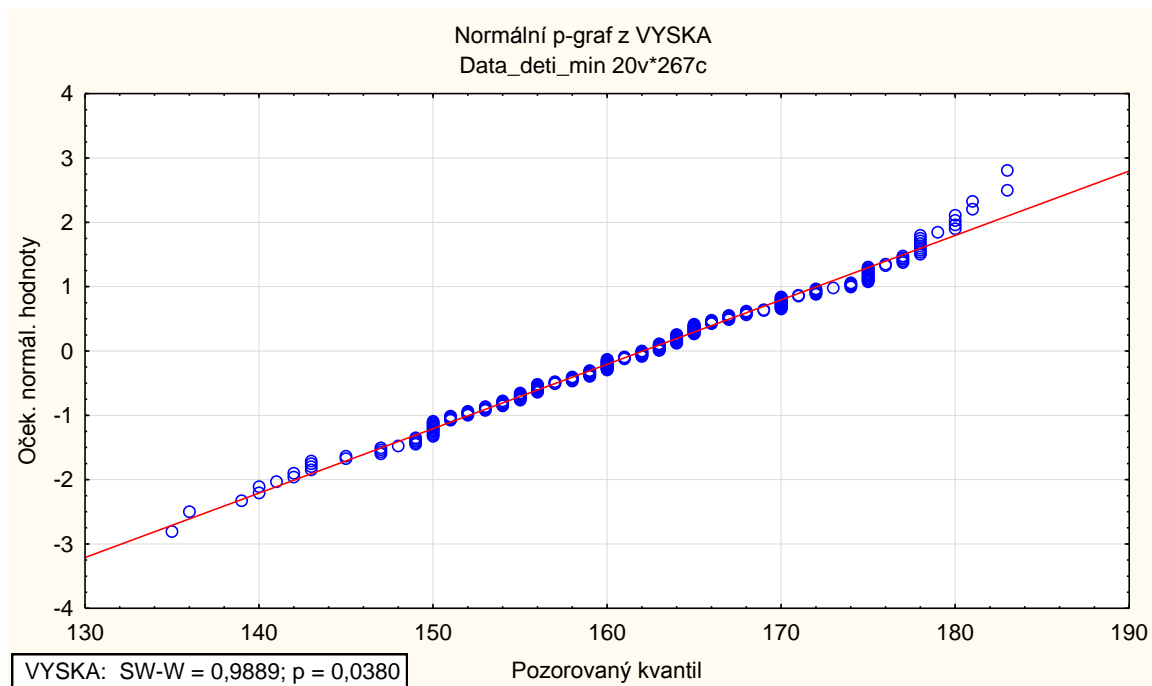
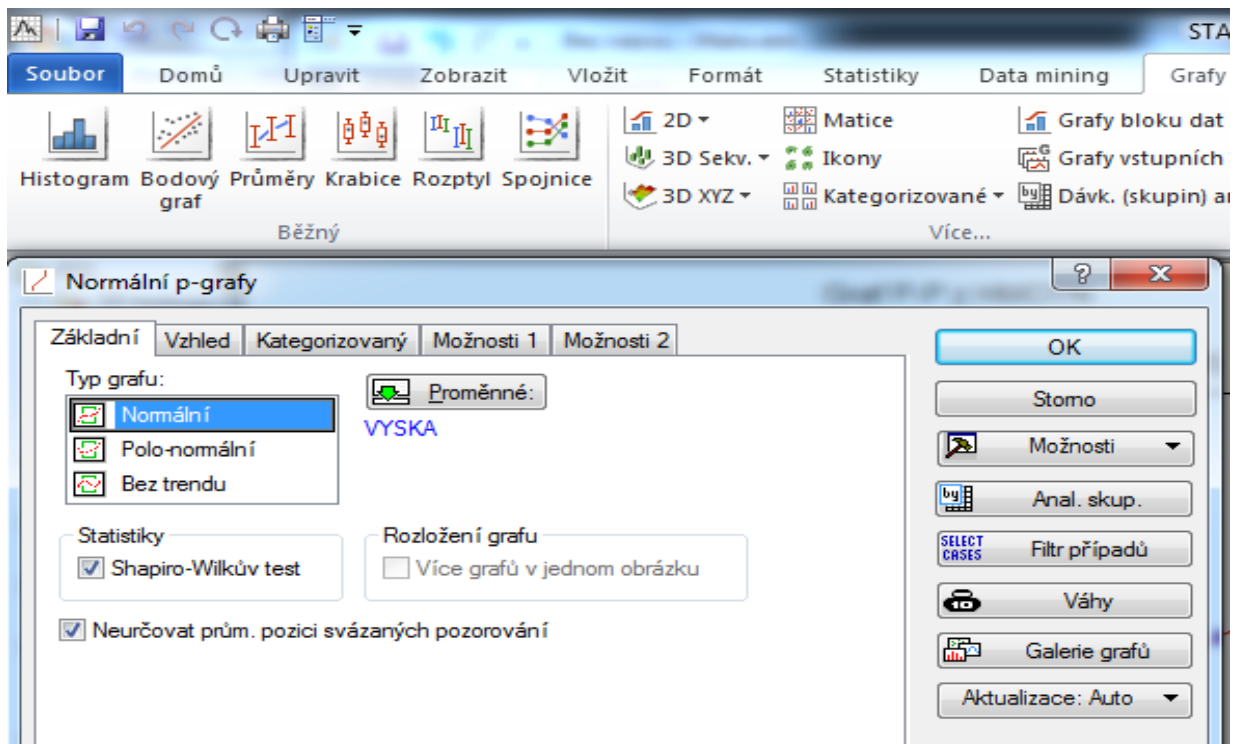


Vidíme, že body se od přímky vzdalují jen pro vysoké kvantily. Další body leží téměř na přímce. Můžeme tedy učinit závěr, že data pochází ze souboru, který má normální rozdělení. Stejný postup provedeme i s proměnnou HMOTN. Získáme následující graf. Vidíme, že v tomto případě je mnohem více bodů mimo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

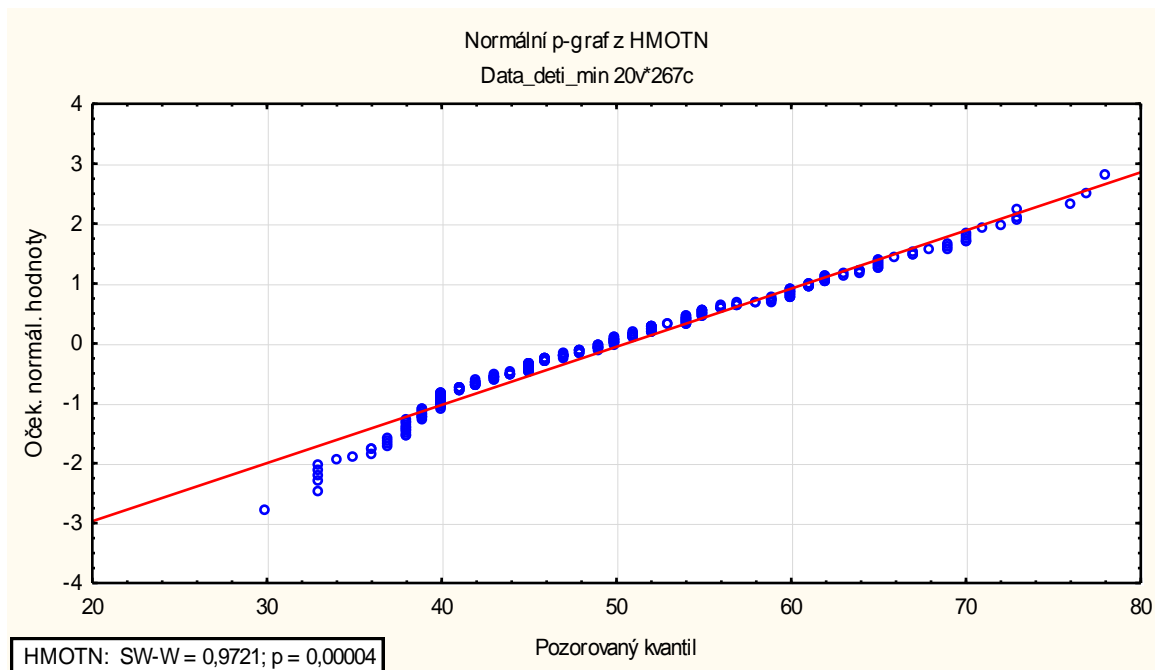


*N-P graf* umožňuje posoudit pouze, zda data pocházejí z normálního rozdělení. Na vodorovnou osu jsou vynesena vzestupně seřazená data a na svislou osu jsou vyneseny kvantily normovaného normálního rozdělení. Pokud data pocházejí z normálního rozdělení, budou ležet na přímce. Pokud rozdělení nebude symetrické, 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 jednocentní hladině významnosti tento závěr podporuje.

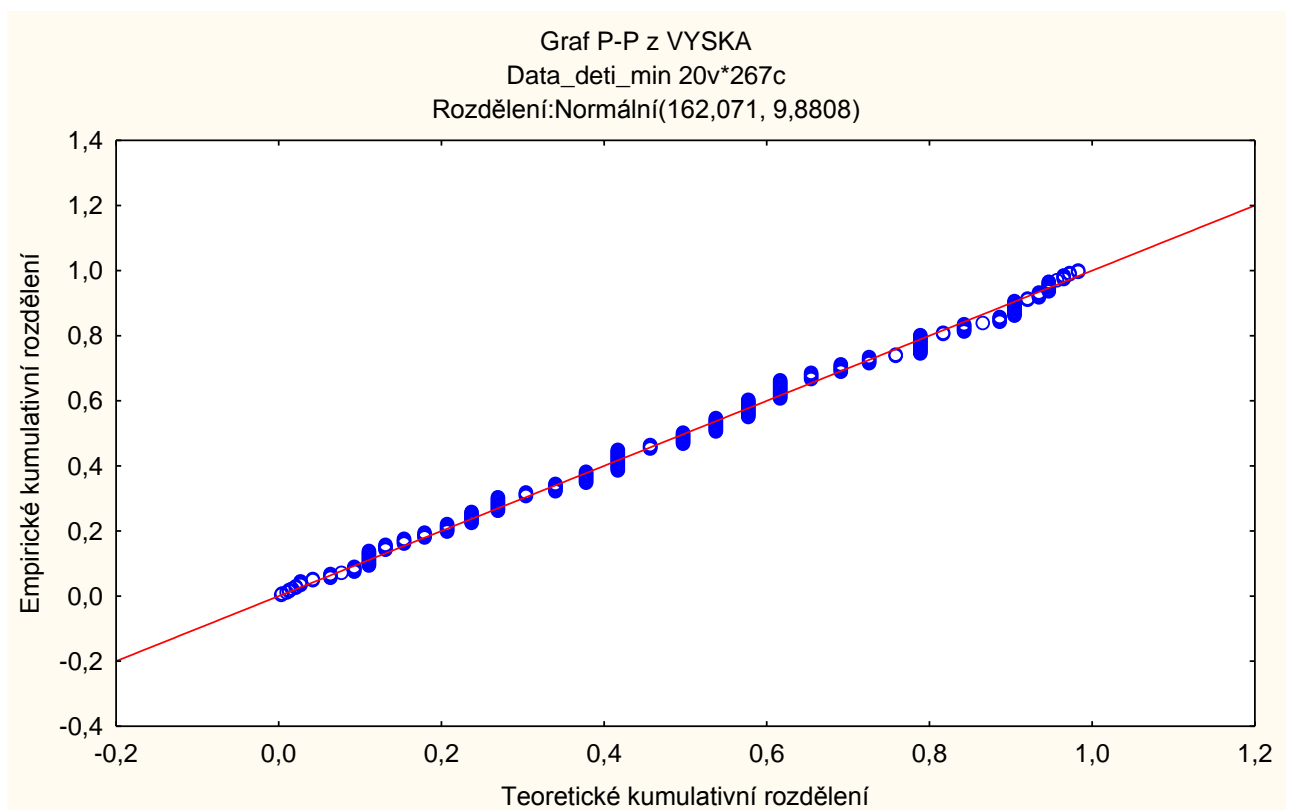
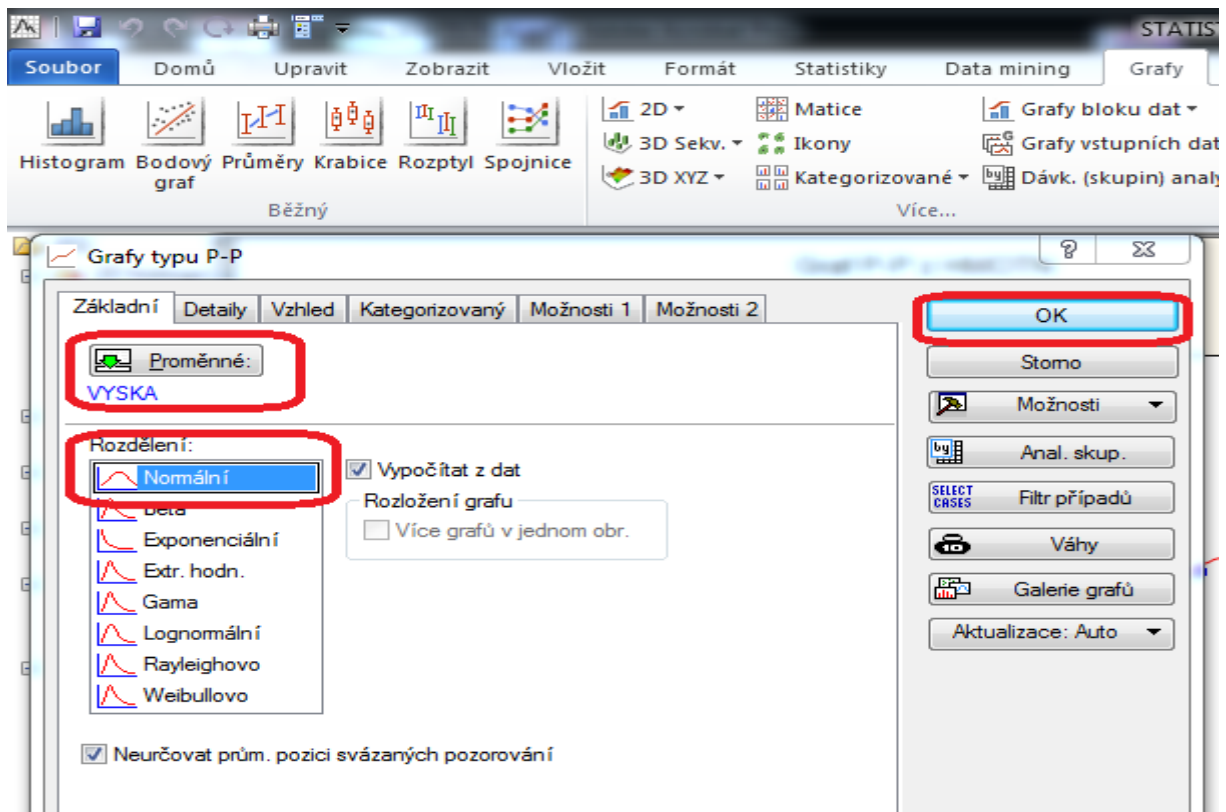


Pro data HMOTN vypadá graf následovně.



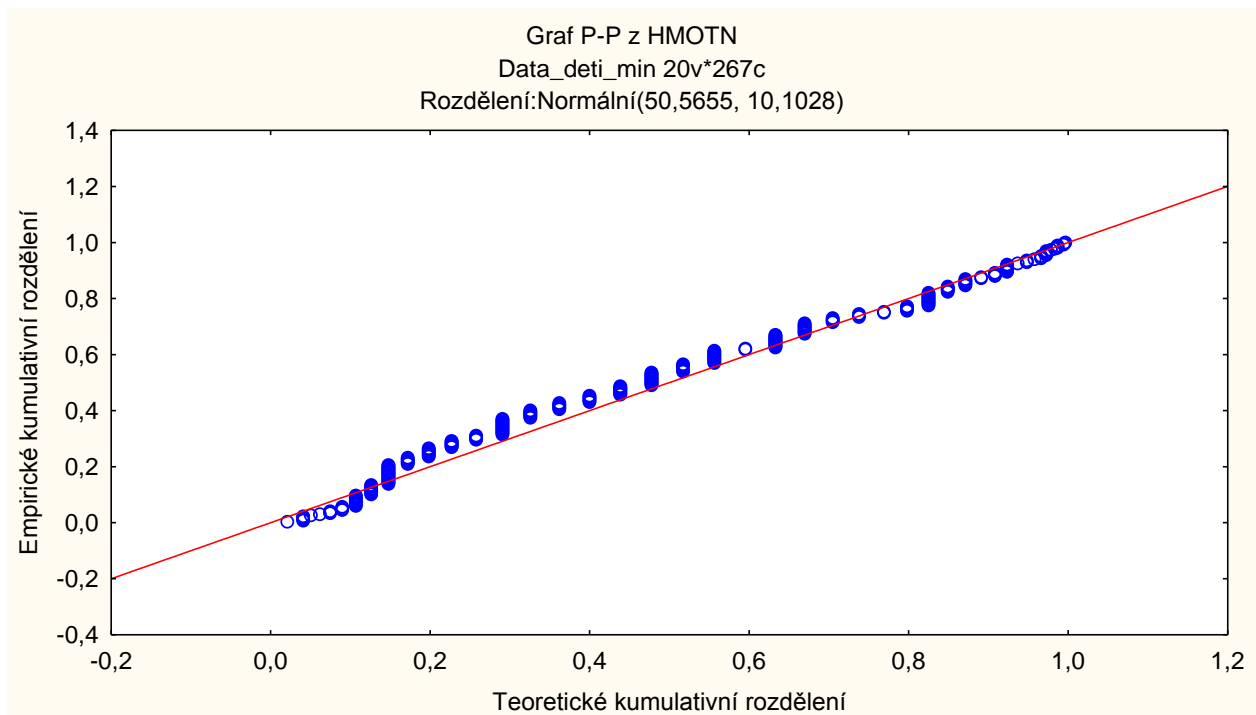
Body již neleží jednoznačně na přímce a nízká p- hodnota testu normality v tabulce vlevo dole na grafu potvrzuje, že data nepochází z populace s normálním rozdělením.

*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 směrnici 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 sestojíme pro veličinu HMOTN. Na následujícím grafu

vidíme, že většina bodů neleží na přímce a z toho bychom usoudili, že data nepochází z normálního rozdělení.



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  $\mu$  a rozptylem  $\sigma^2$ , 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.

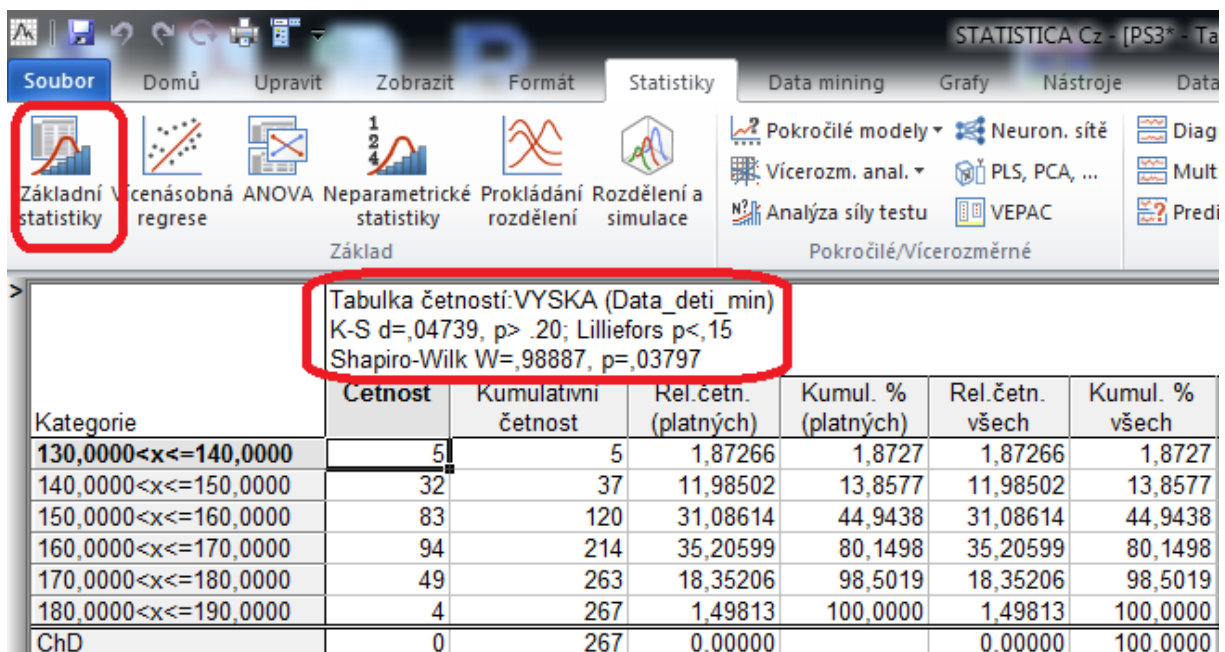
### *Lillieforsův test normality dat*

Jedná se o modifikaci 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.

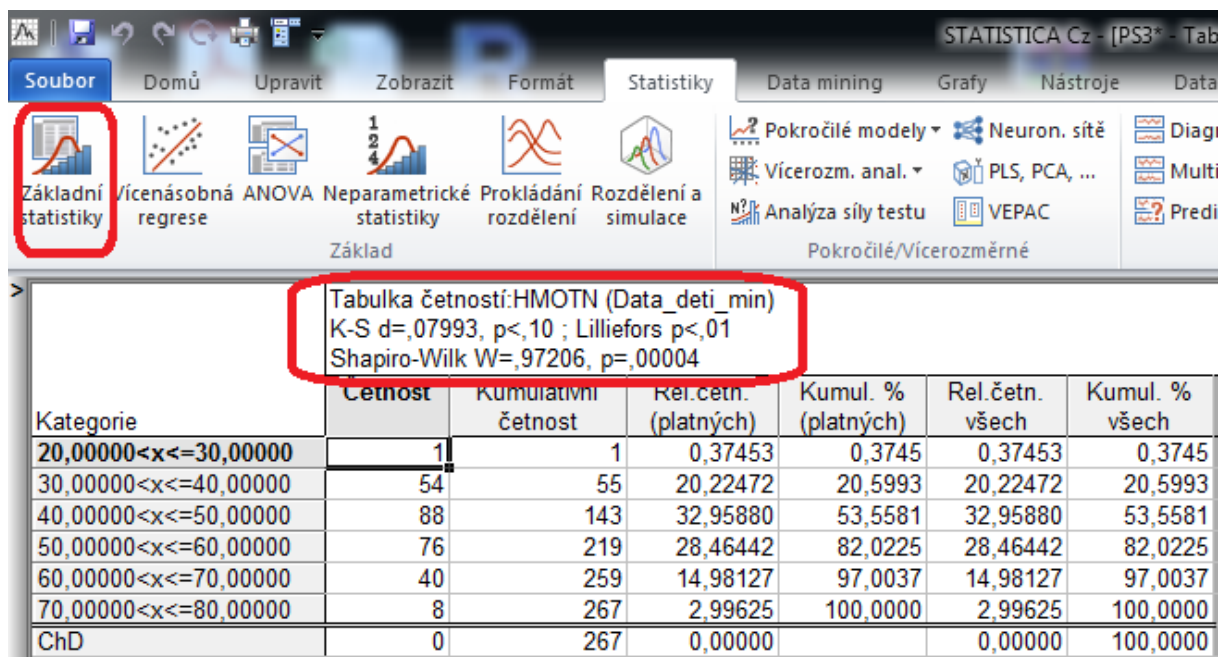


The screenshot shows the SW Statistica software interface. The 'Statistiky' menu is open, and the 'Základní statistiky' option is highlighted. Below the menu, a table titled 'Tabulka četností: VYSKA (Data\_deti\_min)' is displayed. The table shows the frequency distribution for the variable 'VYSKA'. The table has 7 columns: 'Kategorie', 'Četnost', 'Kumulativní četnost', 'Rel.četn. (platných)', 'Kumul. % (platných)', 'Rel.četn. všech', and 'Kumul. % všech'. The data is as follows:

Kategorie	Četnost	Kumulativní četnost	Rel.četn. (platných)	Kumul. % (platných)	Rel.četn. všech	Kumul. % všech
130,0000<x<=140,0000	5	5	1,87266	1,8727	1,87266	1,8727
140,0000<x<=150,0000	32	37	11,98502	13,8577	11,98502	13,8577
150,0000<x<=160,0000	83	120	31,08614	44,9438	31,08614	44,9438
160,0000<x<=170,0000	94	214	35,20599	80,1498	35,20599	80,1498
170,0000<x<=180,0000	49	263	18,35206	98,5019	18,35206	98,5019
180,0000<x<=190,0000	4	267	1,49813	100,0000	1,49813	100,0000
ChD	0	267	0,00000		0,00000	100,0000

Summary statistics for the 'VYSKA' data are provided in a red box above the table:

Tabulka četností: VYSKA (Data\_deti\_min)  
K-S  $d=,04739$ ,  $p> ,20$ ; Lilliefors  $p<,15$   
Shapiro-Wilk  $W=,98887$ ,  $p=,03797$



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.

STATISTICA Cz - [PS2\* - Proměnná: HMOTN, Rozdělení: Normální (Data\_deti\_min)]

Soubor Domů Upravit Zobrazit Formát Statistika Data mining Grafy Nástroje Data Seřit

Základní statistiky Vícenásobná regrese ANOVA Neparаметrické statistiky **Prokládání rozdělení** Rozdělení a simulace Pokročilé modely Neuron. sítě Diagramy řízení kvality Analýza procesu Vícerozm. anal. PLS, PCA, ... Multivariate DOE STA Víceerozm. anal. VEPAC Analyza síly testu VEPAC Predictive Six Sigma Kalk

Proměnná: HMOTN, Rozdělení: Normální (Data\_deti\_min)  
 Chí-kvadrát = 35,95460, sv = 7 (uprav.) , p = 0,00001

Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.	Pozorované - Očekáv.
<= 30,00000	1	1	0,37453	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	24,0025
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	-12,4286
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

STATISTICA Cz - [PS2\* - Proměnná: VYSKA, Rozdělení: Normální (Data\_deti\_min)]

Soubor Domů Upravit Zobrazit Formát Statistika Data mining Grafy Nástroje Data Seřit

Základní statistiky Vícenásobná regrese ANOVA Neparаметrické statistiky **Prokládání rozdělení** Rozdělení a simulace Pokročilé modely Neuron. sítě Diagramy řízení kvality Analýza procesu Vícerozm. anal. PLS, PCA, ... Multivariate DOE STA Víceerozm. anal. VEPAC Analyza síly testu VEPAC Predictive Six Sigma Kalk

Proměnná: VYSKA, Rozdělení: Normální (Data\_deti\_min)  
 Chí-kvadrát = 8,86598, sv = 6 (uprav.) , p = 0,18126

Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.	Pozorované - Očekáv.
<= 135,00000	1	1	0,37453	0,3745	0,82078	0,8208	0,30741	0,3074	0,17922
140,00000	4	5	1,49813	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	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	-1,69278
160,00000	51	120	19,10112	44,9438	48,02779	111,3348	17,98794	41,6984	2,97221
165,00000	57	177	21,34831	66,2921	53,28252	164,6173	19,95600	61,6544	3,71748
170,00000	37	214	13,85768	80,1498	46,00653	210,6238	17,23091	78,8853	-9,00653
175,00000	28	242	10,48689	90,6367	30,91633	241,5402	11,57915	90,4645	-2,91633
180,00000	21	263	7,86517	98,5019	16,16838	257,7086	6,05557	96,5201	4,83162
185,00000	4	267	1,49813	100,0000	6,57987	264,2884	2,46437	98,9844	-2,57987
< Nekonečno	0	267	0,00000	100,0000	2,71158	267,0000	1,01557	100,0000	-2,71158

### 3.2 Vyšetřování závislosti dvou kategoriálních náhodných veličin

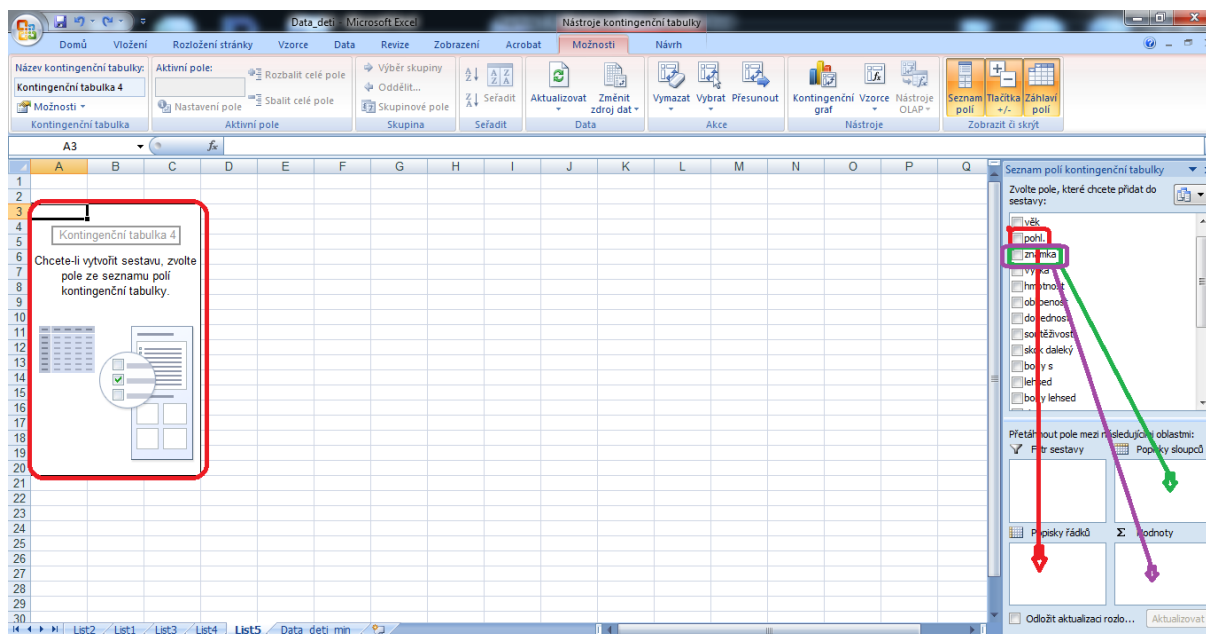
V této části budeme vyšetřovat závislost dvou kategoriálních veličin. Tyto veličiny mohou být nominální, ordinální i kvantitativní. Naučíme se měřit sílu závislosti, která může být symetrická, či asymetrická. K tomuto zjišťování bude třeba mít data uspořádaná v tabulce rozdělení četností podle dvou znaků.

#### 3.2.1 Kontingenční tabulky

Tak jako jsme zadaná data roztřídili do tabulky rozdělení četností podle určitého znaku, stejně je dokážeme roztřídít 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 diskretního 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.



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  $\Sigma$  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			
pohl.	1	2	3	Celkový součet
0	59	32	5	96
1	135	30	6	171
<b>Celkový součet</b>	<b>194</b>	<b>62</b>	<b>11</b>	<b>267</b>

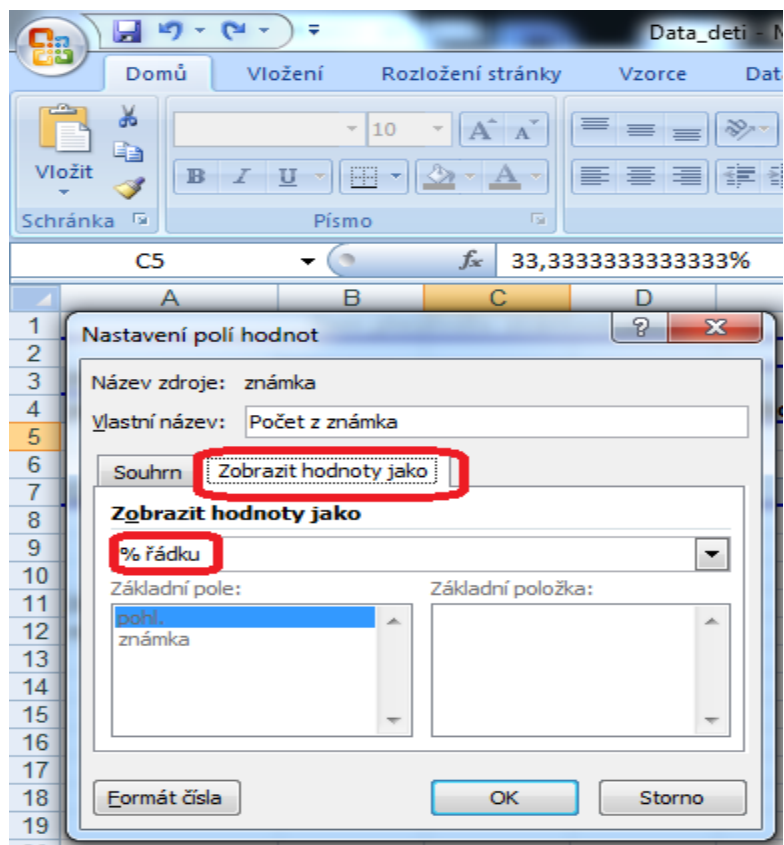
Pro znázornění, že je jedno, zda naše proměnné budou kvalitativní (slovní) či zapsané pomocí číselných hodnot, jsme proměnnou Pohlaví, ve které značila 0 ženu a 1 muže, přepsali do slovních hodnot a proměnnou známka jsme přepsali do proměnné známka z těl. slovně (1-výborně, 2-chvalitebně,3-dobře). Vidíme, 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 TĚL	ZNÁMKA Z TĚL			
POHLAVÍ	VÝBORNĚ	CHVALITEBNĚ	DOBŘE	Celkový součet
ŽENA	59	32	5	96
MUŽ	135	30	6	171
<b>Celkový součet</b>	<b>194</b>	<b>62</b>	<b>11</b>	<b>267</b>

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 danému sloupci. Poslední řádek a sloupec 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  $\Sigma$  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.

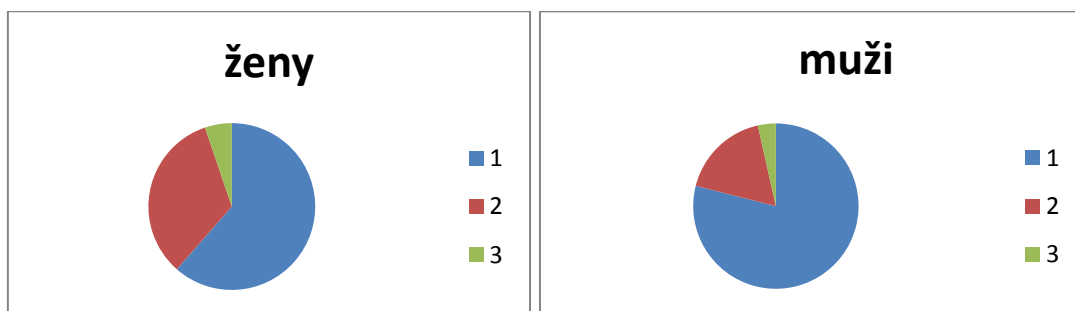




Po odkliknutí se nám vlevo na listě objeví následující tabulka.

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

Z této tabulky vidíme, že mezi ženami bylo 61,46 % jedniček 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 Statistika 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.

Kontingenční tabulka (Data\_deti\_min)  
Četnost označených buněk > 10  
(Marginální součty nejsou označeny)

POHLA	ZNAM 1	ZNAM 2	ZNAM 3	Řádk. součty
0	59	32	5	96
1	135	30	6	171
Vš. skup.	194	62	11	267

### 3.2.2 Testy nezávislosti

K vyšetřování vzájemné závislosti dvou kategoriálních znaků se používá často *test chí-kvadrát 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, opišeme četnosti včetně posledního sloupce a řádku. Toto budou naše aktuální (empirické četnosti). Pod tyto aktuální četnosti budeme psát teoretické, tedy 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 lomenu (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.

The screenshot shows an Excel spreadsheet with the following data:

Počet z	ZNÁMKA Z TĚL	CHVALITEBNĚ	DOBŘE	Celkový součet
ŽENA	59	32	5	96
MUŽ	135	30	6	171
<b>Celkový součet</b>	<b>194</b>	<b>62</b>	<b>11</b>	<b>267</b>

aktuální (empirické) četnosti				celkem
	59	32	5	96
	135	30	6	171
celkem	194	62	11	267

očekávané(teoretické) četnosti				celkem
	69,75280899	22,29213483	3,95505618	96
	124,247191	39,70786517	7,04494382	171
celkem	194	62	11	267

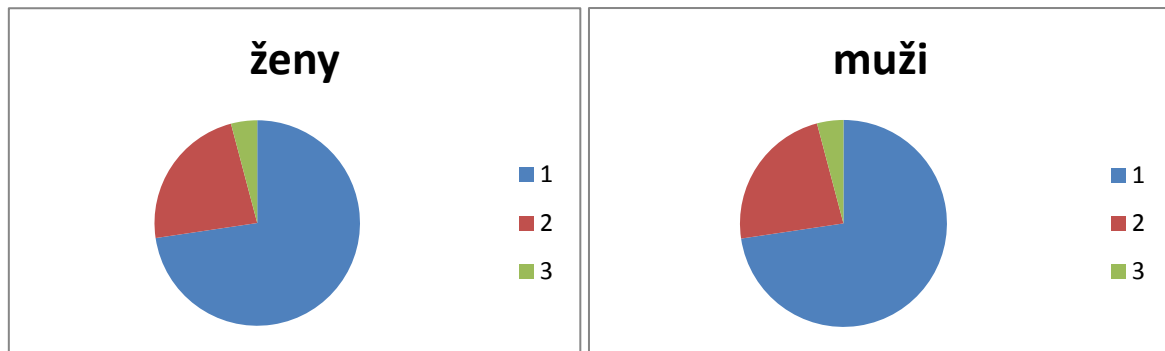
p= 0,008146694 zamítáme H<sub>0</sub>

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 Statistika 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.

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.

Souhrnná tab.: Očekávané četnosti (Data\_deti\_min)  
 Četnost označených buněk > 5  
 Pearsonův chí-kv. : 9,62029, sv=2, p=,008147

POHLA	ZNAM 1	ZNAM 2	ZNAM 3	Řádk. součty
0	69,7528	22,29213	3,95506	96,0000
1	124,2472	39,70787	7,04494	171,0000
Vš.skup.	194,0000	62,00000	11,00000	267,0000

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 měř 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: Statistika -> Základní statistiky -> Kontingenční tabulky -> OK -> Specif. tabulky (vyberte proměnn.) -> OK -> OK -> na záložce Možnosti vybereme například Pearsonův & M-V chí-kvadrát 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.

STATISTICA Cz - Data\_deti\_min

Data: Data\_deti\_min (20s krát 267ř)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	TRIDA	VEK	POHLA	ZNAM	VYSKA	HMOTN	OBLIB	DOVED											VYKON	BMI
1	7	13	0	1	160	52	1												0	20,3
2	7	13	0	1	165	57	1												2	20,9
3	8	13	0	1	172	60	1												1	20,3
4	8	13	0	1	169	55	1												1	19,3
5	8	13	0	1	159	51	1												1	20,2
6	7	13	0	1	162	44	1												0	16,8
7	8	13	0	1	156	53	1												0	21,8
8	8	13	0	1	158	49	1												1	19,6
9	7	13	0	1	159	49	1												2	19,4
10	7	13	0	1	174	67	1												1	22,1
11	7	13	0	1	165	51	1												2	18,7
12	8	13	0	2	166	55	1												1	20,0
13	7	13	0	2	160	55	0												0	21,5
14	7	13	0	1	163	50	1												1	18,8
15	8	13	0	1	164	48	1												1	17,8
16	7	13	0	1	156	46	1												1	18,9
17	7	13	0	1	155	47	1												1	19,6
18	7	13	0	1	160	50	1												1	19,5

Statist. : POHLA(2) x ZNAM(3) (Data\_deti\_min)

Statist.	Chi-kvadr.	sv	p
Pearsonův chí-kv.	9,620286	df=2	p=,00815
M-V chí-kvadr.	9,387711	df=2	p=,00915
Fí	,1898184		
Kontingenční koeficient	,1864885		
Cramér. V	,1898184		

V tabulce vidíme znovu hodnotu statistiky chí-kvadrát a výslednou p-hodnotu testu chí-kvadrát nezávislosti. Dále vidíme tři hodnoty měr síly závislosti. Protože proměnná POHLA má pouze dvě kategorie, hodnota koeficientu  $F_i$  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  $(-1,1)$  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 pořadí 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.

The screenshot shows the Statistica software interface. A contingency table is displayed with columns labeled TRIDA, VEK, POHLA, ZNAM, VYSKA, HMOTN, OBLIB, DOVED, and VYKON. A dialog box titled 'Výsledky: kontingenční tabulky: Data\_deti\_min' is open, showing options for calculating various statistical tests. The 'Detailní výsledky' tab is selected, and several tests are checked: Pearsonův & M-V chí-kvadrát, Kendalovo tau-b & tau-c, Goodmanovo-Kruskalovo gama, Spearmanova korelace, and Somerovo d.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	TRIDA	VEK	POHLA	ZNAM	VYSKA	HMOTN	OBLIB	DOVED												VYKON
1	7	13	0	1	160	52	1													0
2	7	13	0	1	165	57	1													2
3	8	13	0	1	172	60	1													1
4	8	13	0	1	169	55	1													1
5	8	13	0	1	159	51	1													1
6	7	13	0	1	162	44	1													0
7	8	13	0	1	156	53	1													0
8	8	13	0	1	158	49	1													1
9	7	13	0	1	159	49	1													2
10	7	13	0	1	174	67	1													1
11	7	13	0	1	165	51	1													1
12	8	13	0	2	166	55	1													1
13	7	13	0	2	160	55	0													0
14	7	13	0	1	163	50	1													1
15	8	13	0	1	164	48	1													1
16	7	13	0	1	156	46	1													1
17	7	13	0	1	155	47	1													1
18	7	13	0	1	160	50	1													1
									1	146	2	30	2	12,0	3	4,5	2	9		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 měř 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.



Statist.	Chi-kvadr.	sv	p
Pearsonův chí-kv.	5,354532	df=6	p=,49921
M-V chí-kvadr.	5,357775	df=6	p=,49881
Kendall. tau b & c	b=,0209368	c=,0205212	
Somers. D(X Y), D(Y X)	X Y=,02204	Y X=,01988	
Gama	,0319430		
Spearmanovo poř. R	,0244501	t=,39739	p=,69140

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.

The screenshot shows the 'Možnosti' (Options) dialog box for 'Výsledky; kontingenční tabulky: Data\_deti\_min'. The 'Možnosti' tab is active, and the 'Možnosti' section is highlighted. The 'Statistiky detailních 2-rozměrných tab.' (Statistics of detailed 2-dimensional tables) section is checked, and the 'Pearsonův & M-V chí-kvadrát' (Pearson's & M-V chi-square) and 'Fisher exakt., Yates, McNemar (2 x 2)' (Fisher exact, Yates, McNemar (2 x 2)) options are selected. The 'Yates' option is also checked, indicating the Yates correction is applied.

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, McNernan (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.

Statist.	Chí-kvadr.	sv	p
<b>Pearsonův chí-kv.</b>	<b>31,16010</b>	df=1	p=,00000
M-V chí-kvadr.	24,53964	df=1	p=,00000
Yatesův chí-kv.	28,46611	df=1	p=,00000
Fisherův přesný, 1-str.			p=,00000
Fisherův přesný, 2-str.			p=,00000
McNemarův chí-kv. (A/D)	156,5430	df=1	p=0,0000
McNemarův chí-kv. (B/C)	6,282609	df=1	p=,01219

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.

pohlaví	2-rozměrná tabulka: Kouření		Řádkové součty
	0	1	
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: Statistika -> 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.

Statist. : pohlaví(2) x kouření(2) (poi			
Statist.	Chi-kvadr.	sv	p
<b>Yatesův chí-kv.</b>	<b>,7366071</b>	df=1	p=,39075
Fisherův přesný, 1-str.			p=,19697
Fisherův přesný, 2-str.			p=,24242
McNemarův chí-kv. (A/D)	,1250000	df=1	p=,72367
McNemarův chí-kv. (B/C)	0,000000	df=1	p=1,0000

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.

A	B
C	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.



Kontingenční tabulka (List1 v Data\_Pisemky\_zari\_2015)  
 Četnost označených buněk > 10  
 (Marginální součty nejsou označeny)

Pretest4uspesnost	Postest4uspesnost 0	Postest4uspesnost 1	Řádk. součty
0	65	29	94
1	2	0	2
Vš. skup.	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.

Statist.	Chi-kvadr.	sv	p
Yatesův chí-kv.	.0262809	df=1	p=,87122
Fisherův přesný, 1-str.			p=,48487
Fisherův přesný, 2-str.			p=1,0000
McNemarův chí-kv. (A/D)	63,01538	df=1	p=,00000
McNemarův chí-kv. (B/C)	21,80645	df=1	p=,00000

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.