3.1 Dataframe

Ve statistice se často setkáme s tím, že na náhodně vybraném objektu se zjišťuje více než jedna vlastnost. Např. u studentů prvních ročníku středních škol v Olomouckém kraji budeme zkoumat jejich průměrný prospěch a spokojenost s výběrem střední školy. Tato vícerozměrná data můžeme v R uložit do „dataframu“. Pro jednoduchost uvažujme, že jsme náhodně vybrali 5 studentů.

```
x <- c(1.8, 2.4, 2.0, 1.2, 3.5)
y <- c("F","T","T","T","F")
DF <- data.frame(prospech=x, spokojenost=y)
```

Uvědomte si, že vektory x a y mají stejnou délku. Sloupce mohou mít různé datové typy, ale datový typ v jednom sloupci musí být vždy zachován. U dataframu se používá podobný typ indexováni jako u vektoru nebo matice.

► **Příklad 3.1.** Zkuste si vypsat první až třetí řádek, druhý sloupec, prospěch u prvního a posledního studenta, spokojenost u druhého apod. Jaký je datový typ řádků a jaký sloupců?

Pro výběr jen některých řádků z dataframu můžeme použít funkci `subset(dataframe, condition)`.

▷ **Příklad 3.1.** Vyberte jen ty studenty, kteří mají prospěch horší než 2.0. Dále vyberte ty studenty, kteří jsou na škole spokojeni a mají prospěch lepší než 3.0. Řešení:

```
S1 <- subset(DF, prospech > 2.0)
S2 <- subset(DF, spokojenost=="T" & prospech < 3.0)
```

Velmi užitečnou funkcí pro výběr jen některých řádků je funkce `isin(x,y,ordered)`, kde `x`, `y` jsou vektory. Funkce vrací „TRUE“ pokud je každý prvek z vektoru `y` obsažen ve vektrou `x`. Pokud navíc nastavíme parametr `ordered` na „TRUE“ (defaultně je „FALSE“), vrátí funkce „TRUE“ pouze v případě, že jsou prvky vektoru `y` obsaženy ve vektoru `x` ve stejném pořadí. Pozor nemusí být nutně za sebou!

```
> a <- c(1,2,3,4,5)
> b <- c(5,3)
> isin(a,b)
[1] TRUE
> isin(a,b,ordered=TRUE)
[1] FALSE
> c <- c(3,5,3)
> isin(a,c)
[1] FALSE
> d <- c(2,4,5)
> isin(a,d)
[1] TRUE
```

Pokud je prvním argumentem funkce `isin(x,y,ordered)` místo vektoru dataframe, funkce se vyhodnocuje na každém řádku dataframu zvlášť.

Jména sloupců můžeme zjistit pomocí funce `names`. Pak můžeme vybrat sloupec na základě jeho jména pomocí symbolu $.

```
names(DF)
pr <- DF$prospech %$
```

Užitečné jsou rověž funkce `nrow` a `ncol`, které vrátí počet řádků nebo sloupců v dataframu. Vyzkoušejte je.

3.2 Prostor elementárních jevů a jeho podmnožiny

Množina všech možných výsledků náhodného pokusu se nazývá prostor elementárních jevů. Pro náhodné pokusy, které jsou v teorii pravděpodobnosti často používané, byly vytvořeny funkce, které vrací množinu možných výsledků pro n opakování jako dataframe, např. `tosscoin(n)`. Následující příklady vyzkoušejte, změňte parametry a pozorujte výsledky. Nejprve je ale potřeba includovat package „prob" a „combinat".

```
library(prob)
library(combinat)
T <- tosscoin(4)
R1 <- rolldie(2)
R2 <- rolldie(1,nsides=4)
cards()
cards(jokers=FALSE)
```

Poznámka. Po načtení package „prob" se objevilo následující:

```
Attaching package: 'prob'
The following object(s) are masked from 'package:base':
      intersect, setdiff, union
```

V package „base", kterou jsme si nainstalovali současně s R, jsou již definovány funce pro průnik, rozdíl a sjednocení množin (vektorů) (intersect, setdiff, union). V package „prob" jsou tyto funce rozšířeny a umožňují provádět tyto množinové operace i s data framy.

Mezi často používaný náhodný pokus patří výběr objektů (balónků/čísel) z urny. V R existuje funkce `urnsamples(x,size,replace,ordered)`, kde `x` určuje obsah urny (z čeho vybíráme), `size` velikost výběru, parametr `replace` indikuje zda vytažený objekt do urny vracíme či nikoliv (zda-li jde o výběr s opakováním či bez opakování), parametr `ordered` značí, zda-li nám záleží na pořadí objektů ve výběru nebo ne. Představte si, že hodíme dvakrát čtyřstěnnou kostkou, množinu možných výsledků můžeme (kromě použití funkce `rolldie(2,nsides=4)`) vypsat i pomocí výběru z urny.

```
x <- 1:4
U <- urnsamples(x,size=2, replace = TRUE, ordered = TRUE)
```

Pokud nás zajímá jenom počet možných výsledků a nechceme generovat celý prostor elementárních jevů, můžeme použít funkci `nsamp(x,k,replace,ordered)`, která vrátí počet řádků vygenerovaných funkcí `urnsamples`, aniž by v paměti skutečně prostor elementárnich jevů generovala.

Libovolná podmnožina prostoru elementární jevů se nazývá náhodný jev. V R můžeme pro výběr podmnožiny z dataframu použít již dříve zmiňovanou funkci `subset`.

▷ **Příklad 3.2.** Kolika způsoby můžeme sestavit vlajku se třemi různými vodorovnými pruhy, máme-li k dispozici šest barev(M, Č, B, Ž, Z, F)? Protože vlajka s pruhy „M,Č,B" je jiná než vlajka s pruhy v pořadí „Č,B,M", jde o výběr, ve kterém záleží na pořadí (parametr ordered). A protože vlajka musí mít tři různé vodorovné pruhy, jde o výběr bez opakování, tj. nemůžeme vybrat jednu barvu vícekrát.

```
x <- c("M","Č","B","Ž","Z","F")
U <- urnsamples(x, size=3, replace=FALSE, ordered=TRUE)
answer <- nrow(U)
#nebo
nsamp(n=6, k=3, replace=FALSE, ordered=TRUE)
```

Kolik lze sestavit vlajek s červeným pruhem? Červený pruh může být na prvním, druhém nebo třetím místě.

```
nrow(subset(U, (X1=="Č")|(X2=="Č")|(X3=="Č")))
```

Kolik lze sestavit vlajek, které nebudou mít prostřední pruh zelený?

```
nrow(subset(U, !(X2=="Z")))
```

Kolik lze sestavit různých vlajek na kterých bude bílý pruh nad žlutým?

```
S <- subset(U, isin(U,c("B","Ž"),ordered=TRUE))
nrow(S)
```

► **Příklad 3.2.** Na oslavě narozenin je 10 lidí. Při slavnostním přípitku si chce každý ťuknout s každým. Kolikrát cinknou skleničky? [45]

► **Příklad 3.3.** V cukrárně mají tři různé druhy větrníků. Chceme koupit čtyři větrníky. Kolik máme různých možností? [15]

► **Příklad 3.4.** Nechť $\Omega = \{1, 2, 3\}$ je prostor elementárních jevů. Přidejte do následujících tříd množin jen ty nezbyteně nutné množiny, aby se třída stala polem.

1. $\{\{2\}, \{3\}\}$
2. $\{\{1\}\}$
3. $\{\emptyset\}$

► **Příklad 3.5.** Házíme čtyřstěnnou falešnou kostkou. Víme, že $P(\{1\}) = 1/3$, $P(\{2\}) = 1/6$ a $P(\{3\}) = 2P(\{4\})$. Určete pravděpodobnost, že padne sudé číslo. [1/3]

► **Příklad 3.6.** Víte, že $P(A) = 1/2$ a $P(B) = 2/3$. Dokažte, že $1/6 \leq P(A \cap B) \leq 1/2$. Za jakých podmínek nastane $P(A \cap B) = 1/2$ a kdy $P(A \cap B) = 1/6$?

► **Příklad 3.7.** Hazardní hráči vsází na to, zda-li padne alespoň jedna jednička při opakovaném hodu kostkou. Na co je lepší vsadit (padne/nepadne), pokud házíme kostkou třikrát po sobě? Změní se situace nějak, pokud se kostkou bude házet čtyřikrát?

► **Příklad 3.8.** Hodíme třemi kostkami. Kolik je možných výsledků? Kolik výledků splňuje podmínku, že součet ok na všech kostkách je ostře větší než 14? Kolikrát padnou minimálně dvě šestky? V kolika případech padne součet ostře větší než 14 a současně padne maximálně jedna šestka?[216;20;16;10]

Reference

[1] Capinski M., Zastawniak T. J.: Probability Through Problems
Springer 2001, ISBN 978-0-387-95063-1.

[2] Devore J. L.: Probability and Statistics for Engineering and the Sciences
Duxbury Press, 7. vydání 2008, ISBN 978-0-495-55744-9.

[3] Kerns G. J.: Elementary Probability on Finite Sample Spaces, 2009, reference manual package prob, http://CRAN.R-project.org/package=prob

[4] Kerns G. J: Introduction to Probability and Statistics Using R, First Edition
http://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf