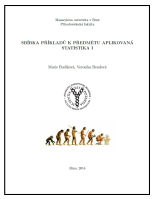
# Úvod do předmětu Aplikovaná statistika I

- **Základní informace o předmětu**
  - Kód předmětu: MAS10c
  - Název předmětu: Aplikovaná statistika I
  - Vyučující předmětu: Mgr. Veronika Bendová
  - Kontakt na vyučující: 375612@mail.muni.cz; bendova.veroonika@gmail.com
  - Konzultační hodiny: Dohodou
- **Podmínky k získání zápočtu**
  - **Aktivní** účast na cvičení s maximálním počtem **tří** absencí (omluvené i neomluvené)
  - Vypracování krátkého domácího úkolu zadaného na prvním cvičení
  - Vypracování sady příkladů zadaných za **domácí úkol**
  - *Poznámka: Zápočtová písemka se **nepíše**.*
- **Domácí úkol**
  - **Obsah:** Šest až osm příkladů (podle náročnosti)
  - **Období zadání úkolu:** Druhá polovina semestru
  - **Doba na vypracování domácího úkolu:** 14 dní
  - **Forma řešení úkolu:** RSkript obsahující kompletní řešení příkladů, komentáře popisující postupy řešení a interpretace získaných výsledků
  - **Hodnocení úkolu:** U každého příkladu se hodnotí správnost řešení, design grafů, interpretace výsledků, komentáře použitých postupů a přehlednost kódu.
  - **Kontrola úkolu:** Vyučující má 14 dní na zkontrolování a opravení domácího úkolu.
  - **Splnění úkolu:** Získání alespoň 75 % bodů z maximálního možného počtu bodů. V případě, že student potřebný počet bodů nezíská, bude mu úkol jedenkrát navrácen k přepracování.
  - **Zvláštní pravidlo:** Student může domácí úkol vypracovávat společně se svými spolužáky a konzultovat s nimi řešení úkolu.
  - **Zvláštní omezení:** Student **nesmí** od spolužáků **kopírovat** řešení, **kopírovat** interpretace výsledků a **kopírovat**  kód. V případě porušení dochází k penalizaci domácího úkolu, jejíž míra je stanovena podle závažnosti porušení: od stržení extra bodů z celkového hodnocení úkolu až po neuznání řešení úkolu a neudělení zápočtu.
- **Další doporučení pro studenty**:
  - Účast na přednáškách
  - Domácí příprava na cvičení z hodiny na hodinu - průběžně si opakujte; principy se budou stále opakovat, průběžné chápání látky usnadní přípravu na zkoušku
  - **Interagujte s vyučující**, ptejte se, e-mailujte, konzultujte, pomáhejte si a hlavně tomu rozumějte :).

- **Doporučená a zajímavá literatura:**
  - **Aplikovaná statistika I: Sbírka řešených příkladů**



Soubor řešených příkladů pokrývající svým obsahem látku probíranou v kursu Aplikovaná statistika I. Sbírka slouží k samostatnému domácímu procvičování probírané látky. Obsahuje zadání příkladů, výpočetní řešení, řešení pomocí softwaru R a interpretace výsledků.
*Citace:* Bendová, V.: Aplikovaná statistika I: Sbírka řešených příkladů; studijní materiál

  - **Průvodce základními statistickými metodami**


Doporučený text doplňující výklad z přednášek. Svým obsahem pokrývá látku probíranou v kursu AS I. Text je sepsán čtivou formou vhodnou pro studenty, kteří se se statistickými pojmy a metodami teprve seznamují. Obsahuje ilustrační příklady vhodné k procvičování výpočtů probíraných metod. Příklady jsou řešeny pomocí softwaru STATISTICA.
*Citace:* Budíková M., Králová M., Maroš B.: *Průvodce základními statistickými metodami*, Praha, Grada, 2010, ISBN 978-80-210-7752-2, 272 s.

  - **Základní statistické metody**


Doporučený text doplňující výklad z přednášek. Obsahově je kniha velmi podobná Průvodci základními statistickými metodami, pokrývá tedy taktéž látku probíranou v kursu Aplikovaná statistika I. Je přehlednější, ale stručnější a neobsahuje tolik ilustračních příkladů. Příklady navíc obsahují pouze výpočetní řešení bez využití jakéhokoli softwaru.
*Citace:* Budíková M., Lerch T., Mikoláš Š.: *Základní statistické metody*, Brno, Masarykova Univerzita, 2009, ISBN 978-80-210-3886-8, 170 s.

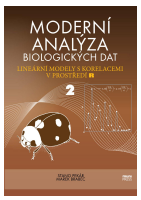
  - **Aplikovaná štatistická inferencia I**


Pokročilá literatura doporučená pro studenty, kteří již absolvovali kurs Aplikovaná statistika I a chtějí si prohloubit znalosti nabyté v kursu. Kniha obsahuje mnoho ilustračních příkladů zaměřených na analýzu antropologických dat s řešením pomocí softwaru R. Text je však více odborný a vyžaduje jistou orientaci ve statistických pojmech.
*Citace:* Katina S., Králík M., Hupková A.: *Aplikovaná štatistická inferencia I.*, Brno: MUNI Press, 2015, ISBN 978-80-210-7752-2, 306 s.

  - **Moderní analýza biologických dat I**


Literatura doporučená pro studenty, kteří absolvovali kurs AS I a chtějí si prohloubit znalosti nabyté v kursu. Zaměřuje se na oblast regresní analýzy, která je v kursu probírána pouze okrajově, ovšem má rozsáhlé využití v praxi. Metody jsou ilustrovány na příkladech analýzy biologických dat. Kniha je psána vstřícnou formou a obsahuje řešení pomocí R.
*Citace:* Pekár S., Brabec M.: *Moderní analýza biologických dat 1*, Praha: Scientia, 2009, ISBN 978-80-86960-44-9, 225 s.

  - **Moderní analýza biologických dat II**

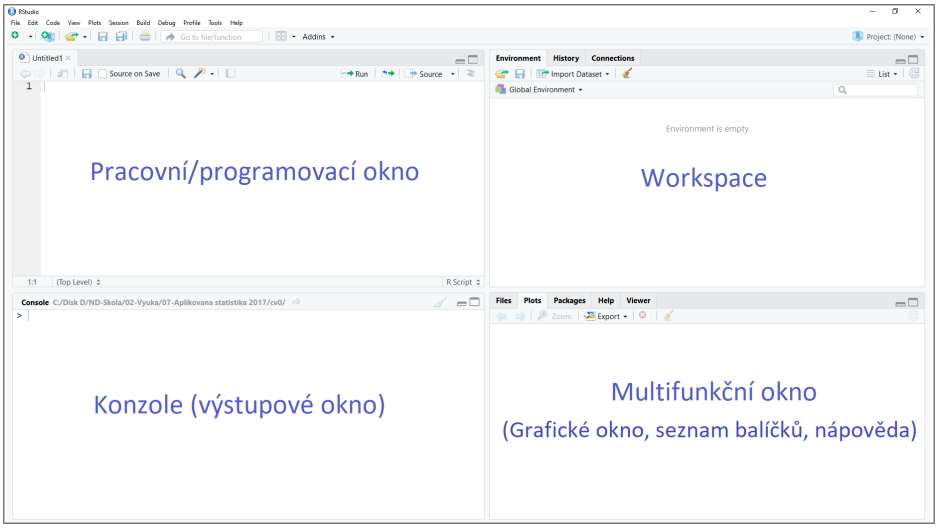



Druhý díl výše uvedené publikace zaměřující se na téma smíšených regresních modelů, které nejsou v kursu probírány, ale mají rozsáhlé využití v praxi. Vhodná jako základní literatura k analýze dat s opakovanými měřeními. Obsahuje opět příklady z oblasti biologie řešení pomocí softwaru R.
*Citace: Citace:* Pekár S., Brabec M.: *Moderní analýza biologických dat 2*, Brno: MUNI Press, 2012, ISBN 978-80-210-5812-5, 256 s.

# 1 Základy práce se statistickým softwarem R

- **Úvod a motivace**
  - Kvalitní analýza dat je v dnešní době často nedílnou součástí kvalitního antropologického výzkumu.
  - Antropolog my měl mít alespoň základní povědomí o principech vybraných statistických metod.
- **Náplň kursu Aplikovaná statistika I:**
  - Představení vybraných statistických metod
  - Aplikace statistických metod na konkrétní datové soubory z oboru antropologie
  - Výuka použití statistického softwaru R při datové analýze
- **Výstup kursu Aplikovaná statistika I:**
  - Orientace v základní statistické terminologii
  - Schopnost samostatně přemýšlet nad vlastními daty
  - Schopnost zvolit vhodné metody k seznámení s daty
  - Schopnost provést samostatně analýzu dat pomocí softwaru R
  - Správná interpretace získaných výsledků
- **Software R a proč právě on?**
  - Existuje mnoho nástrojů umožňujících provedení statistické analýzy dat
    - Tabulkový software MS Excel, statistický software STATISTICA
    - Skriptovací programovací jazyky: Matlab, R
  - **Výhody softwaru R**
    - Živý programovací jazyk, který se neustále vyvíjí.
    - Rozsáhlé množství funkcí a příkazů umožňující provedení libovolné datové analýzy.
    - Volně stažitelná plnohodnotná licence softwaru
    - Volně stažitelná příslušenství k softwaru: balíčky funkcí, nápovědy, RStudio
    - Krásná grafika umožňující vlastní nastavení vzhledu grafů a animací
    - Výborná nápověda + různá diskusní fóra, která pomáhají najít řešení problémů
    - Možnost vytvářet tzv. živé dokumenty: HTML stránky (R HTML), pdf dokumenty (R Sweave), Word dokumenty (R Markdown) a interaktivní aplikace (R Shiny)
  - **Nevýhody softwaru R**
    - Programovací jazyk → je potřebné naučit se jeho syntaxi
    - Balíčky funkcí vytváří uživatelé → je potřebné studovat nápovědu k používaným funkcím pro získání jistoty, že funkce fungují dle našich předpokladů
    - Nápověda dostupná pouze v angličtině, většina diskusí též v angličtině
- **Citace a odkazy**
  - **Citace softwaru R:** R CORE TEAM. R: *A Language and Environment for Statistical Computing. The R Project for Statistical Computing*, 2017 [cited 2018 Sep 21]. URL: https://www.R-project.org/
  - **Html stránka pro software R:** *https://www.r-project.org/*
  - **Html stránka pro RStudio:** *https://www.rstudio.com/*

- **RStudio: Uživatelsky přístupné prostředí pro práci se softwarem**
  - Po otevření RStudia
    - Zahájíme nový projekt: **File** → **New file** → **R Skript**
    - Uložíme projekt: **File** → **Save** → **'Nazev projektu'** → **Save**
    - Nastavíme absolutní cestu do složky s projektem: **Session** → **Set Working Directory** → **To Source File Location**
  - Nyní máme RStudio rozdělené na čtyři okna
    1. **Pracovní/programovací okno**
       - Prostor pro vytváření našeho kódu
       - Kód = posloupnost příkazů a funkcí → vlastnosti funkce specifikujeme volbou jejich argumentů
       - Provedení příkazu nebo funkce: **Ctrl + Enter**
    2. **Workspace**
       - **Environment** - Seznam proměnných uložených v paměti
       - **History** - Seznam naposledy provedených příkazů
       - Vymazání uložených proměnných: **Session** → **Clear Workspace** → **Yes**
    3. **Konzole (výstupové okno)**
       - Zobrazuje provedené příkazy a jejich výstupy
       - Vyčištění konzole: **Ctrl+L**
    4. **Multifunkční okno** (obsahuje více záložek)
       - **Plots** - Grafické okno
         - Prostor pro zobrazení vykreslených grafů
       - **Packages** - Seznam nainstalovaných balíčků + instalace nových balíčků
         - Instalace balíčku **nortest**: **Packages** → **Install** → **nortest** → **Install**
       - **Help** - Nápověda
         - Každá funkce má svou nápovědu obsahující:
           i. **Description**- stručný popis funkce
           ii. **Usage** - tvar příkazu se všemi povinně volitelnými argumenty
           iii. **Arguments** - Předhled všech argumentů (povině volitelných i volitelných)
           iv. **Values** - popis výstupů funkce
           v. **Details** - Bližší detaily, například o vzorcích a metodách, na nichž je funkce založena
           vi. **See Also** - tipy na příbuzné funkce, které by nás dále mohly zajímat
           vii. **Examples** - Ilustrační příklady správného použití funkce
  - Nastavení vzhledu RStudia: **Tools** → **Options** → **Appearance**



- **Přehled základních matematických objektů**
  - **Proměnná** = označení objektu (číslo, vektor, matice, tabulka, funkce)
    - Číslo: a $\leftarrow$ 3
    - Vektor: **a** $\leftarrow$ (3, 6, 9)
    - Matice: **A** $\leftarrow \begin{pmatrix} 2 & 3 & 7 \\ 8 & 4 & 5 \end{pmatrix}$
    - Tabulka: **Tab** $\leftarrow$

| | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| $r_1$ | 2 | 3 | 7 |
| $r_2$ | 8 | 4 | 5 |

    - Funkce: **funkce()**
      - Objekt, do kterého vložíme vstup (IN) a získáme výstup (OUT)
      - **sum**(vektor) $\rightarrow$ číslo
      - **matrix**(vektor, 2, 2) $\rightarrow$ matice
      - **mean**(matice) $\rightarrow$ číslo

- **Úvod do syntaxe programovacího jazyka** 
  - Jazyk  je tzv. **case sensitive** $\rightarrow$ názvy sum, Sum, SUM, sUm, sUM, ... znační různé objekty
  - **Vytvoření proměnné**
    - Číslo:

```
a <- 3
```

    - Vektor:

```
aa <- c(1.2, 5.3, 6.4)
```

    - Matice:

```
A <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 2, ncol = 3, byrow = T)
```

    - Datová tabulka:

```
Tab <- data.frame(A, row.names = c('r1', 'r2'))
names(Tab) <- c('s1', 's2', 's3')
```

  - **Základní operace**
    - +, -, *, /

```
3 + 2 - 6 * 9 / (8 + 9 - 5)
```

    - Operace s čísly

```
a <- 25
b <- 5
a / b
```

    - Operace s vektory

```
x <- c(1, 2, 3)
y <- c(3, 2, 1)
x - y
x + y

z <- c(0, 1, 2, 3)
x + y + z
# !Pozor, vektor z je delsi nez vektory x a y.
#  R sice napise varovnou hlasku, ale vypocet i tak provede!
```

    - Operace s maticemi

```r
B <- matrix(c(1, 1, 1, 1, 1, 1), nrow = 2, ncol = 3)
A - B

C <- matrix(c(1, 1, 1, 1, 1, 1), nrow = 3, ncol = 2)
A - C
# !Matice A ma rozmer 2x3, matice C ma rozmer 3x2.
#  Nyni jiz R vypocet neprovede, pouze zahlasi chybu.
```

– **Pokročilé operace**

* Délka vektoru a dimenze matice

```r
length(z)
dim(A)
```

* Minimální a maximální hodnota vektoru a matice

```r
min(z)
max(z)

min(A)
max(A)
```

* Součet hodnot a aritmetický průměr vektoru a matice

```r
sum(z)
sum(A)

mean(z)
mean(A)
```

* Mocniny a odmocniny

```r
(odmocnina <- sqrt(2))
odmocnina ^ 2
```

* Zakrouhlování

```r
round(odmocnina, digits = 3) # klasicke zaokrouhleni na tri desetinna mista
floor(odmocnina)             # zaokrouhleni na nejblizsi nizsi cele cislo
ceiling(odmocnina)           # zaokrouhleni na nejblizsi vyssi cele cislo
```

– **Vytváření posloupností**

* Posloupnosti čísel se vzdáleností 1

```r
(x <- 1:10)
(y <- 50:55)
```

* Posloupnosti čísel s libovolnou ale ekvidistantní vzdáleností

```r
# Posloupnost cisel s predem zadanou delkou (R dopocita vzdalenost mezi sousedicimi cisly)
(pst <- seq(from = 0, to = 1, length = 12))

# Posloupnost cisel s predem zadanou vzdalenosti mezi sousedicimi cisly (R dopocita delku)
(pst2 <- seq(from = 0, to = 1, by = 0.09))
```

* Posloupnosti opakujících se čísel

```r
vaha     <- c(58, 61, 57, 59, 60, 54, 64, 71, 66, 70)
divky    <- rep(1, 6)
chlapci  <- rep(2, 4)
(pohlavi <- c(divky, chlapci))
```

* Přídání nového řádku (resp. sloupce) ke stávajícímu vektoru

```
(data  <- matrix(c(vaha, pohlavi), nrow = 2, ncol = 10, byrow = T))
(data.r <- rbind(vaha, pohlavi)) # pridani radku k vektoru (vznikne matice dimenze 2x10)
(data.c <- cbind(vaha, pohlavi)) # pridani sloupce k vektoru (vznikce matice dimenze 10x2)

diabetes <- rep(0, 10)
(data.r2 <- rbind(data.r, diabetes)) # pridani radku k matici 2x10 (vznikce matice 3x10)
(data.r3 <- cbind(data.r2, c(62, 2, 1))) # pridani sloupce k matici 3x10 (vznikce matice 3x11)
```

* Přídání nového řádku (resp. sloupce) ke stávající matici

```
data  <- matrix(c(vaha, pohlavi), nrow = 2, ncol = 10, byrow = T)
data.r <- rbind(vaha, pohlavi) # pridani radku k vektoru (vznikne matice dimenze 2x10)
data.c <- cbind(vaha, pohlavi) # pridani sloupce k vektoru (vznikce matice dimenze 10x2)

diabetes <- rep(0, 10)
data.r2 <- rbind(data.r, diabetes) # pridani radku k matici 2x10 (vznikce matice 3x10)
data.r3 <- cbind(data.r2, c(62, 2, 1)) # pridani sloupce k matici 3x10 (vznikce matice 3x11)
```

– **Podmnožiny vektorů a matic**

* Výběr konkrétních hodnot z vektoru

```
vyska <- c(133, 132, 145, 126, 127)
vyska[c(2, 3, 4)]
vyska[2:4]
```

* Výběr konkrétních řádků z matice

```
data.r3[1, ] # vyber prvniho radku
data.r3[2, ] # vyber druheho radku
```

* Výběr konkrétních sloupců z matice

```
data.r3[ ,4] # vyber ctvrteho sloupce
data.r3[ ,8] # vyber osmeho sloupce
data.r3[ , c(3, 5, 8)] # vyber tretiho, pateho a osmeho sloupce najednou

data.r3[1:2, 5:7] # vyber cisel z prvniho a druheho radku a pateho, sesteho a sedmeho sloupce
```

– **Práce s datovým souborem**

* Zjištění absolutní cesty k aktuální složce a výpis všech souborů z této složky

```
getwd() # absolutni cesta k aktualni slozce
dir()   # vypis souboru
```

* Načtení datového souboru a vypsání prvních tří řádků z datového souboru (tabulky)

```
# Nacteni datoveho souboru: read.delim()
# sep: separator sloupcu, napr. tabulator '\t'; strednik ';' nebo carka ','
# dec: oddelovac desetinnych mist pouzity v souboru, napr. carka ',' nebo tecka '.'
data <- read.delim('Zaznam teploty.txt', sep = '\t', dec = '.')
head(data, n = 3) # vypsani prvnich tri radku
```

```
  Hodina Teplota
1      0    38.0
2      2    37.7
3      4    37.3
```

* Výběr konkrétních řádků z tabulky

```
data[4:6, ]
```

* Výběr konkrétních sloupců z tabulky (více možností)

```r
# Vyber prvniho sloupce s nazvem 'Hodina'
data[ ,1]
data[ ,'Hodina']
data$Hodina

# Vyber druheho sloupce s nazvem 'Teplota'
data[ ,2]
data[ , 'Teplota']
data$Teplota

hodiny  <- data$Hodina
teploty <- data$Teplota
```

– **Logické operátory**

* Rovnost ==, větší >, menší <, větší nebo rovno >=, menší nebo rovno <=

```r
(teploty == 37.7)*1
sum((teploty==37.7)*1) # kolikrat byla teplota rovna 37.7

(teploty <= 37.7)*1
(teploty >= 37.7)*1
(teploty < 37.7)*1
(teploty > 37.7)*1
```

* Výběr řádků s konkrétní vlastností

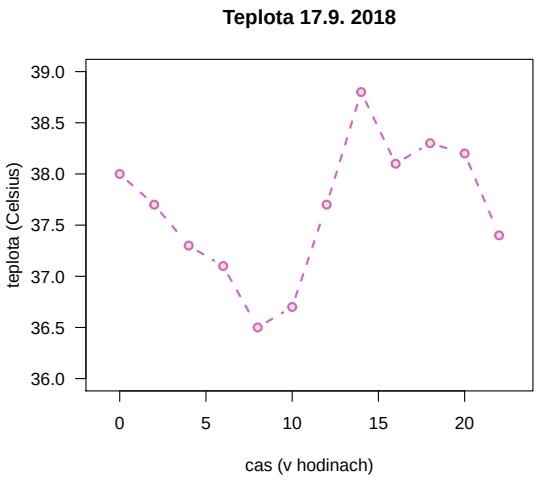
```r
data[data$Teplota == 37.7, ]
data[data$Teplota == 37.7, 'Hodina']
```

– **Tvorba základních grafů příkazem plot(x, y)**

1. **Povinně volitelné argumenty funkce plot(x, y)**
   * **x** - proměnná, která se vykreslí na ose x
   * **y** - proměnná, které se vykreslí na ose y
2. **Volitelné argumenty funkce plot(x, y)**
   * **main** - nadpis grafu (v publikacích nepoužíváme, nahrazujeme popiskem pod grafem)
   * **xlab** - popisek osy x v grafu
   * **ylab** - popisek osy y v grafu
   * **col** - základní barva objektů (body, čáry) v grafu
   * **type** - typ grafu
     · **type = 'p'**: bodový graf
     · **type = 'l'**: spojnicový graf
     · **type = 'b'**: kombinace bodového a spojnicového grafu
     · **type = 'n'**: prázdný graf
   * **pch** - typ bodů
     · **pch = 1**: kruh bez výplně
     · **pch = 19**: výplň kruhu
     · **pch = 21**: kruh s výplní (možnost volby různých barev obvodu kruhu a obsahu kruhu)
     · Další typy bodů viz **nápověda funkce points() → Details → 'pch values'**
   * **bg** - barva vnitřku bodu v případě, že **pch = 21**
   * **cex** - velikost bodů (defaultně **cex = 1**)
   * **lwd** - šířka čáry (defaultně **lwd = 1**)
   * **lty** - typ čáry
     · **lty = 1**: klasický styl

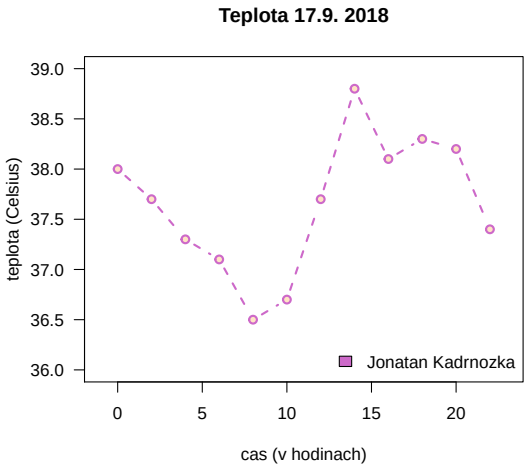
        - **lty = 2**: čárkovaný styl
        - **lty = 3**: tečkovaný styl
        - **lty = 4**: čerchovaný styl
    - **xlim = c(a, b)** - rozsah osy x od a do b
    - **ylim = c(a, b)** - rozsah osy y od a do b
    - **las = 1** - popisky měřítka osy y vodorovně s osou x

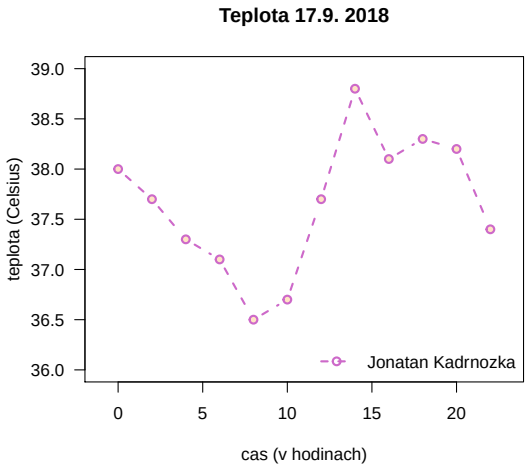
```r
plot(hodiny, teploty,
     main = 'Teplota 17.9. 2018', xlab = 'cas (v hodinach)', ylab = 'teplota (Celsius)',
     col  = 'orchid3', type = 'b',
     pch  = 21, lwd  = 2, lty  = 2, bg = 'bisque',
     xlim = c(-1, 23), ylim = c(36, 39), las = 1)
```




– **Doplnění legendy do grafu příkazem legend()**
    - Umístění legendy
        - **'topright'**: vpravo nahoře
        - **'topleft'**: vlevo nahoře
        - **'bottomright**: vpravo dole
        - **'bottomleft**: vlevo dole
        - **'top'**: nahoře uprostřed
        - **'right'**: vpravo uprostřed
        - ...
    - **fill** - barva výplně legendy
    - **legend** - popisek legendy
    - **bty = 'n'** - odstranění rámečku okolo legendy
    - **pch**, **lwd**, **lty** - analogické argumentům funkce **plot(x, y)**
    - **col** - barva bodu, resp. čáry

```r
legend('bottomright', fill = c('orchid3'), legend = c('Jonatan Kadrnozka'), bty = 'n')
# resp.
legend('bottomright', col = c('orchid3'), pch = c(21), lwd = c(2), lty = c(2),
       legend = c('Jonatan Kadrnozka'), bty = 'n')
```




- **Export grafů do png souboru**
  - **Ruční export: Multifunkční okno → Plots → Export → Save as Image → Maintain aspect ratio: odškrtnout → Save**
  - **Export posloupností příkazů:**

```
png('Nazev grafu.png')
plot(1:5, 1:5)
dev.off() # prikaz dev.off() projizdime, dokud se v konzoli neobjevi hlaska null device 1
```

**Tip na domácí procvičení: Vytvoření analogického grafu příkazy plot(), lines() a points()**

- **plot(x, y, type= 'n', ...)** - příprava prázdného grafu
- **lines(x, y, ...)** - vykreslení čar
- **points(x, y, ...)** - vykreslení bodů

```
plot(hodiny, teploty, main = 'Teplota 17.9. 2018', xlab = 'cas (v hodinach)',
     ylab = 'teplota (Celsius)', type = 'n', xlim = c(-1, 23), ylim = c(36, 39), las = 1)
lines(hodiny, teploty, lwd = 2, lty = 2, col = 'orchid3')
points(hodiny, teploty, pch = 21, col = 'orchid4', bg = 'bisque', lwd = 2)
legend('bottomright', fill = c('bisque'), legend = c('Jonatan Kadrnozka'), bty = 'n')
```