

Jak si pomoci simulacemi

Newsletter Statistica ACADEMY



Téma: Simulace, pravděpodobnostní rozdělení
Typ článku: Příklad

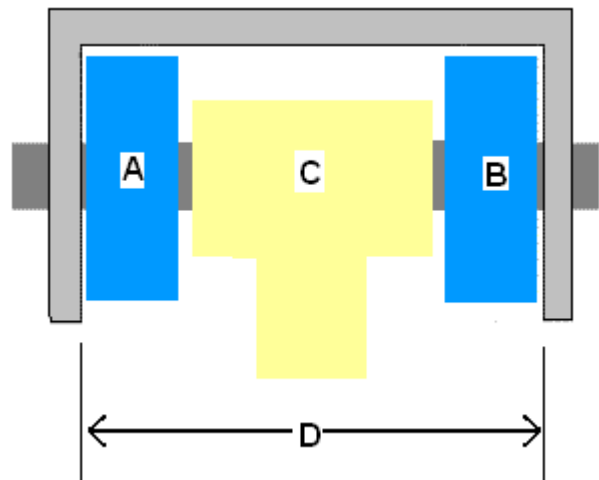
Dnes si ukážeme zajímavý příklad, jak využít simulaci pravděpodobnostního rozdělení k vypočítání složitějšího úkolu, který bychom dokázali vyřešit bez umělého generování dat jen těžko nebo nákladně.

Zadání

Představme si, že vyrábíme součástky pro panty dveří podobné jako na schématu vpravo. Jsou zde tedy 4 součástky, výrobek je považován za zmetek, pokud šířky $A+B+C$ přesáhnou šířku D , tedy první 3 součástky se nevejdou do součástky čtvrté.

Zadáním je zjistit (odhadnout) procento defektních výrobků v budoucí výrobě.

Místo toho, abychom museli čekat na reálnou hromadnou produkci a stanovovat počty na finálních výrobních, můžeme si pomoci simulacemi na základě vzorku dat (může to být například nějaká malá zkušební série).



Nasbírali jsme data na malém výrobním vzorku, informace jsou v souboru [SimulationRiskData.sta](#) v příkladech (**Soubor -> Otevřít příklady -> Datasets**)

Řešení

Abychom získali výsledky, využijeme funkcionalitu **Rozdělení a simulace**. Možnosti, které můžete v této souvislosti využít, jsme si vysvětlovali v článku [Generování dat](#).

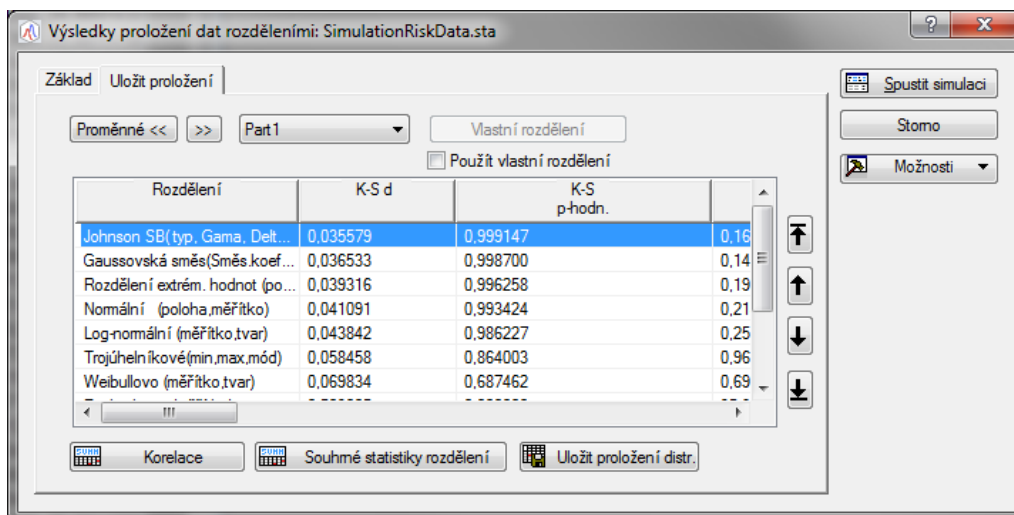
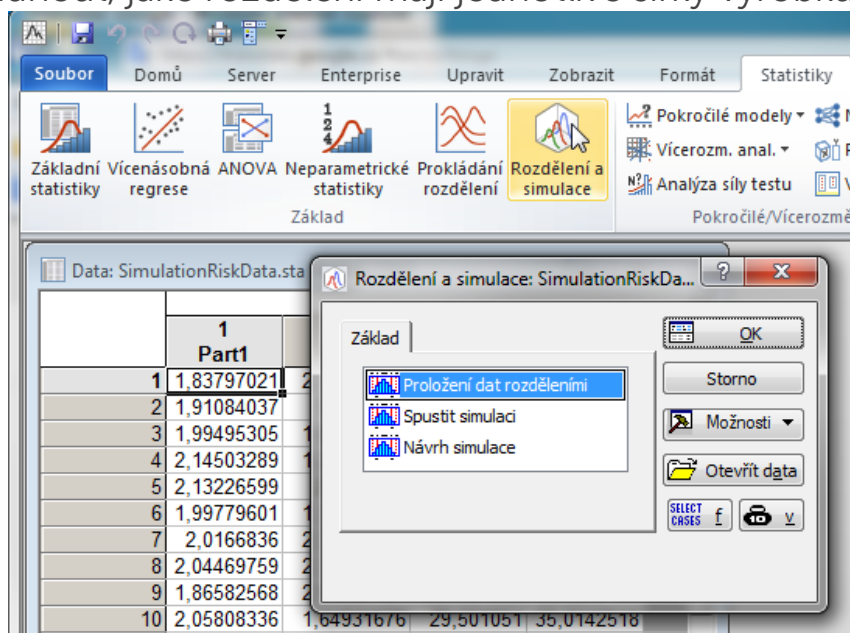
Tímto příkladem si tedy zopakujeme, jak postupovat, abychom odhadli rozdělení dat a následně z něj vygenerovali náhodná data, která použijeme k řešení úlohy.

1. Nejprve zkusíme odhadnout, jaké rozdělení mají jednotlivé šířky výrobků.

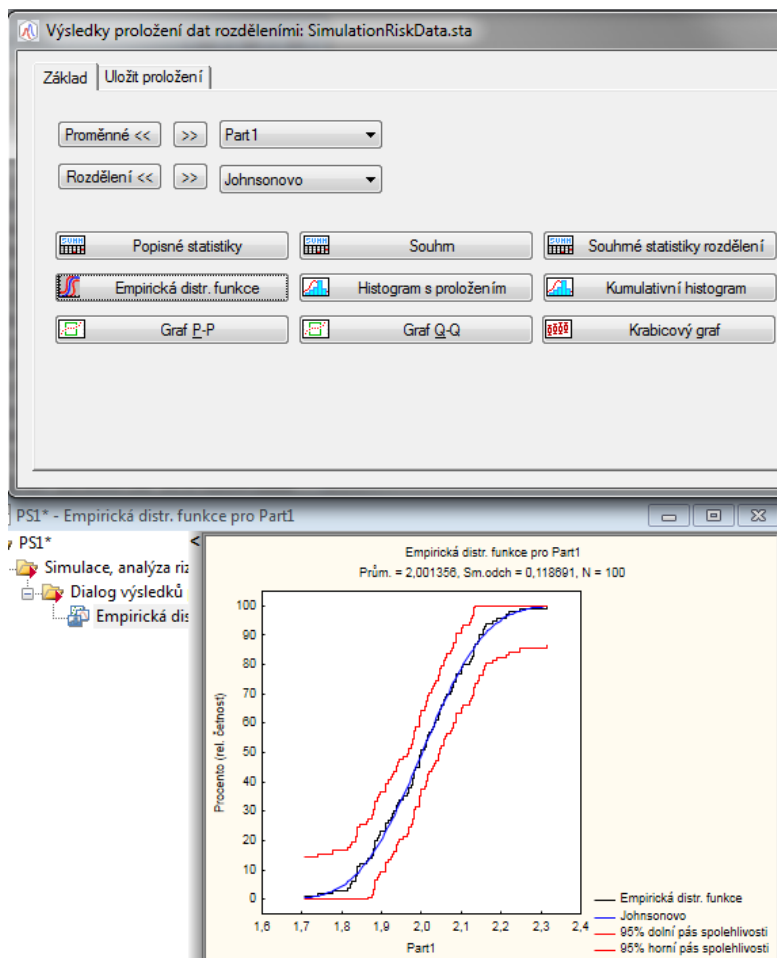
Klikneme na **Statistiky** -> **Rozdělení a simulace** -> **Proložení dat rozděleními** -> **OK**. Jelikož jsou všechny 4 proměnné šířkou nějakého výrobku, jedná se o spojitě veličiny (vybereme tedy všechny jako spojitě veličiny, které budou vstupovat do analýzy). Bez nutnosti dalšího nastavování klikneme **OK** a tím necháme provést výpočty.

V záložce uložit proložení se můžete podívat, která

rozdělení software vybral jako nejvhodnější k té které proměnné (přepínat mezi proměnnými lze v rozevírací liště, kde je nyní defaultně vybrána proměnná *Part1*). Pro *Part1* a *Part4* bylo vybráno jako nepřihodnější (na základě Kolmogorov-Smirnovovy charakteristiky) Johnsonovo rozdělení, které je čtyřparametrovou transformací rozdělení normálního. Pro proměnné *Part2* a *Part3* jde o směs normálních rozdělení.



V záložce **Základ** můžete vizualizovat výsledky analýz. Pokud bychom se chtěli podívat na kumulativní distribuční funkci či nějakou jinou charakteristiku pro zvolené rozdělení pro veličinu *Part1*, pak vybereme ze seznamu proměnných *Part1* a ze seznamu rozdělení **Johnsonovo** a zmáčkneme tlačítko **Empirická distribuční funkce** (či tlačítko jiných diagnostických grafů či charakteristik).



Pokud si detailně projdeme tyto diagnostické grafy pro všechny veličiny pro nejvhodnější rozdělení, pak shledáme, že vybrané distribuce celkem rozumně odpovídají naměřeným datům. Můžeme se koukat na empirické distribuční funkce nebo na Q-Q graf.

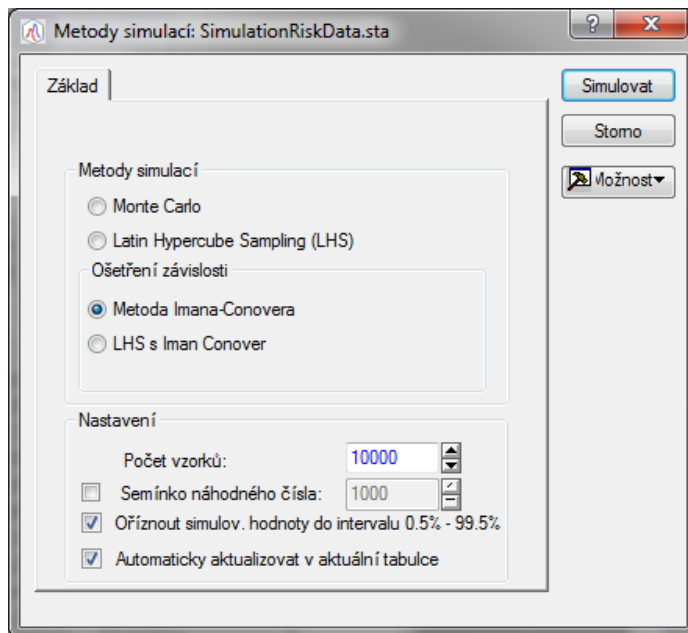
Můžeme tedy přejít ihned ke generování nových dat.

Poznámka: O tom, jak poznat zda rozdělení odpovídá například na základě Q-Q grafu, se můžeme dočíst v [článku o Q-Q grafech](#). I když se jedná o článek o normálním rozdělení, princip, jak interpretovat graf je pořád stejný.

2. Generování dat na základě vybraných rozdělení.

Nyní máme zjištěno teoretické rozdělení dat a nyní bychom chtěli této skutečnosti využít a nagenarovat si na základě tohoto data nová a prověřit to, co nás zajímá, tedy počet zmetků. Generování je možné provést dokončením právě otevřeného dialogu pro výběr rozdělení (další možností je uložit si rozdělení ve formě matice a to později použít v sekci **Návrh simulace** nebo **Spustit simulaci** – jsou v dialogu pro **Rozdělení a simulace**, tento postup je popsán v článku [Generování dat](#)). Máme zvolena rozdělení, která chceme generovat – tedy žádaná rozdělení jsou nejvýše v dialogu **Výsledky proložení dat rozděleními** (záložka **Uložit proložení**) pro každou proměnnou. Poté stačí zmáčknout tlačítko **Spustit simulaci**, které vyvolá dialog pro nastavení simulace. Máme zde několik algoritmů, pokud chceme zachovat korelační matici dat, pak vybereme algoritmus v sekci **Ošetření závislosti**. Dále můžeme zvolit počet dat, která chceme generovat, zda chceme oříznout extrémní hodnoty v datech a jestli bude opakované spuštění generování přepisovat tabulku nebo se budou tvořit tabulky nové.

V našem příkladu zvolíme možnosti následovně: Protože víme, že výroba jednotlivých součástí je na sobě závislá, pak zvolíme metodu ze sekce **Ověření závislosti**. Budeme chtít 100 000 hodnot s vyloučením těch nejdlejších hodnot.



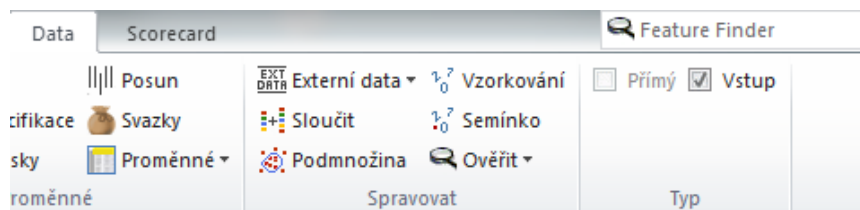
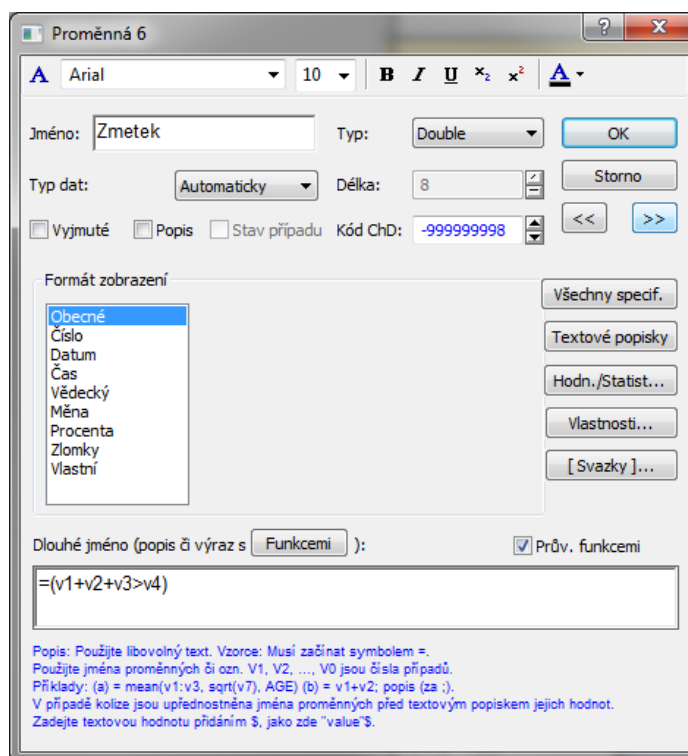
3. Vyvození závěrů.

Nyní máme nagenеровáno opravdu velké množství dat, které simulují vyrobené kusy, a stačí pouze zjistit, kolik z nich je nevyhovujících (byly by shledány jako zmetky).

K tomu vytvoříme novou proměnnou, kterou nazveme například *Zmetek* a vložíme do ní vzorec: $=(v1+v2+v3>v4)$, který do této proměnné vloží jedničku, pokud je šířka první plus druhé plus třetí součástky větší než součástka čtvrtá.

Nyní stačí pouze zanalyzovat nově vzniklou proměnnou *Zmetek* a tedy zjistit procentuální zastoupení zmetků.

Nejprve musíme zaktivnit nagenеровaná data jako data pro analýzu (tabulka výstupu simulace je totiž defaultně jako výstup analýzy a tedy se s ní nedá ihned pracovat, je potřeba ji zvolit jako vstupní tabulku – najedeme do **Data** do sekce **Typ** a zde zaškrtneme **Vstup**).



Abychom získali potřebné četnosti, můžeme postupovat různě, my provedeme výpočet například přes klasickou tabulku četností, tedy najedeme do sekce **Statistiky** -> **Základní statistiky** -> **Tabulky četností**, analyzujeme proměnnou *Zmetek*.

Výsledek je následující: očekávaný počet nevyhovujících součástí se bude pohybovat okolo 10ti procent.

Tabulka četností: Zmetek: =v1+v2+v3>v4 (Metoda Iman Cono				
Kategorie	Četnost	Kumulativní četnost	Rel. četnost	Kumulativní rel. četnost
0	89588	89588	89,58800	89,5880
1	10412	100000	10,41200	100,0000
ChD	0	100000	0,00000	100,0000

Poznámka: Ačkoli budete postupovat stejným způsobem jako my, nedosáhnete úplně stejného výsledku, jedná se o náhodná čísla a tedy pokaždé, co provedete generování náhodných čísel, výsledek Vám vyjde jinak. Aby se tomuto zabránilo a výsledky byly opakovatelné, je možné využít funkcionality nastavení semínka náhodného generátoru (překlad z anglického *seed*). To můžete udělat před vlastním generováním přes záložku **Data** -> **Semínko** (zde nastavíte konkrétní hodnotu semínka) nebo přímo v dialogu nastavení generování, který jsme používali. Můžete si to vyzkoušet, naše data jsme nagerovali s nastaveným parametrem funkce (semínko) náhodného generátoru na hodnotu 1.

Závěr

Tento článek byl ukázkou praktické aplikace, která může ušetřit spoustu peněz především ve fázi zavádění produktu do výroby. Nicméně simulace se dají využít jistě v mnoha dalších odvětvích a aplikacích.

Tento článek spolu s článkem [Generování dat](#) a článkem [Rozdělení náhodné veličiny](#) dá celkem dobrou představu o tom, jak vše funguje a jak si se simulacemi poradit. Dalším zdrojem inspirace může být i jiný příklad [simulací intervalů spolehlivosti](#)