Generování dat

Newsletter Statistica ACADEMY

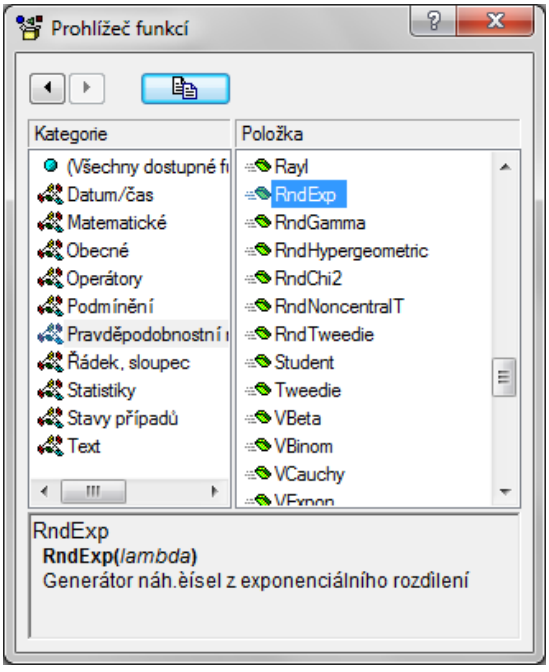
Téma: Simulace, pravděpodobnostní rozdělení
Typ článku: Návody

Pokud se dostanete do situace, kdy budete potřebovat nasimulovat data z nějakého rozdělení, Statistica Vám s tímto úkolem pomůže. Ukážeme si možnosti, jak generovat data z konkrétních rozdělení a to nejen jednorozměrných, ale i vícerozměrných.

Pokud byste si chtěli na začátek zopakovat, co je pravděpodobnostní rozdělení a jaká rozdělení známe, pak se může hodit jeden z našich starších ***článků***.

Generování pomocí funkcí

První možností, jak nagenerovat data z konkrétního rozdělení přímo do proměnné, je pomocí funkcí. Rozkliknete dialog proměnné a do okénka pro vzorec vybereme jednu z funkcí, které generují náhodná čísla. Poznáte je jednoduše, tyto funkce začínají písmeny ***rnd***.



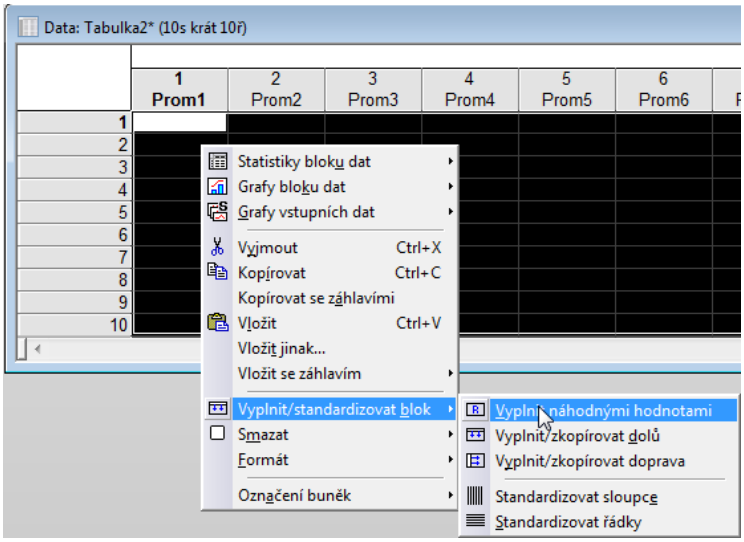
Nejklasičtějším zástupcem je funkce ***rnd(x)***, která generuje čísla z intervalu **(0,x)** a to rovnoměrně. Tato funkce je nejpoužívanější a jistě nejužitečnější. Náhodná čísla takto vygenerovaná lze použít mnoha způsoby pro nejrůznější účely. Zmiňme několik z nich:

- Můžeme využít pro generování dat z jiných rozdělení: simulace hodu mincí: ***=round(rnd(1))***
- Simulace hodu kostkou: ***=floor(rnd(6))***
- Transformace náhodné veličiny ze spojitého rovnoměrného rozdělení v intervalu (0;1) mohou dát i jiná rozdělení.
- Vygenerování identifikátoru trénovací a testovací

množiny. Pokud chceme zastoupení přibližně 70 ku 30, tak to zajistíme například takto: ***=iif(rnd(1)>0,3;"Trénovací";"Testovací")***

- Náhodně vybrat nějaké množství případů – stačí vytvořit pomocnou proměnnou ***=rnd(1)***, seřadit soubor podle této proměnné a vybrat požadovaný počet případů.

Když už jsme u generování náhodných čísel z intervalu (0;1), zmiňme možnost vyplnit blok dat náhodnými čísly – pokud potřebujete rychle vyplnit nějakou oblast daty (například v případě, že chcete mít rychle po ruce data se spojitými proměnnými, na kterých si chcete rychle něco ve Statistice zkoušet), pak můžete využít funkcionalitu: ***Vyplnit náhodnými hodnotami*** (vyberete oblast, kliknete do ní pravým tlačítkem a vyberete ***Vyplnit/standardizovat blok –> Vyplnit náhodnými hodnotami***).



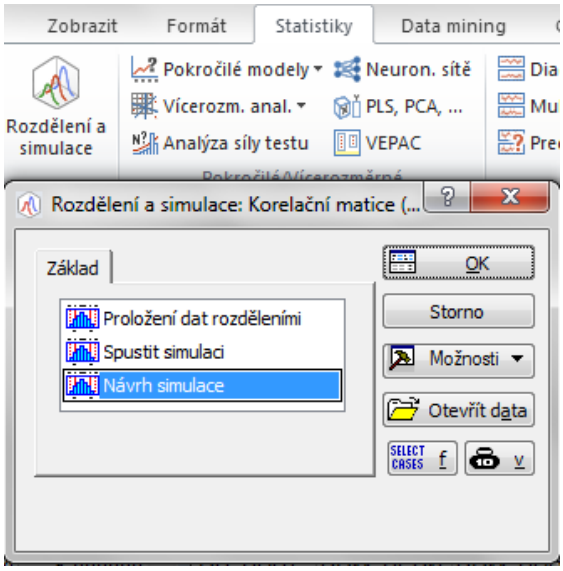
Generování vícerozměrných rozdělení

Pokud potřebujeme nasimulovat nějaké složitější chování, jistě se nám bude hodit funkcionalita pro generování vícerozměrných dat.

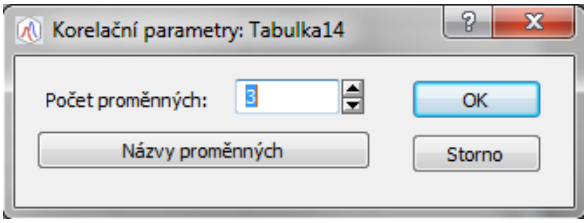
Můžeme si volit typ rozdělení pro danou veličinu, stejně tak korelaci mezi veličinami.

Máme několik možností, jak si zvolit schéma pro generování:

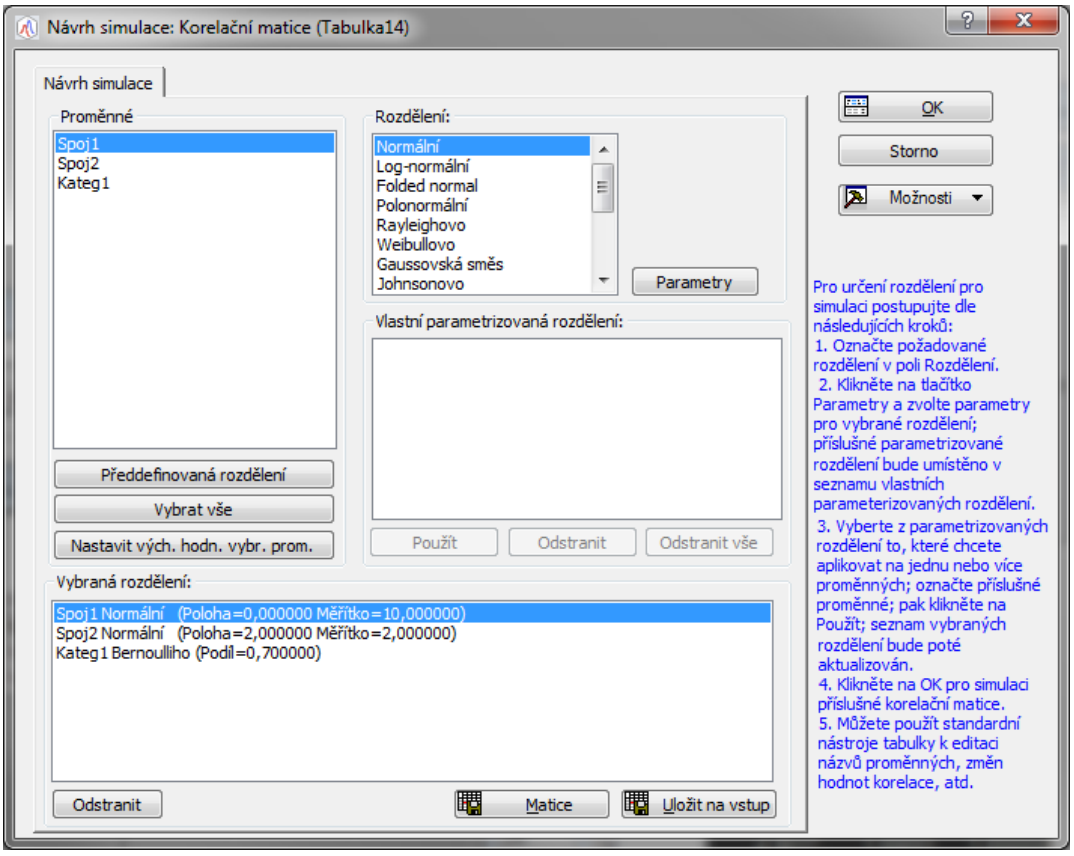
Nastavíme si přesné parametry rozdělení veličin i korelací sami:



Funkcionality pro generování a odhad vícerozměrných rozdělení najdete ve ***Statistiky –> Simulace a rozdělení***. Pro kompletně nový návrh vybereme možnost ***Návrh simulace***. Po stisknutí ***OK*** se objeví hláška, že chybí vstup ve formě matice. Matice je jeden z formátů Statistica a má svůj speciální tvar. Statistica jej v tomto kroku potřebuje, protože v ní jsou uložena rozdělení, stejně jako korelační struktura simulovaných dat.



My ale teprve chceme návrh vytvořit, hláška je tedy v pořádku a zmáčkneme ***OK***. Objeví se nám dialog pro specifikaci počtu a jmen proměnných. Vyplníme ji podle toho, jak potřebujeme, a zmáčkneme ***OK***. Objeví se nám matice návrhu (defaultně nastavená) a také dialog pro specifikaci rozdělení. Nejprve specifikujte korelace mezi veličinami přímo v matici. Poté přejděte do dialogu a nadefinujte zde rozdělení a jejich parametry: na pravé straně dialogu je podrobný návod, co dělat...



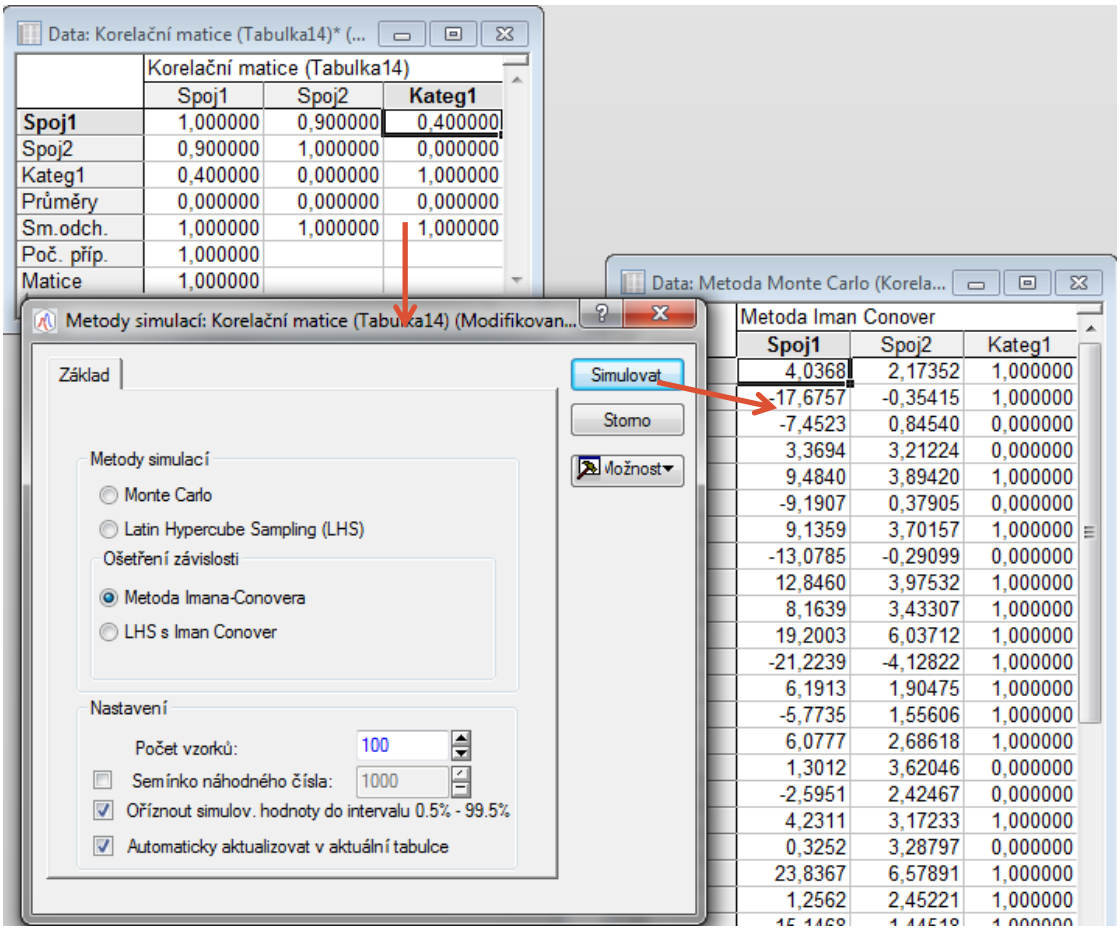
Po nastavení rozdělení klikneme na ***Uložit na vstup***, což uloží vybraná rozdělení do matice. Toto je důležitá informace, formát matice v sobě má uloženu i informaci o rozděleních, i když to není navenek vidět. Dejte si na to pozor, i dvě stejně vypadající matice mohou být různé pro generování náhodných čísel. Poté zmáčkněte ***OK*** v dialogu návrhu simulace.

Objeví se okno pro specifikaci simulace, můžeme si vybrat více simulačních algoritmů. Pokud chceme generovat data s přihlédnutím k nadefinované korelační struktuře, musíme vybrat některou z metod v sekci ***Ověření závislosti***.

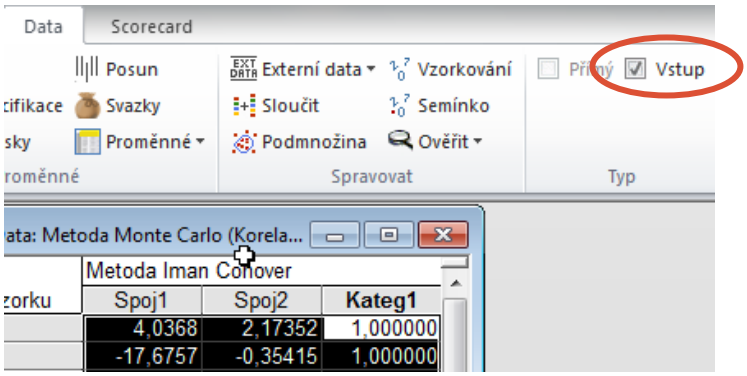
Ostatní nastavení jsou intuitivní. Počet vzorků je počet řádků, které chceme vygenerovat.

Možnost ***Oříznout simul. hodnoty do intervalu***... ořízne nejodlehlejší hodnoty, které vznikly generováním.

Zaškrtnutí ***Automaticky aktualizovat v aktuální tabulce*** říká, že po opakovaném stisknutí tlačítka ***Simulovat*** se budou předchozí nasimulovaná data přepisovat a nebude vytvořen nový soubor.

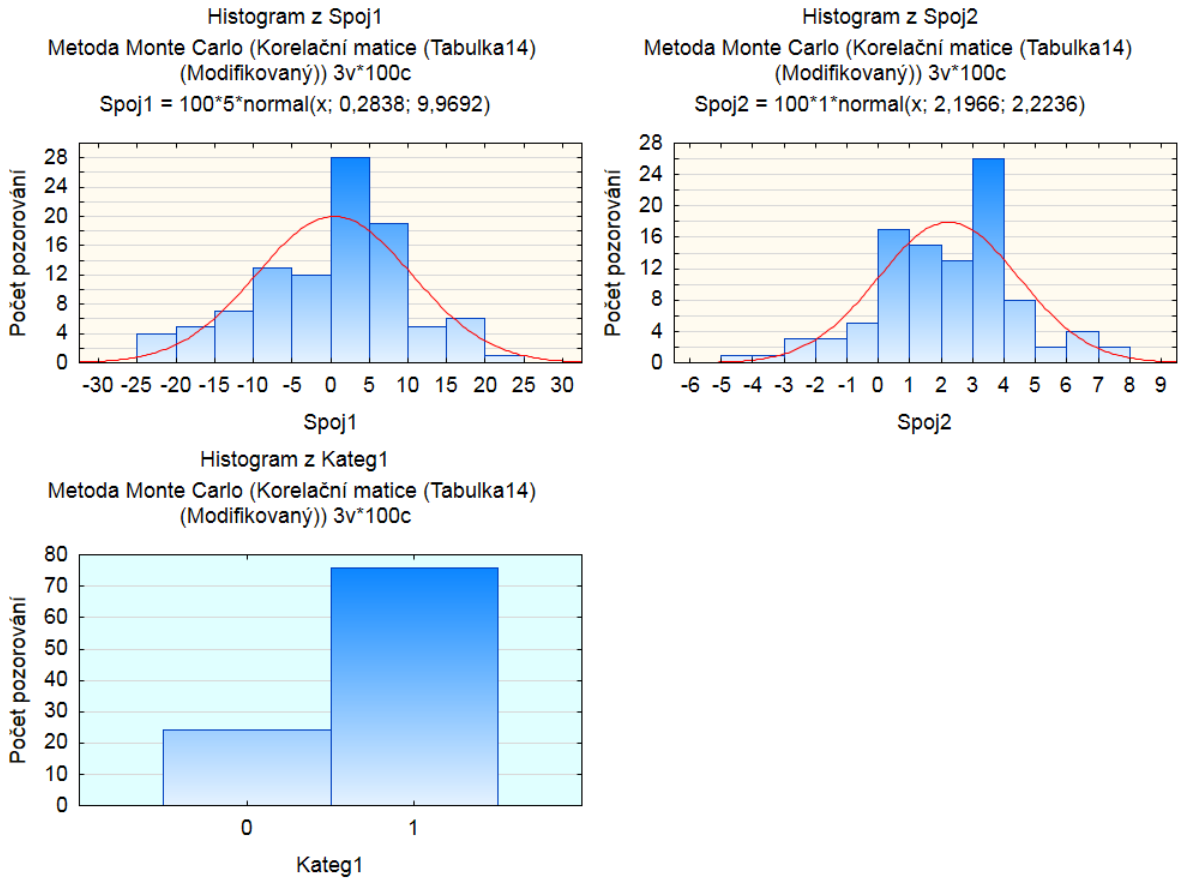


S výslednými daty chcete většinou ihned nějak pracovat. Je potřeba zmínit, že po vygenerování dat není tabulka vnímána jako aktivní tabulka pro analýzu, ale jen jako výsledek. Proto nejdřív musíte zaškrtnout zaškrtávátko ***Vstup*** v záložce ***Data***. Nyní je možné s daty pracovat a dělat na nich analýzy. My nyní zkusíme spočítat korelace a popisné statistiky, abychom zjistili, zda nagenerovaná data splňují to, co jsme do simulace zadali.



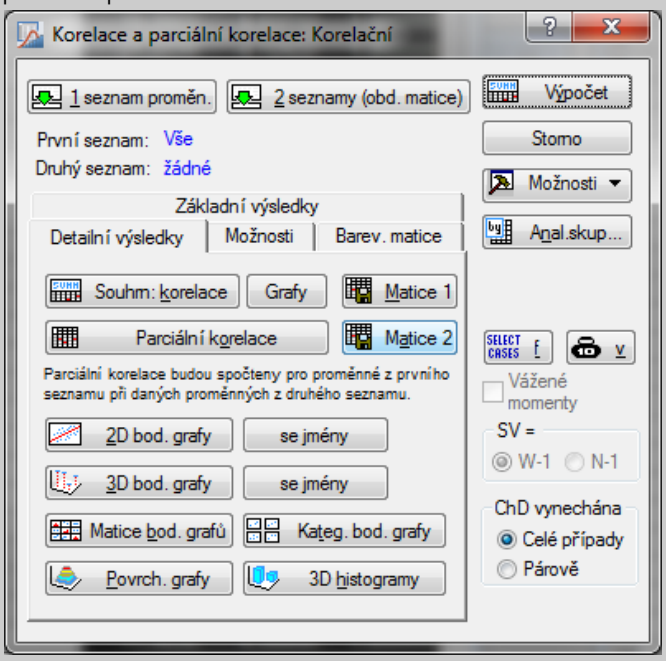
Korelační matice i histogramy odhalují, že data jsou nagenerována přesně podle našeho zadání (všimněte si, že jsme generovali ze spojitých i diskrétních rozdělení):

Proměnná	Průměry	Sm.odch.	Spoj1	Spoj2	Kateg1
Spoj1	0,283764	9,969199	1,000000	0,893977	0,340210
Spoj2	2,196597	2,223638	0,893977	1,000000	0,069122
Kateg1	0,760000	0,429235	0,340210	0,069122	1,000000



Matici, podle které lze generovat data, si můžete uložit ve formátu s koncovkou ***smx***.

Poznámka: Formát matice je možné získat jako výstup u některých metod, jako příklad uveďme výstup z korelační analýzy. Dále je možné příklad formátu matice najít v datových souborech příkladů pod názvem ***CorrelationMatrix.smx***.

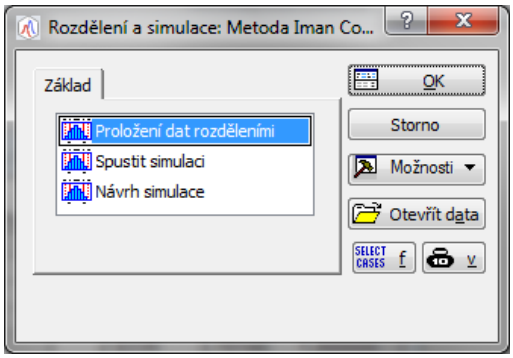


Parametry simulace se nastaví na základě odhadu z dat:

Tato možnost je běžně využívána v nejrůznějších aplikacích, kdy už máme nasbírán vzorek dat a na jejich základě se ptáme například, jak často může nastat nějaká situace. Na reálných datech provedete odhad pravděpodobnostní struktury a poté můžete ze stejné struktury simulovat další a další data, což může pomoci například pro zjištění chování dat, aniž bychom data reálně sbírali či měřili. Krásný příklad na toto téma najdete v nápovědě: ***Statistics – Analyses – Distributions & Simulations – Distributions &Simulation Example.***

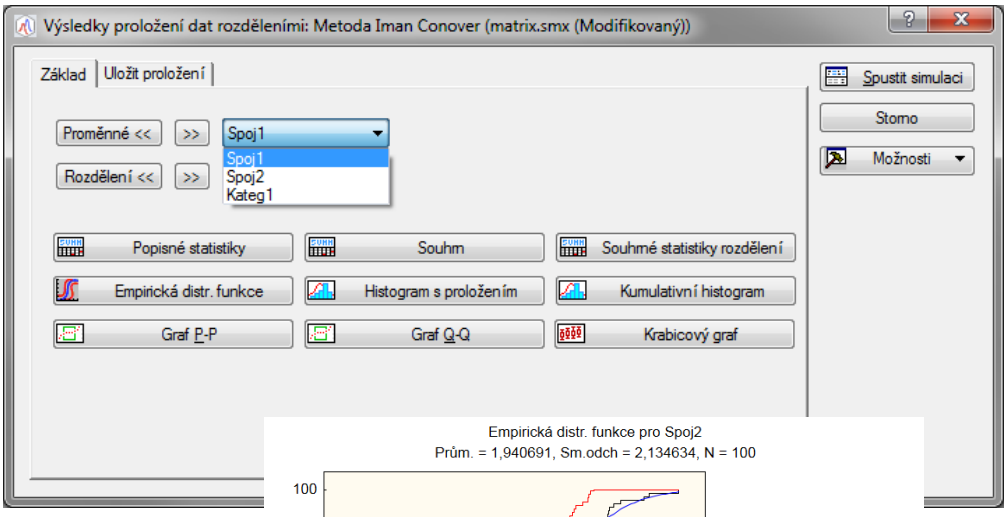
Jak to provedeme v praxi? Ve zkratce: Máme data, na nich zjistíme, jakého jsou rozdělení a jakou mají korelační strukturu a na základě této parametrizace budeme simulovat data nová.

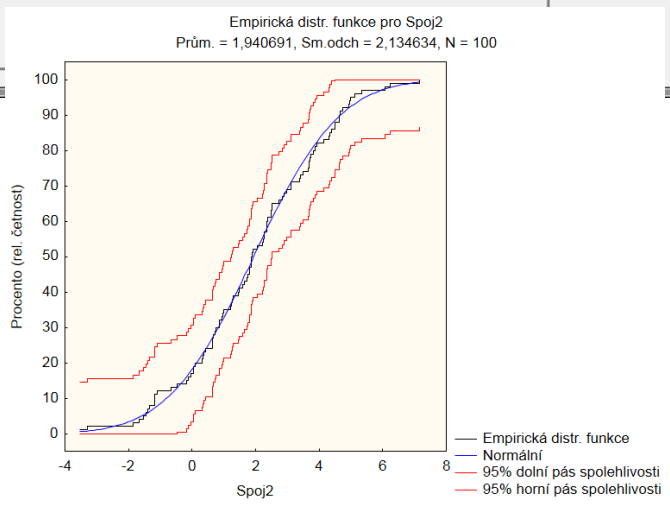
My použijeme pro náš příklad jako náhražku reálných dat výstup z minulého bodu, nicméně Vy můžete pracovat s jakýmkoli souborem, na kterém chcete odhadnout jeho pravděpodobnostní rozdělení. My máme tedy 3 rozměrná data se dvěma spojitými a jednou kategorickou veličinou.



Postup: Máme otevřena jen tato data, matici návrhu budeme teprve tvořit. Vybereme: ***Statistiky – Rozdělení a simulace – Proložení dat rozděleními***. Zvolíme proměnné, pokud máme představu, z jakých rozdělení mají být data, pak zvolíme i prostor posuzovaných rozdělení v záložkách ***Spojité/Diskrétní proměnné***. Klikneme na ***OK***.

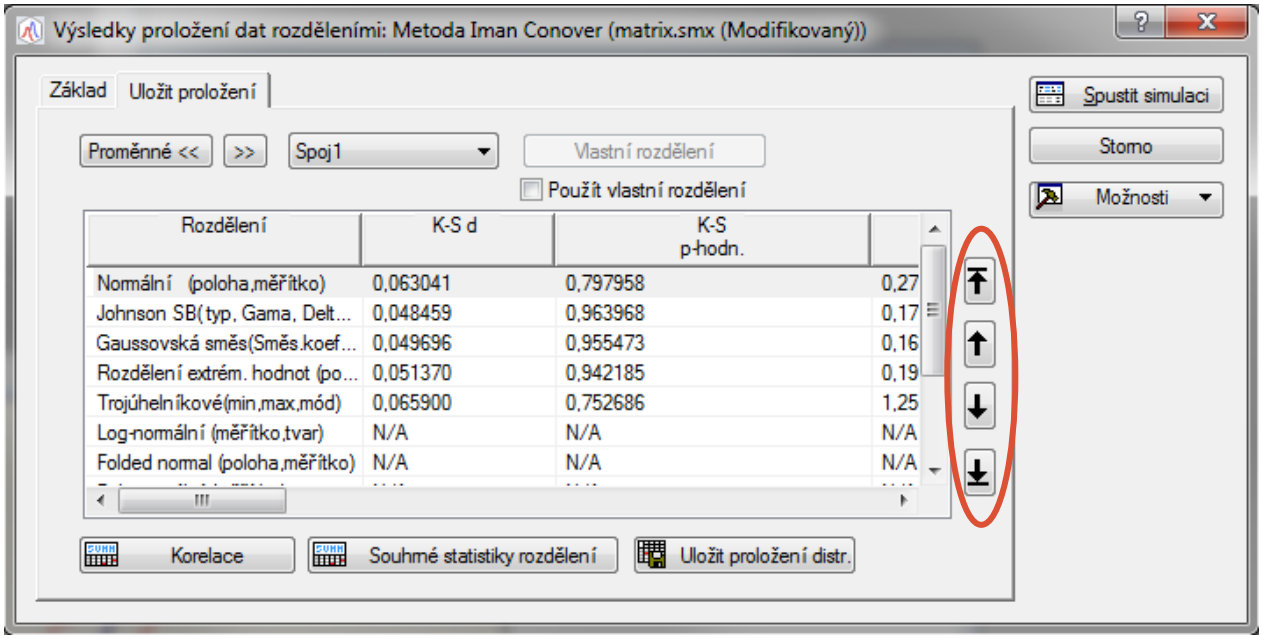
Ve výsledcích si můžeme vybrat proměnnou, provést nejrůznější diagnostické grafy, podívat se na výsledky testů a podle těchto nástrojů určit, jaké je rozdělení každé z veličin.



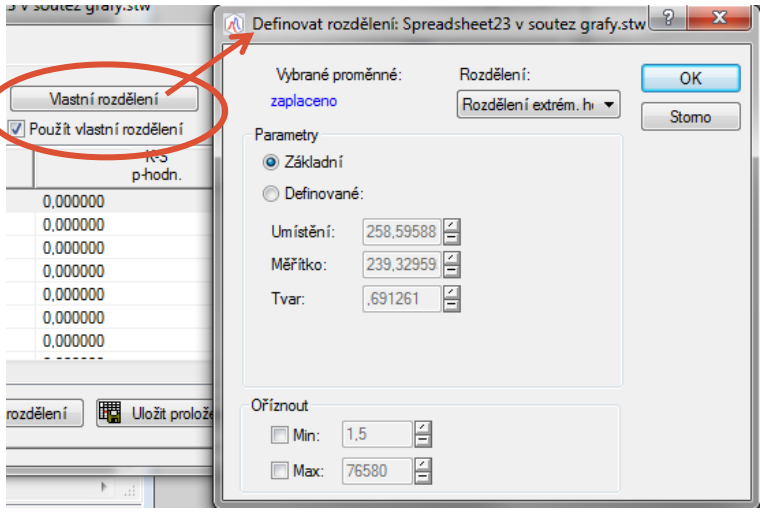


Jak uložit rozdělení:

Pomocí tlačítka ***Uložit proložení distr.*** můžete vygenerovat matici návrhu simulace s vybranými rozděleními, včetně korelační matice. Uloží se distribuce, které byly nejvýše v seznamu rozdělení v záložce ***Uložit proložení***. Pokud chcete změnit pořadí a dát na první místo rozdělení, které podle Vás data popisuje, pak použijte šipky vpravo:



Pokud chcete vybrat konkrétní rozdělení i jeho parametry, pak zaškrtněte zaškrkávátko Použít vlastní rozdělení a po kliknutí na Vlastní rozdělení můžete nadefinovat konkrétní rozdělení, dokonce i s hranicemi odkud kam mohou data sahat, tedy oříznutí:

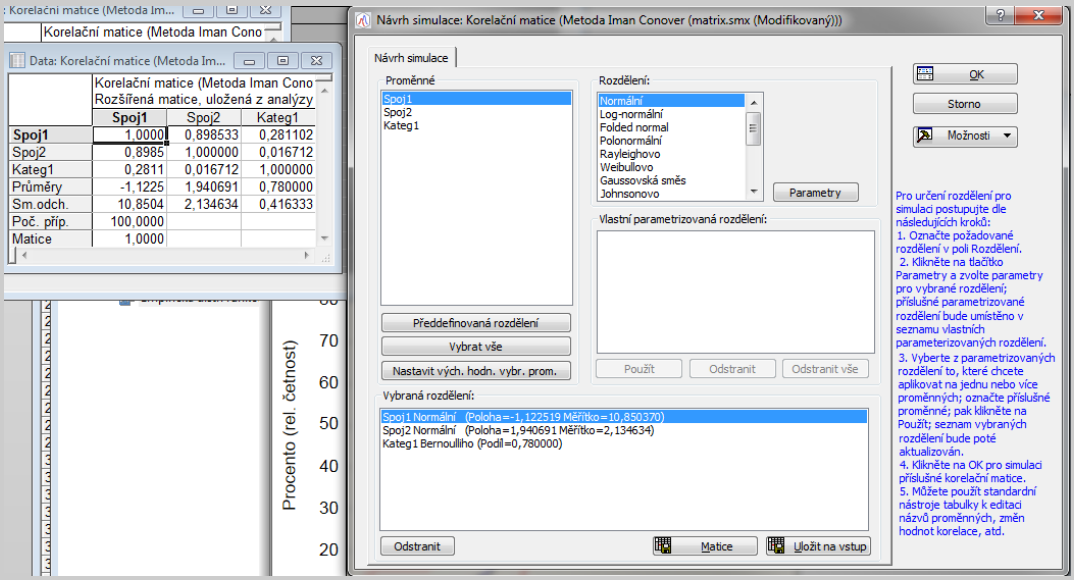


Pokud již máte korelační matici i s rozděleními:

Pak jsou dvě možnosti: buď využít možnost ***Statistiky –> Simulace a rozdělení –> Spustit simulaci*** a přímo pouze vybrat parametry simulování bez možnosti nějak kontrolovat nebo měnit parametry rozdělení. Druhou možností je stejný postup jako v prvním případě, tedy přes ***Statistiky –> Simulace a rozdělení –> Návrh simulace***, jen již není potřeba specifikovat rozdělení, ta se načtou z informací

v souboru matice. Výhodou tohoto postupu je, že si můžeme zkontrolovat či změnit rozdělení, ze kterých generujeme data.

Poznámka: Pokud si chcete ověřit, jaká rozdělení se skrývají pod korelační maticí, pak stačí vyvolat ***Návrh simulace*** na tuto korelační matici.



Závěr

V tomto článku jsme Vám odkryli další část softwaru. Věříme, že ji využijete nejen při simulacích, ale také při hledání rozdělení pro Vaše data. Ukázané funkcionality totiž obsahují mnoho zajímavých grafických výstupů i testů, podle kterých se můžete řídit.

Vystižení struktury dat a možnost generovat data na základě této struktury, může usnadnit mnoho práce a času a také odpovědět na důležité otázky, které Vás trápí.

Věříme, že možnosti generovat náhodná čísla nejen z jednorozměrných, ale také vícerozměrných rozdělení, využijete.

Dell | Information Management | Software Group
V Parku 2325/16, 148 00 Praha 4
t +420 233 325 006, **e** info@statistica.cz
www.dell.cz/statsoft, www.statsoft.cz