StatSoft

Odkud tak asi je?

Ukážeme si, jak bychom mohli vypočítat pravděpodobnosti, na které jsme se ptali **v minulém newsletteru**. Úkolem bylo zjistit, z kterého kraje nejpravděpodobněji pochází náš výherce minulé soutěže pan Petr Hudík.

Data

Abychom mohli úlohu vyřešit, podívejme se nejdříve na to, jaká data máme k dispozici (na adrese: **http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx** si můžete inkriminovaná data stáhnout): Máme zde několik souborů obsahujících:

1. četnosti příjmení v obcích a krajích
2. četnosti příjmení podle roku narození
3. četnosti jmen v obcích a krajích
4. četnosti jmen podle roku narození
5. kontingenční tabulku mezi obcemi a rokem narození
6. soubor s kódováním pro obce, města, kraje

	PŘÍJMENÍ	3101	3102	3103
1	AADI	0	0	0
2	AAFJES	0	0	0
3	AALBREGT	0	0	0
4	AALDERS	0	0	0
5	AALDERSOVÁ	0	0	0
6	AAMANN	0	0	3
7	AANDERUD	0	0	0
8	AANENSENOVÁ	0	0	0
9	AAOUATIF	0	0	0
10	AARDOOM	0	0	0

Soubory četností obsahují i celkové řádkové (proměnná s kódem 3000) a sloupcové četnosti (řádek z názvem SOUČET), ty se určitě budou hodit.

Jak jste si mohli všimnout, chybí kontingenční tabulka mezi jmény a příjmeními dohromady, stejně jako zde není tabulka četností pro jména a příjmení společně. Máme tedy všechny agregované výsledky zvlášť pro jméno a příjmení, ale nemáme žádnou informaci o tom, jak se vyskytují společně.

Nějaké jednoduché pravděpodobnosti

Začneme povídáním o tom, co ze souborů dokážeme zjistit, přesněji jaké pravděpodobnosti. Máme zde několik zajímavých veličin: KRAJ (v soutěžní otázce nás zajímal kraj, detailnější dělení tedy zde nebudeme uvažovat), JMÉNO, PŘÍJMENÍ, ROK NAROZENÍ (ten zde taky nebudeme potřebovat). KRAJ nabývá hodnot 0,1,...,14. Proměnná JMÉNO (resp. PŘÍJMENÍ) nabývá hodnotu z množiny křestních jmen (resp. příjmení).

Důležité je uvědomit si, co všechno z dat víme. Zkusme napočítat nějaké pravděpodobnosti. Nejdříve jednoduché:

$P(\text{KRAJ} = i)$ je pravděpodobnost, že náhodně vybraný člověk náleží do kraje i. Vypočte se jednoduše jako počet lidí v daném kraji děleno celkovým počtem lidí v souboru (České republice). Stačí tedy využít řádek SOUČET v souboru četností, například četností příjmení a četnosti v krajích (kraje jsou v souboru zakódovány čísly 0 až 14, součet všech za všechny kraje pak najdeme ve sloupci s názvem 3000).

	0	14	1	2	3	4	5	6	7	8	9	10	11	12	13	3000
SOUČET	3	1108885	628150	1160587	291195	544175	430486	1226351	634912	507400	553201	1251617	815193	505383	585821	10243359
pravděpodobnost	2,93E-7	0,10825	0,06132	0,1133	0,02843	0,05312	0,04203	0,11972	0,06198	0,04953	0,05401	0,12219	0,07958	0,04934	0,05719	1

Vzniklo vydělením čísla nad ním (počtu lidí v daném kraji) číslem v prvním řádku ve sloupci 3000 (celkový součet řádků i sloupců, tedy počet všech lidí). Například pravděpodobnost, že jste z Prahy (14 je kód pro Prahu, viz soubor s kódy) je 0,108.

$P(\text{JMÉNO} = j)$ je pravděpodobnost, že je Vaše jméno j (řekněme, že jste například Petr). Vypočte se podobně jako u kraje, jen dělíme počty Petrů počtem všech osob – potřebujeme tedy jen sloupec s kódem 3000.

	JMÉNO	3000	pravděpodobnost
41159	PETJA	3	0,00000029
41160	PETKO	14	0,00000137
41161	PETKO NIKOLAJ	1	0,00000010
41162	PETMAT	1	0,00000010
41163	PETR	272852	0,02663697
41164	PETR ADAM	8	0,00000078
55440	ŽÓFIA	5	0,00000049
55441	ŽUDI	1	0,00000010
55442	SOUČET	1,0E+7	1,00000000

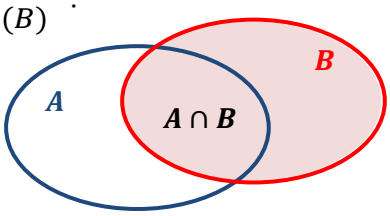
Vzniklo vydělením čísla ve sloupci 3000 v řádku, kde je PETR, číslem v řádku se součtem (celkovým počtem lidí – řádek SOUČET). Pravděpodobnost, že toto čte PETR, je tedy 0,0266 (jinak řečeno, v ČR máme 2,66% lidí, kteří mají křestní jméno Petr).

$P(\text{PŘÍJMENÍ} = k)$ je pravděpodobnost, že je Vaše příjmení rovno j. Vypočte se stejně jako u jmen, jen se použije soubor s četnostmi příjmení, jedinou drobnou komplikací je přítomnost více listů v souboru s příjmeními, nicméně řádek i sloupec se součty je v souboru také.

Podmíněné pravděpodobnosti

Nejprve trochu naznačme, co je podmíněná pravděpodobnost. Není to nic složitého, prostě nás zajímá pravděpodobnost, ale za nějaké podmínky – víme, že něco nastalo nebo se něco předpokládá a pak se ptáme, jaká je pravděpodobnost s tímto předpokladem či podmínkou.

Značení: $P(A\,|B)$ je podmíněná pravděpodobnost jevu A za podmínky, že nastal jev B (podmínky se píší za svislou čáru). A jak se spočítá? Celkem intuitivně jako $P(A\,|B) = \frac{P(A \cap B)}{P(B)}$.



Tedy zjednodušeně řečeno, díváme se na to, jakou část zabírá průnik A a B z celého B. Viz diagram:

Když přeneseme situaci na náš příklad, tak pravděpodobnost $P(\text{KRAJ} = \text{PRAHA}\,|\,\text{JMÉNO} = \text{PETR})$ znamená, že chceme vědět, jaká je pravděpodobnost, že je daný člověk z Prahy, za podmínky, že se jmenuje Petr. Spočítali bychom to tedy tak, že bychom vzali pouze lidi, kteří se jmenují Petr (podmínka) a napočítali, jaká část těchto Petrů je z Prahy.

Pojďme se tedy podívat, jak bychom toto napočítali z poskytnutých dat:

$P(\text{KRAJ} = i \mid \text{JMÉNO} = j)$ a $P(\text{KRAJ} = i \mid \text{PŘÍJMENÍ} = k)$ se dají spočítat stejně, jako tomu bylo u nepodmíněných $P(\text{KRAJ} = i)$, jen bereme v úvahu řádek s daným jménem či příjmením.

	0	14	1	2	3	4	5	6	7	8	9	10	11	12	13	3000
PETR	1	31197	15819	29839	8226	14714	11896	32577	16632	13530	13971	33587	22956	12565	15342	272852
pravděpodobnost	3,66E-6	0,11434	0,05798	0,10936	0,03015	0,05393	0,0436	0,11939	0,06096	0,04959	0,0512	0,1231	0,08413	0,04605	0,05623	1

Chceme zjistit pravděpodobnost kraje, když víme, že se jedná o Petra – logicky tedy vezmeme všechny Petry a podíváme se, jak jsou rozděleni mezi jednotlivé kraje. Nejvíce Petrů je ve středočeském kraji.

$P(\text{JMÉNO} = j \mid \text{KRAJ} = i)$ a $P(\text{PŘÍJMENÍ} = k \mid \text{KRAJ} = i)$ tyto pravděpodobnosti se počítají z poskytnutých dat podobně jako výše úloha pro pravděpodobnosti výskytu jmen a příjmení nepodmíněné krajem. Zde ale musíme vzít v úvahu kraj a tedy napočítat pravděpodobnosti pro každý kraj a jméno. Na tabulce vpravo je vidět, jak bychom počítali pravděpodobnost toho, že je daný jedinec PETR, když víme, že je z kraje číslo 14. $P(\text{JMÉNO} = \text{PETR} \mid \text{KRAJ} = 14) = 0{,}028$.

	JMÉNO	14	pravděpodobnost
41161	PETKO NIKOLAJ	0	0
41162	PETMAT	0	0
41163	PETR	31197	0,0281336658
41164	PETR ADAM	0	0
41165	PETR ADOLF	1	0,00000090180677
41166	PETR ALBERT	0	0
41167	PETR ALEX	0	0
41168	PETR ALEXANDR	2	0,00000180361354
41169	PETR ALLEN	0	0
...	...	...	...
55437	ZOLTY	0	0
55438	ŽORA	0	0
55439	ŽORŽ	0	0
55440	ŽOFIA	0	0
55441	ŽUDI	0	0
55442	SOUČET	1108885	1

Podmíněné pravděpodobnosti – Bayesova věta a nezávislost

A teď, co bychom potřebovali pro řešení úlohy my pro naši úlohu? Potřebovali bychom vypočítat následující pravděpodobnost, to je to, co nás zajímá:

$$P(\text{KRAJ} = i \mid \text{PŘÍJMENÍ} = k, \text{JMÉNO} = j),$$

tedy chceme zjistit, z jakého je člověk kraje, když má jméno j a příjmení k (čárka ve vzorci značí průnik). Jinak řečeno, zajímá nás, z jakého kraje je člověk, když víme, jak se jmenuje.

V našem případě tedy konkrétně chceme pouze následující pravděpodobnost: $P(\text{KRAJ} = i \mid \text{PŘÍJMENÍ} = \text{HUDÍK}, \text{JMÉNO} = \text{PETR})$ pro všechny kraje. Protože podmíněná pravděpodobnost je také pravděpodobnost, je součet těchto pravděpodobností přes všechny kraje roven 1, přeci jenom člověk musí být z nějakého kraje, není-liž pravda.

Některé vzorečky pro práci s podmíněnými pravděpodobnostmi by se mohly hodit:

Přepis podmíněné pravděpodobnosti uvedené výše $P(A \cap B) = P(A \mid B)P(B)$ využijeme v Baysově větě, kterou lze vypočíst opačné podmínění: $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$.

Dále se bude hodit Bayesova věta s využitím rozpisu $P(B)$, nyní máme více (n) disjunktních jevů A_i, kde sjednocení A_i dává všechny možnosti. Pak:

$$P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}.$$

V našem případě je jev A_i roven tomu, že $\text{KRAJ} = i$, B je $\text{PŘÍJMENÍ} = \text{HUDÍK} \cap \text{JMÉNO} = \text{PETR}$.

Problémem je, že jakoukoli pravděpodobnost, kde se vyskytuje $\text{PŘÍJMENÍ} = \text{HUDÍK} \cap \text{JMÉNO} = \text{PETR}$ nejsme schopni z dat určit, protože nemáme data, kde by byly četnosti příjmení i jmen dohromady. Proto nám nezbývá nic jiného, než si pomoci nějakým předpokladem. Jediné, co připadá v úvahu je předpokládat nezávislost - v tomto případě se nám bude hodit podmíněná nezávislost - tedy vezmeme daný kraj a v něm předpokládáme, že to, jaké má člověk křestní jméno nezávisí na tom, jaké má příjmení (to sice asi úplně pravdou nebude, přeci jenom, lidé jistě vybírají dětem takové jméno, aby se trochu k danému příjmení hodilo a ne zcela náhodně). Nicméně těžko tyto pravděpodobnosti bez dat, která by o tomto něco říkala, určíme. Budeme se tedy muset spokojit s nezávislostí jmen. Předpokládáme tedy, že

$$P(\text{PŘÍJMENÍ} = k, \text{JMÉNO} = j \mid \text{KRAJ} = i) = P(\text{JMÉNO} = j \mid \text{KRAJ} = i)P(\text{PŘÍJMENÍ} = k \mid \text{KRAJ} = i).$$

Pokud dáme nyní vzorce dohromady, tak nám vyjde následující:

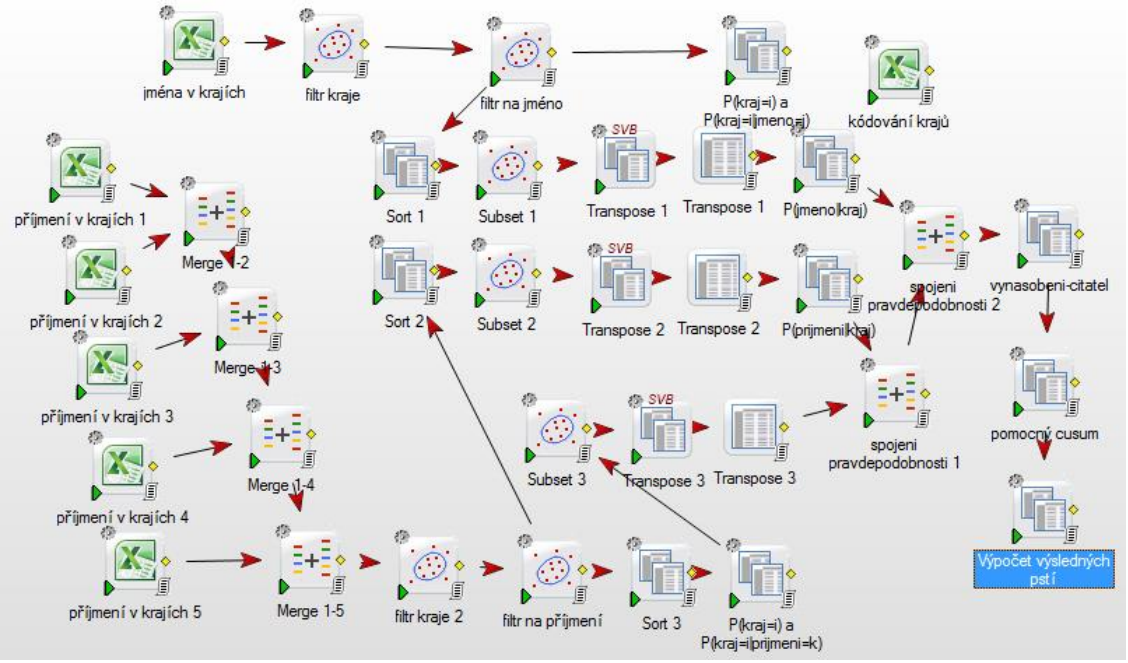
$$P(\text{KRAJ} = i \mid \text{PŘÍJMENÍ} = k, \text{JMÉNO} = j) = \frac{P(\text{JMÉNO=PETR} \mid \text{KRAJ=}i)P(\text{PŘÍJMENÍ=HUDÍK} \mid \text{KRAJ=}i)P(\text{KRAJ=}i)}{\sum_{j=1}^{n} P(\text{JMÉNO=PETR} \mid \text{KRAJ=}j)P(\text{PŘÍJMENÍ=HUDÍK} \mid \text{KRAJ=}j)P(\text{KRAJ=}j)}.$$

Pokud se podíváme na výrazy ve vzorci, zjistíme, že všechny lze napočítat z dat, která máme k dispozici. Máme tedy napůl vyhráno, nyní již víme jakým způsobem pravděpodobnost napočítat.

Implementace v softwaru

Nedalo nám to a nespokojili jsme se s pouhým zjištěním, že úloha má řešení. Chtěli jsme opravdu napočítat podmíněné pravděpodobnosti výskytu pana Petra Hudíka za pomoci reálných dat.

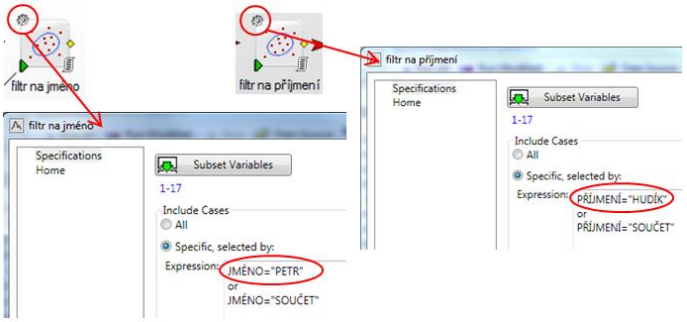
Jako prostředek k realizaci jsme zvolili workspace programu *STATISTICA* (přesněji řečeno modulu Data Miner). Workspace jsme zvolili z několika důvodů, jednak zdrojová data pro příjmení jsou v několika listech, což je pro manipulaci ne moc přínosné, ve workspace jsme je jednoduše spojili do jednoho souboru. Dále je workspace vhodný, protože se jedná vlastně o jakési graficky naprogramované makro, je tedy možné pouhým změněním jména a příjmení v uzlech *„filtr na jméno"* a *„filtr na příjmení"* a spuštěním vypočítat výsledek pro nové změněné jméno. Naprogramovaný workspace najdete **zde** a vypadá následovně:



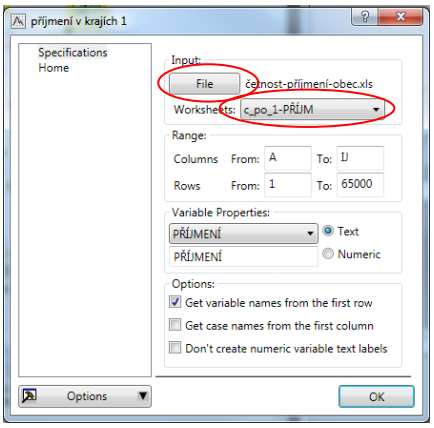
Znalci prostředí workspace ve verzi 10 (prostředí workspace ve verzi 10 jsme ukazovali například v **tomto článku**) jsou teď jistě velmi překvapeni tím, co právě viděli. Největší vývoj softwaru ve verzi 12 se totiž týká právě prostředí workspace, nové typy uzlů, možnost psát si do uzlů poznámky, načítání Excelovských souborů, atd. V jednom z dalších čísel se jistě k workspace vrátíme a popíšeme ho detailně v článku s novinkou.

Výsledky

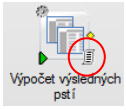
Tak a teď stačí pouze spočítat pravděpodobnosti z naší otázky. Do uzlů *„filtr na jméno"* a *„filtr na příjmení"* tedy vložíme zkoumané jméno a workspace spustíme.



Poznámka: Před prvním spuštěním workspace je potřeba znovu napojit Excelovské datové soubory, protože ty mají nyní jiné umístění. To se provede kliknutím na ozubené kolečko uzlů na načítání Excelu a výběrem umístění souboru pod tlačítkem *File* (u souborů s příjmením je nutné zvolit i list sešitu, postupně v 5 uzlech načteme všechny listy). Zbytek je už nastaven a není potřeba nic jiného dělat.



Výsledky jsou pak k nalezení v uzlu „Výpočet výsledných pstí", výsledné pravděpodobnosti jsou v posledním sloupci tohoto souboru.



Odpovědí na naši otázku je tedy sdělení, že pan Petr Hudík pochází nejpravděpodobněji z Libereckého kraje. Ostatní pravděpodobnosti najdete v tabulce vpravo.

	KRAJ	pstí pro PETR HUDÍK
0	Externí region	0,0000
14	Hlavní město Praha	0,0594
1	Jihočeský	0,0333
2	Jihomoravský	0,0339
3	Karlovarský	0,0373
4	Královéhradecký	0,0714
5	Liberecký	**0,1824**
6	Moravskoslezský	0,0210
7	Olomoucký	0,0208
8	Pardubický	0,0704
9	Plzeňský	0,0267
10	Středočeský	0,1205
11	Ústecký	**0,1785**
12	Vysočina	**0,1444**
13	Zlínský	0,0000

A když už máme hotovo „makro“ na výpočet takovýchto pravděpodobností, byla by škoda ho nevyužít k nějakým zajímavým výpočtům (nabízí se například zjišťovaní, odkud tak asi pocházejí kolegové z práce nebo slavné osobnosti). My jsme ale zkusili spustit výpočet pro několik slavných pohádkových jmén. Když už jsou Vánoce, za dveřmi, tak ať se dozvíme něco víc o našich pohádkových hrdinech.

	KRAJ	psti pro PETR MÁCHAL	psti pro MATĚJ KOTRBA	psti pro HLOUPÝ HONZA	psti pro KRÁL JAROSLAV
0	Externí region	0,0000	0,0000	0,0000	0,0000
14	Hlavní město Praha	**0,1823**	0,1458	0,0865	0,0960
1	Jihočeský	0,1632	**0,2410**	0,0000	0,0789
2	Jihomoravský	**0,2017**	0,0123	0,1653	0,0714
3	Karlovarský	0,0000	0,0080	0,0000	0,0218
4	Královéhradecký	0,0277	0,0173	**0,7049**	0,0497
5	Liberecký	0,1225	0,0356	0,0000	0,0426
6	Moravskoslezský	0,0181	0,0268	0,0000	0,0625
7	Olomoucký	0,0000	0,0083	0,0000	0,0721
8	Pardubický	0,0182	0,0104	0,0000	0,0573
9	Plzeňský	0,0172	0,0447	0,0433	0,0890
10	Středočeský	0,0732	0,1589	0,0000	**0,1879**
11	Ústecký	0,0480	0,0917	0,0000	**0,1065**
12	Vysočina	0,1187	0,1870	0,0000	0,0451
13	Zlínský	0,0089	0,0124	0,0000	0,0191

Vidíme tedy, že například Petra Máchala bychom hledali v Jihomoravském kraji. Pekař, který se stal císařem, pravděpodobně přišel péct chleba na dvůr Rudolfa II z Jihočeského kraje. Hloupý Honza vyráží k poznávání světa nejpravděpodobněji z Královéhradeckého kraje. A nakonec krásná, ale velmi pyšná princezna Krasomila, dcera vladaře Půlnočního království, se zpočátku odmítne provdat za šlechetného krále Miroslava nejpravděpodobněji ze Středočeského kraje – nejšpíš má tedy tomto kraji jiné favority, kdo ví, třeba nějakého úspěšného podnikatele, navíc se není čemu divit, když král Miroslav přijel na jednom koni, zatímco dnešní opulentní SUV jich mají i 300!

A to je konec naší pohádky o pravděpodobnostech a zajímavém využití workspace.

StatSoft CR s.r.o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz