SEZNAMY (`list`)

- specifické objekty sdružující různé jiné objekty různých typů;
- jednotlivé položky seznamu mohou mít vlastní název (není povinný) - v rámci jednotlivých položek jsou uloženy dané hodnoty: v každé položce pouze jeden typ
  - ```
    Doctor <- list(name = c("William
    Hartnell","Patrick Troughton","Jon
    Pertwee","Tom Baker","Peter Davison","Colin
    Baker","Sylvester McCoy","Paul McGann",
    "John Hurt","Christopher Eccleston"),
    year=c("55","45","50","40","29","40","44","3
    6","73","41"),
    episodes=c("134","119","128","172","69","31"
    ,"42","1","2","13")); Doctor
    ```
    ```

Přidání položek do seznamu

- `Doctor <- list(name = c(Doctor$name,"David Tennat","Matt Smith","Peter Capaldi"), year = c(Doctor$year,"34","27","55"), episodes = c(Doctor$episodes,"47","44","26")); Doctor`

```
> Doctor <- list(name = c(Doctor$name,"David Tennat","Matt Smith","Peter Capaldi"), year = c(Doctor$year,"34","27","55
"),
+                episodes = c(Doctor$episodes,"47","44","26")); Doctor
$name
 [1] "William Hartnell"      "Patrick Troughton"     "Jon Pertwee"           "Tom Baker"             "Peter Davison"
       "Colin Baker"
 [7] "Sylvester McCoy"       "Paul McGann"           "John Hurt"             "Christopher Eccleston" "David Tennat"
       "Matt Smith"
[13] "Peter Capaldi"

$year
 [1] "55" "45" "50" "40" "29" "40" "44" "36" "73" "41" "34" "27" "55"

$episodes
 [1] "134" "119" "128" "172" "69"  "31"  "42"  "1"   "2"   "13"  "47"  "44"  "26"
```

DATOVÁ TABULKA (data.frame)

- seznam uložený jako dvojdimenzionální struktura – dvojrozměrná tabulka dat, kde jednotlivé sloupce (pole) mohou obsahovat různé typy dat
  - `vek <- c(27,35,30,47,42);vek`
  - `sex <- c("female","male","male","female","female"); sex`
  - `strana <-c("sin","sin","dx","dx","dx"); strana`
  - `ID <- c("SC_001", "SC_007", "SC_013", "SC_014", "SC_020"); ID`
  - `VZOREK <- data.frame(vek, strana, sex, row.names = ID); VZOREK`
  - `vzorek <- data.frame(vek, strana, sex, ID); vzorek`

Výběr v data.frame

- `VZOREK$vek`
- `VZOREK$strana`
- `VZOREK$sex`
- `VZOREK[,2:3]`
- `VZOREK[2:5,]`
- `VZOREK[sex=="male",]`
- `VZOREK[sex=="male" & strana=="sin",]`
- `VZOREK[c("SC_001","SC_013"),]`

- *Vyzkoušejte si jednotlivé výběry a navrhněte vlastní.*

Základní funkce data.frame

- `order()` – seřazení tabulky
  - `VZOREK[order(VZOREK$vek),]`
  - `VZOREK[order(VZOREK$sex, VZOREK$vek),]`
- `merge()` – slučování tabulek
  - `vyzkum <- c("A22","A22","A02","A02","A02")`
  - `ID <- c("SC_001", "SC_007", "SC_013", "SC_014", "SC_020")`
  - `rok_vyzkumu <- data.frame(ID, vyzkum); rok_vyzkumu`
  - `merge(vzorek, rok_vyzkumu, by = "ID")`

- **Otázka**: co se stane, když je ve spojovaných tabulkách více sloupců se shodným názvem?

- `rok_vyzkumu <- data.frame(ID, strana, vyzkum); rok_vyzkumu`
- `merge(vzorek, rok_vyzkumu, by = "ID")`

```
> merge(vzorek, rok_vyzkumu, by = "ID")
     ID vek strana.x    sex strana.y vyzkum
1 SC_001  27      sin female      sin    A22
2 SC_007  35      sin   male      sin    A22
3 SC_013  30       dx   male       dx    A02
4 SC_014  47       dx female       dx    A02
5 SC_020  42       dx female       dx    A02
```

- `subset()` – podvýběr datové tabulky; specifikuje řádky, které splňují danou podmínku
  - `subset(VZOREK,VZOREK$vek>=35)[,c("strana","sex")]`

- Přidání nové proměnné do `data.frame`
  - `VZOREK$vyska <- c(155,170,187,165,195)`
  - `VZOREK$mereni <- NA`

```
       vek strana    sex vyska mereni
SC_001  27    sin female   155     NA
SC_007  35    sin   male   170     NA
SC_013  30     dx   male   187     NA
SC_014  47     dx female   165     NA
SC_020  42     dx female   195     NA
```

Chybějící a neexistující hodnoty

- Je rozdíl mezi: `NA`, `0`, `""`, `NULL` a `NaN`
- `NA` – chybějící hodnota (ne všechny funkce dokáží s `NA` hodnotami pracovat – viz odstranění `NA` hodnot)
- `NaN` – 'Not a Number'; nečíselná hodnota
  - `Inf` a `-Inf` : kladné či záporné nekonečno, ale `NaN` číslo není př.:
    - `0/0`
    - `1/0 + 1/0`

- `NULL` – neexistující hodnota, hodnota tam není – neplést s `NA` kdy hodnota chybí! `NULL` reprezentuje nulový objekt
    - `100+NULL`

ULOŽENÍ OBJEKTŮ

- funkce `sink()`
  - `sink("vysledky.txt")`
  - *`to-co-chceme-uložit`*
  - `sink()`

- `write.table(VZOREK, "tabulka.csv", sep=";")`
- další funkce: `write.csv()`, `write.csv2()`, `write.matrix()`, `write()`

- *Note*: možnost Compile Report (Ctrl+Shift+K) v Rstudiu.

DOMÁCÍ ÚKOL 4

- A) Dle tabulky 1 vytvořte `data.frame` s názvem `Soubor`.

  Snažte se údaje „slepě" nevypisovat, ale použijte funkce z kapitoly Vektory, výjimku tvoří poslední dva sloupce).

- B) Seřaďte objekt `Soubor` podle proměnné `strana` a vytvořte dva samostatné objekty:

  `prava` obsahující pouze hodnoty pravé strany (`dx`)

  `leva` obsahující hodnoty pouze levé strany (`sin`)

- C) Objekty `prava`, `leva` spojte dohromady podle proměnné `ID` a proměnné `sex` do objektu `Soubor_S`.

ID	sex	strana	H1	F1
1	f	dx	35.3	49.7
1	f	sin	35.4	49.7
2	f	dx	32.3	47.5
2	f	sin	32.5	47.9
3	f	dx	37.1	53.7
3	f	sin	37.3	53.2
4	m	dx	29.8	44.4
4	m	sin	31.5	44.4
5	m	dx	28.4	43.2
5	m	sin	27.5	44.1
6	m	dx	30.7	46.3
6	m	sin	30.9	45.8

Tab. 1 Soubor měření

Antropologická data

- Tabulky většinou obsahují:
  - Hlavičku (názvy sloupců)
  - Pozor na ZDVOJENÉ proměnné
  - Pozor na sloučené buňky apod.
    - Doporučuje se co nejjednodušší formát
    - Nedoporučuje s v rámci hodnot používat `";"`
    - V případě textových proměnných se doporučuje co nejjednodušší název (bez mezer, pomlček, háčků a čárek)
  - Vždy data po načtení kontrolujeme!

- **R používá desetinnou tečku ne desetinnou čárku!**
  - Note: pozor při převádění dat např. z `.xlsx` do `.txt`
    - `Soubor -> Ulož jako -> Text (oddělený tabulátory)(*.txt)`

NAČTENÍ OBJEKTŮ

- `read.table()`
- `read.csv()`
- `read.delim()`
- `read.delim2()`

- Nejčastěji používané parametry:
  - `sep()` oddělovač sloupců, př.: `"\t"`, `"."`, `";"`
  - `header()` nabývá logických hodnot `TRUE` nebo `FALSE`, říká, zda mají sloupce tabulky názvy
  - `dec` oddělovač desetinných míst
  - `row.names` názvy řádků
  - `col.names` názvy sloupců
  - `skip` kolik řádků při načítání tabulky přeskočit
- Funkce načítají tabulky jako data.frame

Kontrola po načtení

- Po načtení `data.frame` doporučujeme:
  - `head()` zobrazí prvních pár položek načteného objektu
  - `dim()` vypíše dimenze načteného objektu
  - `dimnames()` vypisuje nebo nastavuje názvy jednotlivých dimenzí objektu
  - `names()` má-li objekt položky typu tag (`list` nebo `data.frame`), pak vypíše názvy položek
  - `str()` zobrazí strukturu objektu
  - `summary()` výpočet základních charakteristik
  - `levels()` množina hodnot, kterých může kategorická proměnná nabývat

- Stáhněte a načtěte data:
  - `dataP.txt`
  - `dataP.csv`
    - soubor `data.xls` obsahuje list 2 s vysvětlením daných proměnných
    - Odkaz: *učební materiály v ISu*
    - Zkontrolujte načtené objekty, případně opravte chyby.

Variable	Description	Variable	Description
Age	Age	Inner_End_W	At Inner End - Widths
Sex	Sex (male, female)	Inner_A_W	At inner Angle - Widths
No	Number	Least_W_W	Least Width - Widths
No_sub	Number of subject	Distance1_W	Distance 1 - Widths
Side	Side (sin, dx)	Conoid_W	At Conoid - Widths
Length_cl	Length of clavicle	Distance2_W	Distance 2 - Widths
Shoulder_w	Shoulder Width	Outer_End_W	At Outer End - Widths
Proportion	Proportion	Inner_End_D	At Inner End - Depths
Inner_C	Inner Angle - Curvature	Middle_D	At Middle - Depths
Outer_C	Outer Angle - Curvature	Conoid_D	At Conoid - Depths
Total_C	Total - Curvature	Outer_D	Least Outer Dept - Depths
Inner_S	Inner Segments	Circumference	Circumference at Middle
Middle_S	Middle Segments	Index_End	Index of Inner End
Outer_S	Outer Segments		

- `DATA <- read.csv("dataP.csv", sep=";", header = TRUE)`
- `attach(DATA)`
- `head(DATA)`
- `dim(DATA)`
- `dimnames(DATA)`
- `names(DATA)`
- `str(DATA)`

- v rámci proměnných `sex`, `side` a `age` jsou neočekávané hodnoty (viz výstup funkce `str(DATA)`)
  - `> levels(DATA$Sex)`
  `[1] "female" "male" "muz" "zena"`
  - `> levels(DATA$Side)`
  `[1] "dx" "dy" "si" "sin" "sni"`
- opravíme pomocí `levels`, např.
  - `levels(DATA$Sex)[levels(DATA$Sex)=="zena"] <- "female"`
  - `levels(DATA$Age)[levels(DATA$Age)=="old"] <- NA`

Note: po úpravě doporučuji tabulku uložit, aby uživatel nemusel úpravu dat opakovat pokaždé, když bude chtít se souborem dataP.csv pracovat.

- `summary(DATA)` základní statistiky
- `summary(DATA$Age)` základní statistiky sloupce Age
- `summary(DATA[,7])` základní statistiky 7 sloupce

- Jiné základní statistiky pomocí balíčku *pastecs*
  - `install.packages("pastecs")`
  - `library(pastecs)`
  - `stat.desc(DATA[Sex=="female","Length_cl"])`
  - `stat.desc(DATA[Sex=="male","Length_cl"])`

- Totéž, ale bez hodnot NA:
  - `stat.desc(na.omit(DATA[Sex=="female","Length_cl"]))`
  - `stat.desc(na.omit(DATA[Sex=="male","Length_cl"]))`

- Využití balíčku *psych* – základní popisné statistiky dle skupin:
  - `library(psych)`
  - `describeBy(DATA[,"Length_cl"], group=DATA[,"Sex"])`

```
 Descriptive statistics by group
group: female
   vars   n   mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 121 137.86 8.42    137   137.1 7.41 121 162    41 0.84     0.53 0.77
------------------------------------------------------------------------------
group: male
   vars   n   mean   sd median trimmed   mad min max range skew kurtosis   se
X1    1 127 152.32 9.28    153  152.15 10.38 126 176    50 0.04    -0.25 0.82
```

  - `describeBy(DATA[,"Length_cl"], group=DATA[,c("Sex","Side")])`

- Vymazání proměnné:
  - `DATA$Inner_End_W <- NULL`
- Přidání proměnné:
  - `DATA$new <- NA`
    - *Note: přidání i vymazání proměnné lze přes odkazování typu* `[,]` *ne jen* `$`*!*

DOMÁCÍ ÚKOL 5

- A) Z ISu si stáhněte data: `du_5.txt` a načtěte je do RStudia (jméno proměnné zvolte libovolně).
- B) Zkontrolujte načtená data, případné chyby opravte.
  - *Při opravách chyb vycházejte z předpokladu, že hodnota s vyšším zastoupením je ta správná.*
- C) Zjistěte dimenze načteného objektu a vypočítejte základní statistiky jednotlivých proměnných
- D) Vymažte prázdné sloupce (`19` až `31`).
- E) Vytvořte dva objekty `DX` a `SIN`:
  - `DX` bude obsahovat pouze měření z pravé strany (`Side` bude `dx`)
  - `SIN` bude obsahovat pouze měření z levé strany (`Side` bude `sin`)

  A spojte je dohromady do objektu `DXSIN` (*note*: funkce `merge`).
- F) Vypočítejte základní statistiky pro muže a ženy k rozměrům `F1.x`, `F1.y`.