StatSoft

Náhodný výběr

Jistě jste se již setkali s úlohou vybrat náhodně z nějaké množiny. Tento článek se bude zabývat právě tématem náhodného výběru.

Náhodným výběrem (vzorkem) označujeme podmnožinu prvků ze všech prvků (základního souboru), přičemž zahrnutí prvku do podmnožiny je náhodné a každý prvek má stejnou pravděpodobnost, že bude vybrán (zahrnut do výběru).

Soutěžní úloha

Jako motivaci začněme toto téma řešením úlohy z **minulého newsletteru**, která měla za úkol prověřit, zda poznáte, co je náhodný výběr a co už nikoli.
Úloha zněla takto: Představte si, že máte 7 předmětů a chcete z nich vybrat náhodně jen jeden. Jako generátor náhodného čísla máte k dispozici pouze klasickou šestistěnnou kostku. Otázkou a trochu i hádankou je, jakým způsobem použít Vaši kostku tak, aby mělo vybrání každého z předmětů stejnou pravděpodobnost.

Řešení

Řešení je několik, některé si ukážeme. Ukážeme si také příklady, které nejsou řešeními. Zatímco pro výběr ze 2, 3, 4, 6 nebo 8 předmětů je jednoduchý, vybrat ze sedmi pomocí šestistěnné kostky, je již trochu výzva.

První postup:

1. Pro každý předmět si hodím kostkou.
2. Vyberu ty předměty, kterým padlo největší číslo, s ostatními už nepočítám.
3. Vezmu tyto vybrané předměty a opakuji bod 1 a 2, dokud mi nezbyde jen jeden předmět, ten vyberu jako výsledný vybraný předmět.

Druhý postup:

1. Přidám si jeden předmět, mám jich tedy 8 a vybírám náhodně z osmi, což je jednoduché pomocí třech hodů kostkou, kdy vždy podle sudého/lichého čísla vyřadím polovinu předmětů.
2. Pokud bych náhodou vybral osmý předmět, tak opakuji bod 1.

Třetí postup:

1. Hodím postupně 2x kostkou. Zaznamenám výsledek. Pokud padne v prvním hodu 1 a ve druhém něco z 1 až 5 (1-1 až 1-5), pak vyberu první předmět. Pokud to bude 1-6 až 2-4, pak druhý, atd. Pokud by náhodou padly dvě šestky, tedy 6-6, tak celý pokus opakuji.

Čtvrtý postup:

Ten není příliš reálný, ale v rámci teorie by byl v pořádku, takže pro zajímavost uvedeme: nekonečněkrát hodím kostkou, při každém hodu se posouvám v kruhu o tolik míst, kolik padlo. Kruh je složen z předmětů 1-7, po 7 tedy opět následuje 1. Pravděpodobnost toho, kde zrovna jsem po tolika mnoho hodech, je rovnoměrná.

Špatné řešení:

1. Uspořádám předměty do kruhu, tedy 1-7 a po 7 je 1. Hodím si kostkou, tím vyberu, kde budu začínat, tedy na pozici 1 až 6.
2. Poté hodím znova a odpočítám počet, který padl od pozice, kterou jsem vybral (opět beru předměty cyklicky, po 7 je 1). Například pokud hodím 5 a potom 3, tak jsem vybral první předmět. Toto ale není správně, zkuste si tipnout, který předmět bude mít tímto principem lepší možnost vybrání….

…už máte svůj tip nebo výpočet?

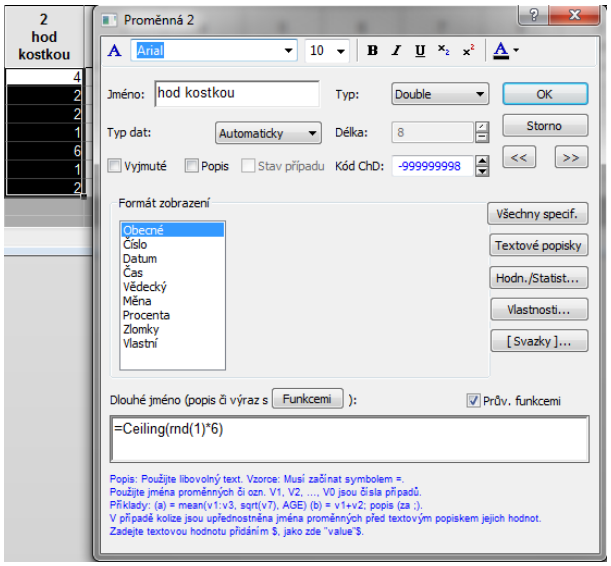
Předmět číslo 7 bude mít největší pravděpodobnost výběru, protože prvním hodem nemůžeme nijak diskvalifikovat to, že by v druhém nemohl být vybrán, pro ostatní předměty tomu tak ale není. Dá se dopočítat, že pravděpodobnost výběru sedmého předmětu je 1/6, zatímco u ostatních je pravděpodobnost jen 5/36.

Perlička: U všech správných postupů výše si na začátku nemůžete být jisti, kolikrát budete muset házet. Bohužel postup, který by zajišťoval, že rozhodneme o výběru přesným pevným počtem hodů, jsme nenašli. Spíš asi ani neexistuje, ale pokud byste na něj přeci jen přišli, určitě nám tento postup pošlete a my Vás za to nějak odměníme.

Jak nasimulovat kostku v softwaru?

Pokud byste neměli po ruce kostku a chtěli si výše popsaný experiment vyzkoušet, můžete použít samozřejmě software Statistica – jaký by to byl statistický software, kdyby neuměl generovat náhodná čísla. Pro tento účel nám bude stačit funkce *rnd(1)*, která generuje náhodné číslo z intervalu 0 až 1, přičemž rozdělení, ze kterého se generuje je rovnoměrné. My si ji jen trochu obalíme, aby generovala náhodně právě čísla 1, 2, 3, 4, 5 a 6. Funkce *rnd(1)*6* generuje čísla z intervalu 0 až 6. Funkce *Ceiling* zaokrouhlí výsledek na nejbližší celé číslo. Hodnoty mezi 0 a 1 jsou 1, mezi 1 a 2 jsou dva,…

Poznámka: Ve Statistica lze generovat i z jiných rozdělení, tyto funkce začínají na *rnd*.



Způsoby výběru vzorku

Nyní se budeme věnovat tomu, jak vybrat náhodný výběr – jak se to dělá v praxi například při výběrových šetřeních.
Druhy pravděpodobnostních výběrů:

1. „Losování“: znáhodňovací proces známý z loterijních her. Prvky je potřeba důkladně promíchat a poté stačí vybrat požadovaný počet. Ve statistice se jedná o **prostý (jednoduchý) náhodný výběr**, tedy z prvků vybereme

náhodně jeden, pak ze zbytku další,… Technik, jak tyto prvky vybrat, je více než jen zmíněné vybírání po jedné, důležité je, že každý prvek má stejnou možnost být zařazen.

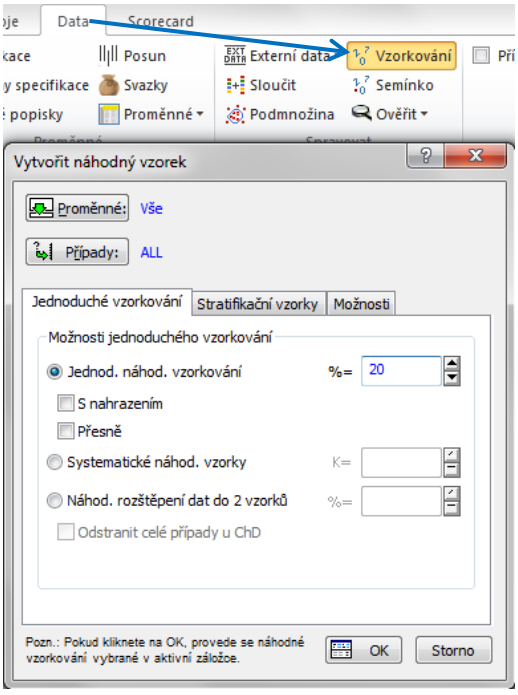
2. **Systematický výběr**: Vyžaduje, aby byly prvky seřazeny do posloupnosti, a poté se vybere každý K-tý prvek od náhodně vybraného prvku v rozmezí 1 až K. Pokud chceme vybrat 20 % prvků, pak K=5 a pokud je náhodně vybráno číslo 2, pak do výběru zahrneme 2., 7., 12.,… prvek posloupnosti. Každý prvek má pravděpodobnost zařazení stejnou a to 1/5. Podmínkou je, aby prvky byly seřazeny do posloupnosti nezávisle na zkoumané vlastnosti (Například můžeme vybírat pacienty z abecedně uspořádané kartotéky u praktického lékaře).
3. **Oblastní (stratifikovaný) výběr**: Základní populace je rozdělena do oblastí. Z každé z nich se pak vybírají prvky prostým náhodným výběrem. Výběrový soubor je spojením vybraných z každé oblasti. Cílem je vybrat skupiny, ve kterých by se sledované znaky moc nelišily a naopak, aby mezi skupinami byl rozdíl velký. Procento vybraných prvků z jednotlivých oblastí se může lišit. (Příkladem stratifikovaného náhodného výběru je například výběr osob z každého kraje ČR úměrně velikosti populace tohoto kraje).
4. **Skupinový výběr**: Nevybíráme jednotlivé prvky, ale přímo celé skupiny prvků (skupinami mohou být rodina, škola, firma, okres,…). Je vhodné, aby byly skupiny pokud možno stejně velké a prvky uvnitř různorodé (tedy právě opačné jako u oblastního výběru). Poté co vybereme skupinu, tak můžeme zahrnout do výběru všechny prvky nebo opět náhodně vybereme.
5. **Vícestupňový náhodný výběr**: Existuje nějaká hierarchie ve vlastnostech prvků základního souboru. Například Kraj->Okres->Obec->Volební obvod->Jedinci. Postupně vybíráme vlastnosti nejvýše v hierarchii, pak o jednu úroveň níže atd., až se dostaneme jednotlivým prvkům. Postupné výběry provádíme metodou prostého náhodného výběru. Tento výběr je vhodný z ekonomických důvodů a také v situacích, kdy neznáme předem celý soubor prvků.

Náhodné vzorkování v programu Statistica

Výše zmíněné metody jsou k dispozici přímo i v softwaru.

1. Jednoduchý (prostý) náhodný výběr

V záložce *Data – Vzorkování – Jednoduché vzorkování – Jednod. náhod. vzorkování*. Na obrázku vpravo vidíte nastavení, kdy bude ve výběru přibližně 20 procent dat z výchozího datového souboru (vybírá se ze všech případů) a výsledek bude mít stejné proměnné jako výchozí soubor (vybrány všechny proměnné). Možnost ***Přesně*** by znamenala přesně 20 procent. V této chvíli je to ale nastaveno tak, že každý prvek bude sám o sobě dotazován a pravděpodobnost výběru tohoto prvku je 0,2.

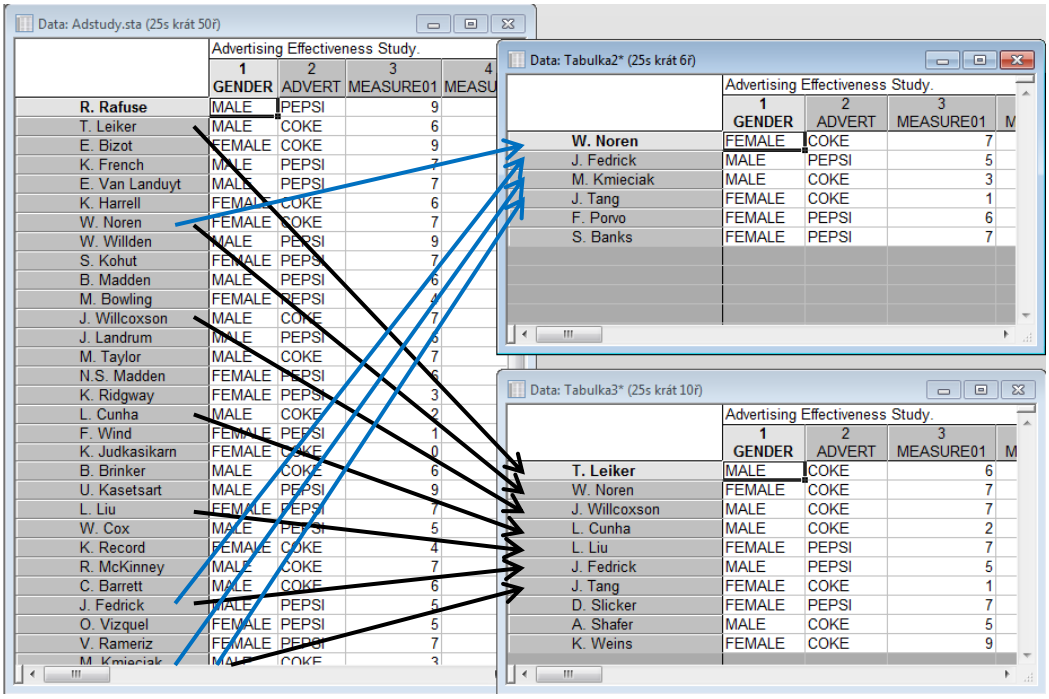


Poznámka: Je to stejné, jako bychom se rozhodovali podle velikosti nové proměnné *rnd(1)* a to tak, že pro velikost této proměnné větší než 0,8 bychom případ zařadili.

Pokud bychom chtěli místo procent přesné počty, pak se to dá nastavit v záložce *Možnosti*.

2. Systematický výběr

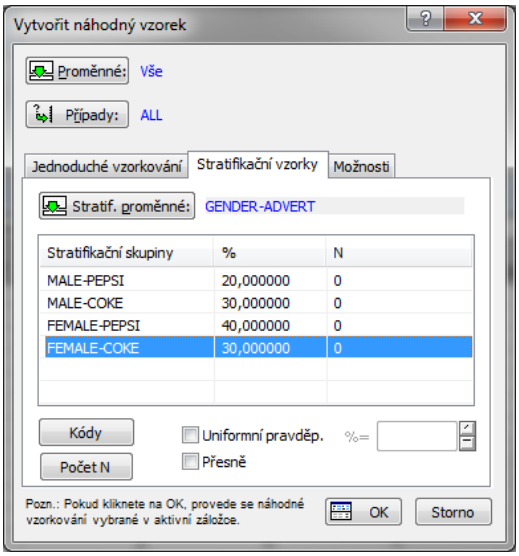
V záložce *Jednoduché vzorkování* prostě vybereme možnost *Systematické náhod. vzorky* . Je potřeba zvolit *K*, pokud chceme opět 20 procent dat, pak zvolíme K=5 (je to jedna pětina dat). Rozdíl výsledného výběru pro jednoduché vzorkování a systematický výběr vidíte na příkladu níže:



3. Stratifikovaný výběr

Najdeme jej v záložce *Stratifikační vzorky*. Na obrázku vpravo vidíme nastavení u souboru *Adstudy.sta* (v příkladech softwaru pod *Data - Otevřít příklady - Datasets*). Oblasti v tomto příkladu určují proměnné *GENDER* a *ADVERT*. Celkem tedy vybíráme přibližně 20 procent z těch, kteří jsou muži a byla na ně směřována reklama na Pepsi, 30 procent z těch, co jsou muži a zároveň na ně byla směrována reklama na Colu (Coke), atd.

Je možné míst procent přímo zadávat počty (do kolonky *N*). Abychom měli přesné počty, potřebovali bychom zaškrtnout možnost *Přesně*.



4.-5. Skupinový a vícestupňový výběr

Tyto výběry nejsou přímo implementovány v programu, ale opět bychom si mohli pomoci generováním náhodných čísel či aplikací jednoduchého náhodného výběru. Skupiny vybereme náhodně podobně jako vybíráme prvky, jen budeme nyní vybírat ze skupin místo přímo z prvků. Vícestupňový výběr je pak pracnější, protože potřebujeme vybírat ve více krocích. V prvním vybereme skupiny podle nejvyšší kategorizace, poté z vybraných podle druhé nejvyšší, atd.

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