Úvod, základné pojmy, testovanie bieleho šumu: CVIČENIA

Beáta Stehlíková

2-PMS-10 Časové rady

Fakulta matematiky, fyziky a informatiky, UK v Bratislave

Cvičenie 1: Ljung-Boxov test - opakovanie

```
##
##  Box-Ljung test
##
## data:  x
## X-squared = 4.4469, df = 3, p-value = 0.2171
```

- Aká hypotéza sa tu testuje?
- Vysvetlite, čo presne znamenajú hodnoty uvedené s popisom `X-squared`, `df` a `p-value`.
- Zamietame nulovú hypotézu alebo nie?
- Čo nám tento test hovorí o dátach? Môže ísť o biely šum alebo je to nepravdepodobné? Aké ďalšie testy by ste spravili, aby ste vedeli lepšie odpovedať na túto otázku?

- Ako sa počíta testovacia štatistika?
- Vypočítajte jej hodnotu, ak viete, že dáta obsahujú 50 pozorovaní a prvé hodnoty autokorelačnej funkcie sú nasledovné:

```
acf(x, lag.max = 5, plot = FALSE)
```

```
##
## Autocorrelations of series 'x', by lag
##
##     0     1     2     3     4     5
## 1.000 0.026 -0.171 0.226 -0.106 0.048
```

Až na zaokrúhľovacie chyby nám má výjsť výsledok z výstupu testu (ten sa počíta z presných, nie zaokrúhlených hodnôt ACF).

Cvičenie 2: Ljung-Boxov test pre prietoky Nílu

- Uvažujme dáta o prietoku Nílu z prednášky po postavení priehrad, napr. od roku 1905. Sú v premennej `Nile`, ktorá je typu `ts` (*time series*):

```
class(Nile)
```

```
## [1] "ts"
```

- Kratší časový rad z nej vytvoríme pomocou funkcie `window` s parametrami `start` a/alebo `end`:

```
x <- window(Nile, start = 1905)
```

- Zobrazte ACF

- LB testom budeme testovať hypotézu, že hladiny Nílu sú nekorelované, pre viac lagov a výsledné p-hodnoty zobrazíme graficky (ako na konci prednáškových slajdov).
- Opakovanie:
  - Ljung-Boxovym testom testujte, že je prvých 5 autokorelácií nulových
  - Ako z výstupu Ljung-Boxovho testu dostaneme p-hodnotu?
- Budeme testovať nulovosť prvých k autokorelácií postupne pre $k \in \{1, 2, \ldots, 15\}$.
  - **Postup 1:** Napíšte for-cyklus, ktorým do pripraveného vektora vložíte postupne jednotlivé p-hodnoty
  - **Postup 2:** Vo všeobecnosti sú for-cykly v R-ku pomalé a odporúča sa vyhýbať sa im. Často je užitočná trieda funkcií `apply`.

- Pre nás bude teraz užitočná funkcia `sapply` (s je zo slova *simplify*, výstup sa zjednoduší, v tomto prípade do tvaru vektora)
- Príklad použitia:

```
# sucet 1^2 + 2^2 + ... + k^2
k <- 1:10
sapply(k, FUN = function(n) sum((1:n)^2))
```

```
##  [1]   1   5  14  30  55  91 140 204 285 385
```

```
# sucet 1^p + 2^p + ... + k^p
k <- 1:10
sapply(k, FUN = function(n, p) sum((1:n)^p), p = 2)
```

```
##  [1]   1   5  14  30  55  91 140 204 285 385
```

- Vypočítajte pomocou funkcie `sapply` všetky p-hodnoty Ljung-Boxovho testu. Použité dáta môžu byť:
  - pevne zvolené (analógia prvého vzorového príkladu z predchádzajúceho slajdu)
  - alebo môžu byť parametrom funkcie `FUN` (analógia druhého vzorového príkladu)

Cvičenie 3: Ljung-Boxov test pre úrokové miery

- Budeme analyzovať dáta dlhodobých úrokových mierach z Európskej centrálnej banky, pričom použijeme mesačné dáta z obdobia 20 rokov (2001-2020). Využijeme ich aj neskôr, pri ďalších témach.
  https://sdw.ecb.europa.eu/browse.do?node=bbn4864
- Na stránke predmetu je csv-súbor stiahnutý z webu ECB (pri aktuálnom stiahnutí bude obsahovať aj novšie dáta - budú potrebné napríklad v domácej úlohe).

	A	B	C	
1	Data Source in SDW: https://sdw.ecb.europa.eu/browse.do?node=bbn4864			
2		IRS.M.AT.L.L40.CI.0000.EUR.N.Z	IRS.M.BE.L.L40.CI.0000.EUR.N.Z	IRS.M.BG.L.L40.C
3		Austria, Euro	Belgium, Euro	Bulgaria, Bulgariar
4	Collection:	Average of observations through period (A)	Average of observations through period (A)	Average of observ
5	Period\Unit:	[Percent]	[Percent]	[Percent]
6	2022Jul	1.70	1.80	1.85
7	2022Jun	2.07	2.13	1.77
8	2022May	1.54	1.58	1.62
9	2022Apr	1.29	1.30	1.62
10	2022Mar	0.72	0.79	1.09
11	2022Feb	0.54	0.59	0.61
12	2022Jan	0.18	0.26	0.57

- Načítame dáta:

```
# DOPLNTE
data <- read.csv(...,       # cesta k suboru
                 skip =   , # vynechanie riadkov,
                            # chceme iba data
                 header = FALSE, # lebo sme vynechali riadk
                 row.names = 1)  # prvy sltpec nie su data
                                 # ale pouzijeme ich
                                 # ako nazvy riadkov
```

```
head(data)
```

```
##              V2   V3   V4   V5   V6   V7   V8   V9  V10  V1
## 2022Jul 1.70 1.80 1.85 3.41 4.40 1.08 1.59 2.66 2.31 1.7
## 2022Jun 2.07 2.13 1.77 3.42 5.12 1.45 1.83 2.48 2.63 2.0
## 2022May 1.54 1.58 1.62 2.69 4.61 0.95 1.30 1.91 2.04 1.4
```

- Získajte zo súboru s dátami názvy (štáty a meny):

```
# DOPLNTE
nazvy <- read.csv(...)
```

```
head(nazvy)
```

```
##       [,1]
## [1,] "Austria, Euro"
## [2,] "Belgium, Euro"
## [3,] "Bulgaria, Bulgarian lev"
## [4,] "Cyprus, Euro"
## [5,] "Czech Republic, Czech koruna"
## [6,] "Germany, Euro"
```

- Teraz môžeme priradiť stĺpcom ich názvy a získať dáta, s ktorými budeme ďalej pracovať:

```
colnames(data) <- nazvy
head(data)
```

```
##         Austria, Euro Belgium, Euro Bulgaria, Bulgarian
## 2022Jul          1.70           1.80                   1
## 2022Jun          2.07           2.13                   1
## 2022May          1.54           1.58                   1
## 2022Apr          1.29           1.30                   1
## 2022Mar          0.72           0.79                   1
## 2022Feb          0.54           0.59                   0
##         Czech Republic, Czech koruna Germany, Euro Denma
## 2022Jul                         4.40          1.08
## 2022Jun                         5.12          1.45
## 2022May                         4.61          0.95
```

- V tomto cvičení zoberieme dáta pre Rakúsko:

```
x <- data[, "Austria, Euro"]
x <- x[length(x):1]  # usporadame od najstarsich
```

- Z vektora spravíme časový rad funkciou `ts` s parametrami:
  - použité dáta
  - `frequency` - frekvencia dát: 1 pre ročné dáta, 4 pre kvartálne, 12 pre mesačné
  - `start` alebo `end` - v závislosti od frekvencie, napr. `start = 2000` pri ročných dátach, `start = c(2000, 1)` pri kvartálnych znamená prvý kvartál roku 2000
- Ako sme už videli: menšiu časť časového radu vyberieme funkciou `window` s parametrami:
  - použité dáta
  - začiatok a koniec výberu - `start` alebo `end` (ak niektoré nie je zadané, berú sa všetky dostupné)

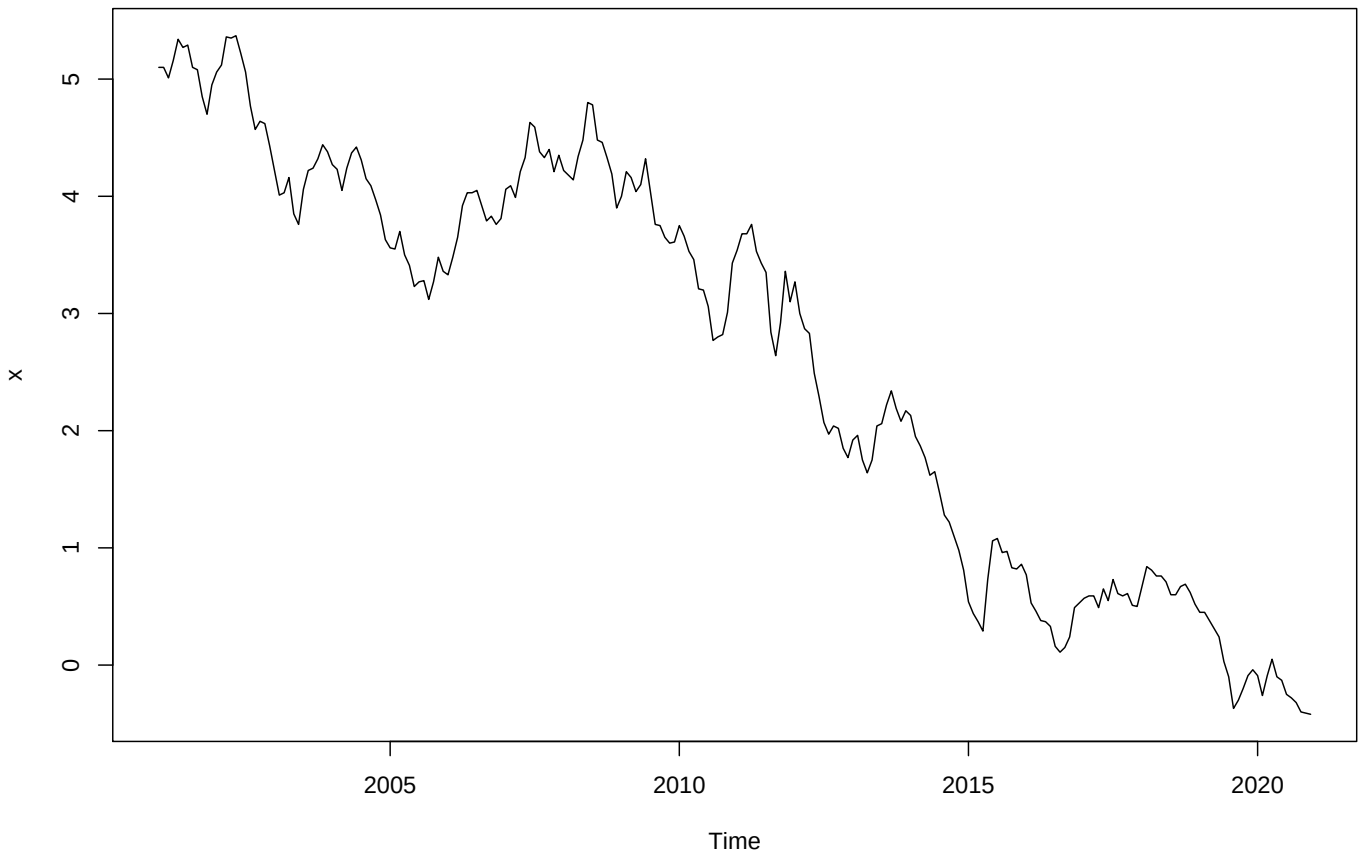
- Chceme mesačné dáta so začiatkom v januári 2001 a koncom v decembri 2020

```
# DOPLNTE
x <- ts(x, ....)
x <- window(x, ...)
```

```r
# kontrola - 20 rokov mesacnych dat
length(x)
```

```
## [1] 240
```

```
plot(x)
```



- Vidíme klesajúci trend, dáta preto nie sú stacionárne (a nemôžeme napríklad odhadovať ACF). **Budeme preto pracovať s diferenciami** (teda medzimesačnými zmenami).
- Vytvoríme diferencie pomocou funkcie `diff` - napríklad:

```
# numericky vektor
y <- c(1, 4, 6, 2)
diff(y)
```

```
## [1]  3  2 -4
```

```
# casove rady
z <- ts(y, frequency = 12, start = c(2022, 1))
diff(z)
```

```
##      Feb Mar Apr
## 2022   3   2  -4
```

- Vytvorte vektor diferencií úrokových mier a zobrazte ich.

```
# DOPLNTE
xDif <- ...
plot(xDif)
```

- **Zaujíma nás, či sú tieto diferencie korelované alebo či ich môžeme považovať za biely šum posunutý o konštantu.**
  - Zobrazte ACF. Ktoré autokorelácie sú signifikantné?
  - Zobrazte výsledky LB testu pre rôzne lagy a skomentujte.
  - Zhodnoťte výsledky. Aký je váš záver ohľadom charakteru daného časového radu?

Cvičenie 4: Testovacia štatistika Ljung-Boxovho testu

- Vygenerujme si dáta, ktoré **sú** nekorelované. Napríklad:

```
x <- rnorm(100, mean = 10, sd = 3)  # iid N(10, 3^2)
```

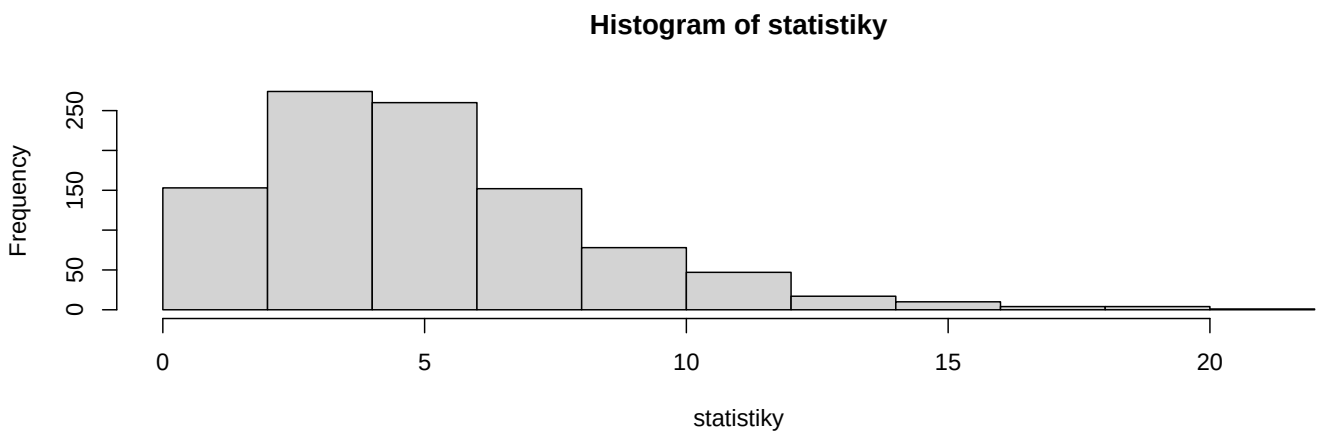
- Budeme testovať hypotézu, že prvých 5 autokorelácií je nulových a zaznamenáme hodnotu testovacej štatistiky:

```
# DOPLNTE
simulacia <- function(){
  x <- rnorm(100, mean = 10, sd = 3)
  LB <- Box.test(x, ...)
  return(...)
}
```

- Zopakujeme 1000 krát - pri opakovaní simulácií je užitočná funkcia `replicate` - a zobrazíme histogram:

```
statistiky <- replicate(1000, simulacia())
hist(statistiky)
```

- Aká je teoretická hustota rozdelenia štatistiky v Ljung-Boxovom teste? Porovnajte so získaným histogramom.
- Ukážka výstupu (kvôli náhodnosti môžete dostať iný):



- Zopakujte simulácie:
  - pre iný počet dát,
  - pre iné rozdelenie generovaných náhodných veličín, ktoré sú bielym šumom,
  - pre iný počet lagov.
- Zobrazte:
  - histogram a jeho porovnanie s hustotou,
  - kvantil a jeho porovnanie s kritickou hodnotou LB testu.

Cvičenie 5: Sila Ljung-Boxovho testu

- Vypočítajte ACF procesu $x_t = u_t - \beta u_{t-1}$, kde u_t je biely šum (na prednáške sme mali prípad $\beta = -1$).
- Pre $\beta \neq 0$ dostaneme nenulovú prvú hodnotu ACF. LB testom budeme testovať, či sú prvé dve hodnoty ACF nulové. Skutočnosť je samozreme taká, že nie sú.

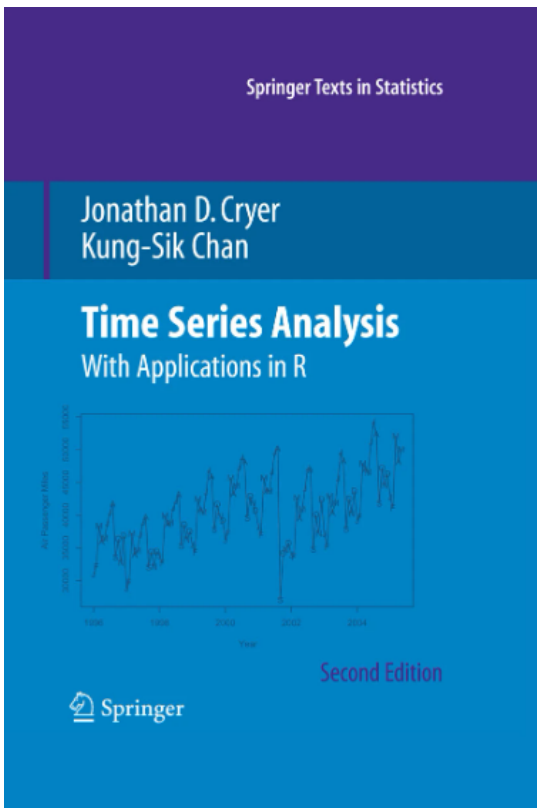
```
set.seed(123) # kvoli reprodukovatelnosti
N <- 101
u <- rnorm(N, mean = 0, sd = 1)
x <- u[1:(N - 1)] + 3*u[2:N] # 100 hodnot procesu
                             # beta = -3
LB <- Box.test(x, lag = 2, type = "Ljung-Box")
LB$p.value
```

```
## [1] 0.03218388
```

- Dal Ljung-Boxov test v predchádzajúcej simulácii správny výsledok?
- Napíšte funkciu, ktorá pre zadaný parameter procesu β a dĺžku časového radu vráti TRUE/FALSE podľa toho, či sa nulová hypotéza zamietla. Pomocou `simulate` odhadnite, aká je pravdepodobnosť toho, že sa testovaná hypotéza zamietne.
- Graficky zobrazte odhadnutú silu testu ako funkciu parametra procesu β.
- Zobrazte aj funkciu, ktorá vyjadruje závislosť ACF(1) od parametra β a vysvetlite, ako súvisia priebehy týchto dvoch funkcií .

Cvičenie 6: Teoretické príklady I.

- Výpočet strednej hodnoty, disperzie, autokovariancií a autokorelácií.
- Overovanie stacionarity.



- Jonathan D. Cryer, Kung-Sik Chan: *Time Series Analysis With Applications in R. Second Edition.* Springer, New York, 2008
- Cvičenia ku kapitole 2: **2.4–2.16**
- Dostupné zo školskej siete: https://link.springer.com/book/10.1007/978-0-387-75959-3

Cvičenie 7: Teoretické príklady II.

Príklad 1. Výpočítajte ACF procesu zo str. 36 prednáškových slajdov. Porovnajte s výsledkom simulácie procesu a odhadnutia ACF vygenerovaných hodnôt.

Príklad 2. Majme dané postupnosti nezávislých rovnako rozdelených náhodných premenných X_t, Y_t, pričom:

- $P(X_t = 0) = P(X_t = 1) = 1/2$,
- $P(Y_t = -1) = P(Y_t = 1) = 1/2$,
- Pre ľubovoľné s, t sú X_t a Y_s nezávislé.

Definujme $Z_t = X_t(1 - X_{t-1})Y_t$. Dokážte, že

- Z_t je biely šum,
- Z_t nie je postupnosť nezávislých rovnako rozdelených náhodných premenných.