# Časové řady – dekompozice, analýza trendu a sezónnosti

—

Supplementum ke cvičení 4ST201 Statistika

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova, Praha

ČVUT FBMI

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

# Obsah

1. Opakování
2. Složky časové řady, dekompozice
3. Trendová složka časové řady
4. Sezónní složka časové řady
5. Literatura

# Příklad

- V přiloženém souboru `_12_cviceni_.xlsx` jsou v záložce `pocty_pracovniku` hodnoty počty pracovníků daného podniku vždy k prvnímu dni každého měsíce v rámci roku. Spočítejme
  - (i) diference prvního řádu,
  - (ii) meziměsíční tempa růstu,
  - (iii) průměrné teplo růstu
  - (iv) a chronologický průměr počtu pracovníků[1].

[1] Jako váhy sčítanců v průměru použijme vždy délky měsíců ve dnech.

# Časová řada

- časová řada je posloupnost hodnot uspořádaných v čase od minulosti do současnosti
  - obvykle (ne však vždy) se předpokládá konstantní časový krok mezi každými dvěma sousedními hodnotami časové řady

# Dělení časových řad

- dle délky časového kroku mezi sousedními hodnotami řady
  - **krátkodobé řady** – je-li časový krok kratší než jeden rok[2]
    - např. čtvrtletní, měsíční, týdenní, denní, minutová časová řada, atd.
  - **dlouhodobé řady** – je-li časový krok alespoň jeden rok
- dle vztahu hodnoty k časovému kroku
  - **okamžikové řady** – hodnota v každém časovém okamžiku představuje nepředpočítanou velikost dané veličiny
    - např. cena akcií na burze, počet obyvatel na Zemi, směnný kurz €–U.S.$ na FOREXu, atd.
    - nemá smysl nad hodnotami takové řady provádět součty
  - **intervalové řady** – hodnota v každém časovém okamžiku představuje předpočítanou velikost dané veličiny ve smyslu úhrnu (součtu) či průměru
    - např. HDP dle čtvrtletí, úmrtnost na karcinom plic ročně za posledních deset let, počet prodaných automobilů měsíčně v daném regionu atd.

[2] jde o historické dělení, v současnosti v podstatě překonané kvůli *high-frequency* časovým řadám např. u algoritmického obchodování s peněžními deriváty, kde je časový krok řádově v nanosekundách

# Složky časové řady

- časovou řadu $\{y\}_{t=1}^{n} = \{y_1, y_2, y_3, \ldots, y_n\}$ lze dekomponovat na několik složek
  - (i) trend $T_t$, tedy setrvalou tendenci vývoje
  - (ii) sezónnost $S_t$, tedy pravidelně se opakující odchylku s periodou kratší než jeden rok
  - (iii) cyklus $C_t$, tedy pravidelně se opakující odchylku s periodou alespoň jednoho roku
  - (iv) náhodnou složku $\varepsilon_t$, kterou nelze vysvětlit jinak
- dekompozice může být aditivní

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

- nebo multiplikativní

$$y_t = T_t \cdot S_t \cdot C_t \cdot \varepsilon_t$$

# Parametrické modelování trendu $T_t$ časové řady

- trend modelujeme jako parametrickou křivku, tedy např. přímku, parabolu, exponenciálu apod.
- parametry zvolené parametrické křivky je třeba odhadnout tak, aby křivka „dobře“ odpovídala[3] hodnotám časové řady
  - odhad parametrů křivky může být založen na minimalizaci součtu čtverců odchylek mezi skutečnými hodnotami časové řady a vždy příslušnými hodnotami časové řady vypočtenými na základě křivky trendu

[3] v určitém slova smyslu

# Volba parametrické křivky trendu $T_t$ časové řady

- založena na věcné znalosti dané problematiky
- může odrážet charakteru rozložení hodnot časové řady v bodovém diagramu $[t, y_t]_{t=i}^{n}$
- jsou-li diference prvního řádu konstantní a diference druhého řádu nulové, volíme *přímku*
- jsou-li diference druhého řádu konstantní a diference třetího řádu nulové, volíme *parabolu*
- jsou-li mezihodnotová tempa růstu konstantní, volíme *exponenciálu*

# Příklad

- V přiloženém souboru `_12_cviceni_.xlsx` jsou v záložce `kolorektalni_karcinom` je incidence nových případů kolorektálního karcinomu v Praze pro roky 1998 až 2005. Proložme časovou řadu přímkovým trendem. Nyní zanedbáváme jakoukoliv možnost přítomnosti sezónní složky.

# Příklad

- V přiloženém souboru `_12_cviceni_.xlsx` jsou v záložce `obrat_startupu` jsou roční úhrny obratu jistého startupu v milionech korun. Proložme časovou řadu exponenciálním trendem. Nyní zanedbáváme jakoukoliv možnost přítomnosti sezónní složky.

# Neparametrické modelování trendu $T_t$ časové řady

- též zvané *vyhlazování*
- existují různé sofistikované metody (exponenciální vyhlazování i jiné)
- běžně se ale používá pouze *klouzavý průměr*
- trend tedy nemodelujeme jako parametrickou křivku pomocí regrese, ale pouze počítáme průměrnou hodnotu vždy pomocí několika sousedních hodnot časové řady
- prostý klouzavý průměr pro vyhlazení sezónnosti liché délky

$$\bar{y}_t = \frac{y_{t-p} + y_{t-p+1} + \cdots + y_t + \cdots + y_{t+p-1} + y_{t+p}}{m}$$

kde $m = 2p + 1$ je okénko klouzavého průměru

- centrovaný klouzavý průměr pro vyhlazení sezónnosti sudé délky

$$\bar{y}_t = \frac{y_{t-p} + 2y_{t-p+1} + \cdots + 2y_t + \cdots + 2y_{t+p-1} + y_{t+p}}{2m}$$

kde $m = 2p$ je okénko klouzavého průměru

# Příklad

- V přiloženém souboru `_12_cviceni_.xlsx` jsou v záložce `HDP_1994_az_2000` je čtvrtletní časová řada HDP České republiky v období od 1. čtvrtletí 1994 do 4. čtvrtletí 2000. Vyrovnejme tuto časovou řadu klouzavými průměry délky 5.

# Sezónní složka $S_t$ časové řady

- sezónnost je pravidelně se opakující odchylka s periodou kratší než jeden rok
- modeluje se různě, včetně regresního přístupu
- uvažujeme-li u časové řady $\{y\}_{t=1}^{n} = \{y_1, y_2, y_3, \ldots, y_n\}$ aditivní dekompozici a přítomnost trendu $T_t$ a sezónnosti $S_t$, můžeme psát

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

kde $\varepsilon_t$ je náhodná složka

- po přepisu do podoby regresní křivky dostáváme pro sezónnost délky $k$ časových kroků

$$y_t = \beta_0 + \beta_1 t + \gamma_1 D_1 + \gamma_2 D_2 + \ldots + \gamma_{k-1} D_{k-1} + \varepsilon_t$$

kde $t$ je (lineární) trend a $\gamma_j$ a $D_j$ je lineární koeficient a dummy proměnná $j$-tého časového kroku sezónnosti, kde $j \in \{1, 2, \ldots, k-1\}$; hodnota sezónnosti $k$-ho časového kroku je $\beta_0 + \beta_1 t$

# Sezónní složka $S_t$ časové řady pro čtvrtletí ($k = 4$)

- uvažujeme-li sezónnost pro čtvrtletí, pak $k = 4$
- regresní křivka má pro sezónnost délky $k = 4$ časových kroků tvar

$$y_t = \beta_0 + \beta_1 t + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \varepsilon_t$$

kde $t$ je (lineární) trend a $\gamma_j$ a $D_j$ je lineární koeficient a dummy proměnná $j$-tého čtvrtletí, kde $j \in \{1, 2, 3\}$; hodnota sezónnosti $4$-ho čtvrtletí je $\beta_0 + \beta_1 t$

# Korekce sezónní složky $S_t$ časové řady pro čtvrtletí $(k = 4)$

- odhadneme-li regresní křivku pro čtvrtletí následovně

$$y_t = \beta_0 + \beta_1 t + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \varepsilon_t$$

pak je nutné provést přepočet pomocí $\hat{s} = \frac{\hat{\gamma}_1^* + \hat{\gamma}_2^* + \hat{\gamma}_3^* + 0}{4}$, abychom získali skutečné hodnoty odhadů pro jednotlivá čtvrtletí $S_{j+4i}$, kde $i \in \mathbb{N}$

$$\begin{aligned} S_{1+4i} &= \hat{\gamma}_1^* - \hat{s} \\ S_{2+4i} &= \hat{\gamma}_2^* - \hat{s} \\ S_{3+4i} &= \hat{\gamma}_3^* - \hat{s} \\ S_{4+4i} &= -\hat{s} \end{aligned}$$

- trendová složka má tvar $\hat{T}_t = \left(\hat{\beta}_0^* + \hat{s}\right) + \hat{\beta}_1^* t$ a sezónní složka má tvar $\hat{S}_t = \widehat{S_{1+4i}} Q_1 + \widehat{S_{2+4i}} Q_2 + \widehat{S_{3+4i}} Q_3 + \widehat{S_{4+4i}} Q_4$, kde $Q_1, \ldots, Q_4$ identifikují čtvrtletí

# Příklad

- V přiloženém souboru `_12_cviceni_.xlsx` jsou v záložce `HDP_sezonnost` je čtvrtletní časová řada HDP České republiky v období od 1. čtvrtletí 2002 do 4. čtvrtletí 2006. Modelujme trendovou a sezónní složku pomocí regresního přístupu a odhadněme vývoj této časové řady na rok 2007.

# Literatura


Hindls, Richard, Stanislava Hronová, Jan Seger a Jakub Fischer. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. ISBN: 978-80-86946-43-6.


Marek, Luboš. *Statistika v příkladech*. Praha: Professional Publishing, 2015. ISBN: 978-80-7431-153-6.

## Děkuji za pozornost!

lubomir.stepanek@vse.cz
lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

https://github.com/LStepanek/4ST201_Statistika