Математички факултет
Универзитет у Београду

Мастер рад

Тестови случајности у временским серијама

Ментор:
проф.др Весна Јевремовић

Студент:
Јована Трајковић

Београд, Септембар 2013.

Садржај

Предговор

У овом мастер раду бавимо се тестирањем случајности временских серија. Рад се састоји из пет поглавља.

У уводном делу објашњава се појам временске серије и дефинише се проблематика мастер рада.

Друго поглавље уводи појам случајног процеса, даје формалну дефиницију временске серије, као и основне појмове у анализи временских серија. У овом поглављу се даље бавимо и класификацијом и компонентама временских серија.

У трећем поглављу се уводи појам случајности и објашњавају појмови везани за статистичке тестове. У наставку су објашњени неки конкретни тестови за тестирање случајности временских серија.

У четвртом поглављу се поменути тестови из претходног поглавља примењују на конкретним временским серијама. Овај практични део је заснован на реалним подацима који су обрађени у програмском пакету R.

Коначно, у последњем поглављу су наведени закључци до којих смо дошли примењивањем теорије на конкретне податке.

За избор и помоћ у изради мастер рада велику захвалност дугујем ментору, проф. др Весни Јевремовић. Захваљујем се и осталим професорима, асистентима и сарадницима у настави на указаном знању током студирања.

Посебну захвалност дугујем својој породици, мајци Живки и оцу Милану, Јасмини и Ивану.

Београд, септембар 2013. *Јована Трајковић*

1 Увод

У свакодневном животу се сусрећемо са појавама које имају свој ток у времену и које су предмет изучавања различитих наука. Тако се демографија бави прикупљањем података о годишњим стопама наталитета и морталитета, односно проучавањем природног прираштаја. Економија се бави анализом променљивих као што су цене производа, плате, девизни курс, запосленост, индустријска производња, итд. У метеорологији се сваког сата региструје брзина ветра, прати дневна температура, бележи месечна или годишња количина падавина. Пољопривреда прати годишње кретање приноса појединих пољопривредних култура, као и њихове откупне и продајне цене. Све ове појаве су примери временских серија.

Временска серија се може схватити као уређени низ података. Најчешће се уређење врши у односу на време и у једнаким временским интервалима. Избор јединице времена врши се у односу на саму посматрану појаву па то може бити секунда, минут, сат, дан, месец, година, али и деценија, век или миленијум. Осим по времену уређење података може бити извршено и по некој другој карактеристици тока појаве као што су дубина, висина, растојање, количина неког присутног хемијског елемента, итд. Овакве серије се називају **линијске серије** и сматрају се врстом временских серија.

Изучавање временских серија засновано на теорији вероватноће и математичкој статистици почело је тек тридесетих година двадесетог века, а због интеракције са многим дисциплинама, посебно економијом, оно бележи динамични развој последњих неколико деценија. Статистичка дисциплина која се бави овим изучавањима јесте **анализа временских серија**. Њени принципи одступају од уобичајених претпоставки теорије статистичког закључивања па се временске серије не могу изучавати методама класичне теорије вероватноће и класичне математичке статистике. Наиме, у класичној теорији вероватноће и математичкој статистици је основни појам случајан узорак који представља низ независних случајних величина. Насупрот томе временска серија је низ случајних величина које су међусобно зависне, а најчешће и корелисане.

Основни циљеви у анализи временских серија су описивање, објашњење и предвиђање временске серије. Како је процес доношења одлука у различитим областима често повезан са предвиђањем будућих вредности променљивих које зависе од времена, временске серије и њихова анализа представљају погодно средство. Предвиђање подразумева анализу историјских података о датој појави и екстраполацију истих у будућност користећи неки математички модел. Циљ изучавања временске серије може бити и установљавање механизма који њу генерише како бисмо схватили утицај појединих фактора на ток временске серије. То нам даље омогућава формирање одговарајућег математичког модела.

Приликом изучавања временских серија једно од првих питања које се јавља јесте питање случајности временске серије, односно да ли временска серија представља реализацију једног низа случајних величина које су међусобно независне и имају исту расподелу вероватноћа. Тестирање случајности је често од фундаменталног значаја у статистичкој анализи података због чега је развијен велики број различитих тестова. Управо ти тестови и њихова практична примена су предмет изучавања овог рада.

2 Временске серије

Временска серија се обично дефинише као низ података са одређеним хронолошким или линијским редоследом. Међутим дефиницији се може прићи и са становишта случајних процеса као што ће бити описано у наставку.

2.1 Случајни процеси

До појма случајног процеса долази се проширивањем појма случајне променљиве па уводимо следећих неколико дефиниција.

Дефиниција 2.1.1 Ω *је скуп свих исхода једног експеримента. Елементи овог скупа се означавају са* ω_i, $i = 1, 2, 3, ...$ *и називају елементарни догађаји.*

Дефиниција 2.1.2 *Подскуп* $\mathcal{F}$ *партитивног скупа* $\mathcal{P}(\Omega)$ *је алгебра над* Ω *ако важе услови:*

- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F} \Rightarrow \overline{A} \in \mathcal{F}$
- $A \in \mathcal{F} \bigwedge B \in \mathcal{F} \Rightarrow A \bigcup B \in \mathcal{F}$.

Дефиниција 2.1.3 *Подскуп* $\mathcal{F}$ *партитивног скупа* $\mathcal{P}(\Omega)$ *је* σ- *алгебра над* Ω *ако важе услови:*

- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F} \Rightarrow \overline{A} \in \mathcal{F}$
- $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F} \Rightarrow \bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}$

Дефиниција 2.1.4 *Ако је* $\mathcal{K}$ *нека фамилија подскупова простора исхода* Ω *(колекција), тада пресек свих* σ- *алгебри које садрже фамилију* $\mathcal{K}$ *називамо минимална* σ- *алгебра генерисана са* $\mathcal{K}$ *и означавамо је* $\sigma[\mathcal{K}]$.

Аналогно се дефинише минимална алгебра која садржи све скупове дате колекције.

Дефиниција 2.1.5 *Нека је простор исхода једнак скупу реалних бројева* $\mathbb{R}$ *и нека је колекција* $\mathcal{K}$ *дата са* $\mathcal{K} = \{(a, b] : a, b \in \mathbb{R}, a < b\}$. *Минималну алгебру која садржи колекцију* $\mathcal{K}$ *означавамо са* $\mathcal{B}_0$. *Минимална* σ- *алгебра која садржи колекцију* $\mathcal{K}$ *зове се Борелова* σ- *алгебра подскупова реалне праве. Означавамо је са* $\mathcal{B}$. *Њени елементи зову се Борелови скупови на реалној правој.*

Дефиниција 2.1.6 *Нека је Ω простор исхода случајног експеримента и $\mathcal{F}$ σ- алгебра догађаја. Уређени пар $(\Omega, \mathcal{F})$ се зове мерљив простор.*

Дефиниција 2.1.7 *Нека је Ω скуп свих елементарних догађаја и нека је $\mathcal{F}$ σ-алгебра над Ω. Функција $P : \mathcal{F} \to \mathbb{R}$ се зове вероватноћа или вероватносна мера на мерљивом простору $(\Omega, \mathcal{F})$ ако има следећа својства:*

- *својство нормираности - $P(\Omega) = 1$*
- *својство ненегативности - $P(A) \geq 0$, за сваки догађај $A \in \mathcal{F}$*
- *својство σ адитивности вероватноће - ако $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$, $A_i \bigcap A_j = \emptyset$, $i \neq j$, $i = 1, 2, 3, ...$, $j = 1, 2, 3, ...$, онда $P\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \sum\limits_{i=1}^{\infty} P(A_i)$*

Простор вероватноћа је уређена тројка $(\Omega, \mathcal{F}, P)$.

Дефиниција 2.1.8 *Нека је $(\Omega, \mathcal{F}, P)$ простор вероватноћа. Реално пресликавање $X : \Omega \to \mathbb{R}$ се зове случајна променљива ако $\forall B \in \mathcal{B}$ важи да је $X^{-1}(B) \in \mathcal{F}$ где је $\mathcal{B}$ Борелова σ-алгебра. За X кажемо да је $\mathcal{F}$-мерљиво.*

Дефиниција 2.1.9 *Фамилија случајних променљивих $\{X_t(\omega) : t \in T, \omega \in \Omega\}$ дефинисаних над истим простором вероватноће $(\Omega, \mathcal{F}, P)$ се зове случајни (стохастички) процес са индексним скупом T.*

Приликом записа се често изоставља променљива ω, па се уместо тога случајни процес означава са $\{X_t, t \in T\}$. Процес очигледно зависи од две променљиве t и ω. За фиксирано $t \in T$ случајни процес је једна променљива. За фиксирано $\omega \in \Omega$ функција $X_t(\omega)$, $t \in T$ се зове *трајекторија* случајног процеса. Ако пак фиксирамо обе променљиве X_t је реалан број или једна реализација случајног процеса.

Скуп T из претходне дефиниције се још назива и *параметарски скуп*. Ако је $T = \mathbb{R} = (-\infty, \infty)$, $T = [0, \infty)$ или $T = [a, b] \subseteq \mathbb{R}$, онда се параметар $t \in T$ обично интерпретира као време, а случајна функција $\{X_t, t \in T\}$ зове се *случајан процес са непрекидним временом*. Ако је $T \subseteq \mathbb{Z}$ или $T \subseteq \mathbb{N}$, онда се случајна функција $\{X_t, t \in T\}$ зове *случајан процес са дискретним временом* или *случајни низ*.

Дефиниција 2.1.10 *За сваки природан број n и произвољне елементе t_1,...,$t_n \in T$, $t_1 < t_2 < ... < t_n$, расподела вероватноћа случајног вектора $(X_{t_1}, X_{t_2}, ..., X_{t_n})$ одређена је функцијом расподеле*

$$F_{t_1,...,t_n}(x_1, ..., x_n) = P\{X_{t_1} \leq x_1, ..., X_{t_n} \leq x_n\}.$$

Фамилија свих ових расподела зове се фамилија коначнодимензионих расподела случајног процеса $\{X_t, t \in T\}$.

Фамилија коначнодимензионих расподела задовољава следеће услове:

- **услов симетрије** - ако је $i_1, ..., i_n$ једна пермутација бројева од 1 до n онда важи
$$F_{t_{i_1},...,t_{i_n}}(x_{i_1}, ..., x_{i_n}) = F_{t_1,...,t_n}(x_1, ..., x_n)$$
- **услов сагласности** - ако је $m < n$ за произвољне $t_{m+1}, ..., t_n \in T$ важи
$$F_{t_1,...,t_m,t_{m+1},...,t_n}(x_1, ..., x_m, \infty, ..., \infty) = F_{t_1,...,t_m}(x_1, ..., x_m)$$

Нека је $\{X_t, t \in T\}$ случајан процес. Тада је:

- средња вредност случајног процеса
$$E(X_t) = \mu_t, t \in T$$
- варијанса случајног процеса
$$D(X_t) = \sigma_t^2, t \in T$$
- коваријанса случајног процеса
$$\gamma(r, s) = cov(X_r, X_s) = E((X_r - E(X_r))(X_s - E(X_s))), r, s \in T$$
- корелација случајног процеса
$$\rho(r, s) = \frac{cov(X_r, X_s)}{\sqrt{D(X_r)D(X_s)}}$$

Дефиниција 2.1.11 *Случајан процес је строго стационаран ако су његове коначнодимензионе функције расподеле инваријантне у односу на t, односно ако за све $t_i, t_i + t \in T$, $i = 1, 2, ...$ важи:*
$$F_{t_1,...,t_n}(x_1, ..., x_n) = F_{t_1+t,...,t_n+t}(x_1, ..., x_n).$$

Дефиниција 2.1.12 *Случајан процес $\{X_t, t \in T\}$ је слабо стационаран ако су испуњени следећи услови:*

- $E(X_t) = m = const$, $\forall t \in T$
- $E(X_t^2) < \infty$, $\forall t \in T$
- $\gamma(r, s) = \gamma(r + t, s + t)$, $\forall r, s, t \in T$.

Сада можемо увести и формалну дефиницију временских серија.

Дефиниција 2.1.13 *Временска серија је једна реализација случајног процеса.*

У смислу претходне дефиниције однос временске серије и случајног процеса одговара односу узорка и основног скупа у стандардној теорији статистичког закључивања. Као што узорак представља део основног скуп на основу кога се доносе закључци о самом скупу, тако и временска серија представља део случајног процеса који треба да омогући уочавање карактеристика тог процеса.

2.2 Класификација временских серија

Можемо извршити класификацију временских серија користећи различите критеријуме. Најједноставнија је подела на непрекидне и дискретне (прекидне) временске серије. Као и код случајних процеса та подела зависи од параметарског скупа. Временска серија је непрекидна ако њене вредности можемо регистровати у сваком временском тренутку. Прекидна временска серија је она чије вредности бележимо у једнаким временским интервалима (дан, сат, месец, година). Прекидну временску серију можемо добити и од непрекидне тако што ћемо бележење резултата посматране појаве вршити само у одређеним временским интервалима.

Кључна подела временских серија јесте подела на стационарне и нестационарне. Слободно речено, серија је стационарна уколико је њено кретање предвидиво током времена, тј. уколико њено кретање испуњава сличан образац понашања током времена. Супротно, ако су параметри кретања временске серије функције временског тренутка, тада је она нестационарна. Као што је већ поменуто постоје два вида стационарности. У случају дефиниције 2.1.11 серија је стационарна ако се њена својства не мењају транслирањем у времену. То значи да случајне променљиве које припадају тој серији имају исту очекивану вредност, варијансу и моменте вишег реда. Ако пак узмемо дефиницију 2.1.12 у обзир онда за стационарну серију важи да се очекивана вредност и варијанса не мењају током времена док је коваријанса између свака два члана серије само функција временског растојања између њих. На графику 2.2.1 се може видети разлика између стационарне и нестационарне временске серије.

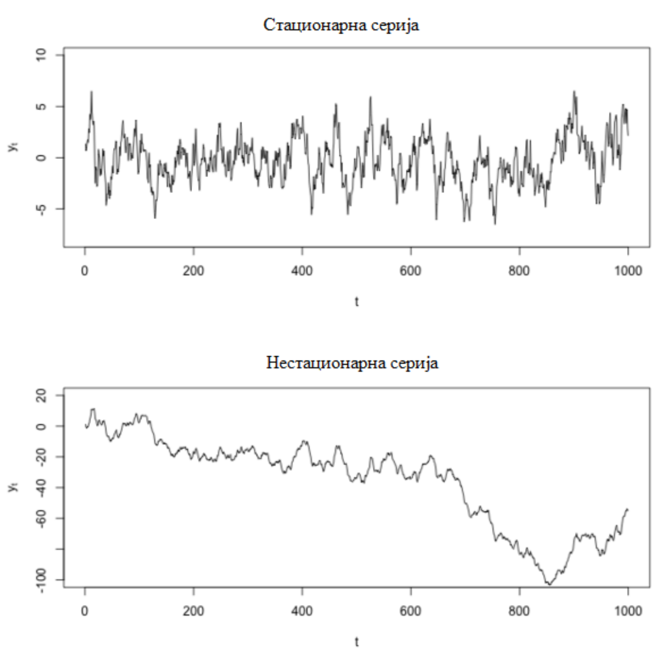

График 2.2.1: Пример стационарне и нестационарне временске серије (преузето са *http : //en.wikipedia.org/wiki/File : Stationarycomparison.png*)

2.3 Компоненте временске серије

Временске серије, у зависности од појаве, се могу састојати од различитих компоненти. Традиционално се претпоставља да се временска серија може разложити на четири компоненте: тренд $f(t)$, цикличну компоненту $c(t)$, сезонску компоненту $s(t)$ и случајну компоненту $\varepsilon(t)$. У случају када серија садржи све ове компоненте типови модела који се најчешће користе су:

- **адитивни модел** - $X_t = f(t) + s(t) + c(t) + \varepsilon(t)$
- **мултипликативни модел** - $X_t = f(t) \cdot s(t) \cdot c(t) \cdot \varepsilon(t)$
- **мешовити модел** - $X_t = f(t) \cdot s(t) \cdot c(t) + \varepsilon(t)$.

Тренд временске серије описује њено основно понашање у дужем временском периоду, или прецизније у читавом временском периоду за који је временска серија позната. У зависности од тога да ли вредности временске серије током времена систематски расту или опадају, тренд може бити растући или опадајући. Осим тога, тренд може бити детерминистички или стохастички у зависности од тога да ли се промене серије могу предвидети или не.

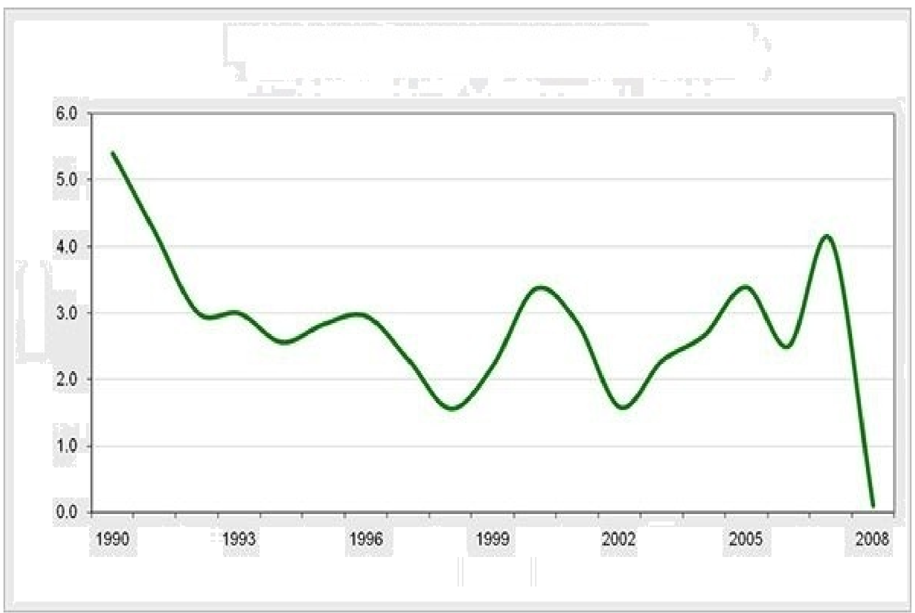

График 2.3.1: Пример временске серије са стохастичким трендом: *Процентуална промена потрошачког индекса у САД*

У практичним применама већина временских серија представљају податке прикупљене у узастопним временским интервалима као што су дани, месеци, квартали, итд. Због тога поједине временске серије испољавају правилности у свом кретању које се јављају у току једне календарске године. За такве серије кажемо да имају сезонску компоненту. Присуство сезонске компоненте утиче на то да постоји већи степен корелације између података добијених у истим месецима различитих година него између оних добијених у суседним месецима исте године.

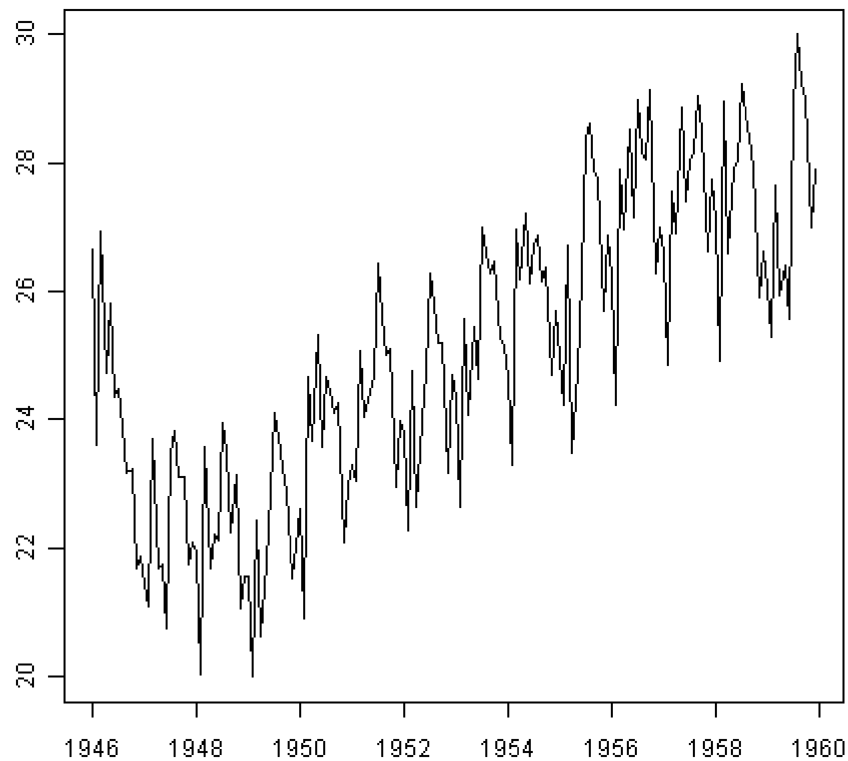

График 2.3.2: Пример временске серије са сезонском компонентом: *Број рођених по месецу у Њујорку од јануара 1946. до децембра 1959.*

Циклична компонента представља слично понашање које се јавља у серији током неког периода времена, односно можемо рећи да се у серији запажа периодични карактер промене. Овакве серије имају своју периоду, односно циклус.

Претходне три компоненте се једним именом називају неслучајне компоненте. Ниједна од ових компоненти није обавезна у моделу. Једина компонента коју сваки модел мора да има је случајна компонента. Ова компонента служи за објашњавање непредвидивих флуктуација.

3 Тестови случајности

3.1 Статистички тестови

Случајне појаве у математичкој статистици описују се математичким моделом. Избор тог модела зависи од предзнања о посматраној појави. У многим истраживањима се региструју вредности неке случајне величине чија расподела није унапред позната. У таквим случајевима се статистички модел задаје фамилијом допустивих расподела. Параметарска фамилија допустивих расподела, која зависи од једног или више непознатих параметара, у општем случају се означава са $\mathcal{P} = \{P_\upsilon : \upsilon \in \Theta\}$, где је υ параметар, а Θ параметарски простор. Свака претпоставка која сужава класу допустивих расподела назива се **статистичка хипотеза**.

Разликујемо параметарске и непараметарске хипотезе. *Параметарска* статистичка хипотеза сужава скуп допустивих вредности неких параметара од којих зависи расподела. У *непараметарској* статистичкој хипотези претпоставља се да расподела обележја има дати тип, а не тиче се вредности параметара. Ако статистичка хипотеза једнозначно одређује расподелу обележја, онда се она назива *проста* хипотеза. У супротном хипотеза је *сложена.*

Правило помоћу кога се на основу узорка одлучује да ли хипотезу треба прихватити као тачну или одбацити назива се **статистички тест**. У наставку ће бити описан општи поступак извођења једног теста.

Означимо почетну хипотезу са H_0. Нека је дат случајан узорак $(X_1, X_2, ..., X_n)$ који узима вредности у скупу $\mathbb{R}^n$. Нека је α број из интервала $(0, 1)$ и W подскуп скупа $\mathbb{R}^n$. Нека још важи да вероватноћа да узорак $(X_1, X_2, ..., X_n)$ узме вредност из скупа W ако је почетна хипотеза тачна, није већа од броја α, односно:

$$P\{(X_1, X_2, ..., X_n) \in W | H_0\} \leq \alpha.$$

Дефиниција 3.1.1 *Скуп $W \subset \mathbb{R}^n$ зове се критична област за хипотезу H_0, а број α зове се праг значајности или величина критичне области. Статистички тест на основу критичне области W одређујемо на следећи начин: Ако реализовани узорак $(x_1, x_2, ..., x_n)$ припада критичној области, онда хипотезу H_0 одбацујемо. Ако реализовани узорак не припада критичној области, онда хипотезу H_0 прихватамо.*

Критична област обично има неки од следећих облика:

$$\{(x_1, x_2, ..., x_n) : T_n(x_1, x_2, ..., x_n) \geq C\},$$

$$\{(x_1, x_2, ..., x_n) : T_n(x_1, x_2, ..., x_n) \leq C\},$$

$$\{(x_1, x_2, ..., x_n) : |T_n(x_1, x_2, ..., x_n|) \geq C\}.$$

У прва два случаја кажемо да је критична област *једнострана*, а у последњем да је *двострана*. Величина C представља константу која се добија из статистичких таблица за одговарајућу расподелу.

Критична област се обично одређује помоћу неке статистике $T_n(X_1, X_2, ..., X_n)$ која се назива **тест статитика**. Под претпоставком да је тачна хипотеза H_0 позната је тачна расподела или приближна расподела за узорке великог обима тест статистике. Тест статистика представља одступање реализованог узорка од вредности које је природно очекивати под претпоставком да је почетна хипотеза тачна.

Хипотеза H_0 која се тестира се назива *нулта хипотеза*. Претпоставка да обележје има расподелу која припада оном делу фамилије допустивих расподела који није обухваћен нултом хипотезом назива се *алтернативна хипотеза* и обележава са H_1. У случају одбацивања нулте хипотезе прихвата се алтернативна хипотеза.

Пре почетка тестирања задаје се величина критичне области α. За α се обично узимају мале вредности, $\alpha = 0.05$, $\alpha = 0.01$. Ако обележимо:

$$P\{T_n \notin W|H_1\} = \beta$$

онда ћемо вредност $1 - \beta$ звати **моћ теста**.

Треба још напоменути да при извођењу сваког теста постоји могућност прављења грешке. Дефинише се грешка прве врсте као одбацивање тачне хипотезе H_0. Вероватноћа да дође до ове грешке је једнака α. Грешка друге врсте је прихватање хипотезе H_0 када она није тачна. Вероватноћа да дође до ове грешке је једнака β. Много озбиљнија је грешка прве врсте због чега и бирамо α тако да буде што мање.

Како би до грешака долазило са малом вероватноћом пожељно је да статистички тест има својства центрираности и постојаности.

Дефиниција 3.1.2 *Нека је $\mathcal{F}$ класа допустивих функција расподела обележја X и нека је $W \subset \mathbb{R}^n$ критична област величине α за тестирање хипотезе да стварна функција расподеле обележја X припада скупу $H_0 \subset \mathcal{F}$. Свака функција расподеле F за коју важи $F \in \mathcal{F} \backslash H_0$ назива се алтернативна функција расподеле.*

Дефиниција 3.1.3 *Нека је $\mathcal{F}$ класа допустивих функција расподеле обележја X и нека је $W \subset \mathbb{R}^n$ критична област за тестирање хипотезе $H_0 \subset \mathcal{F}$ са прагом значајности α на основу узорка $(X_1, ..., X_n)$. Статистички тест заснован на критичној области W је центриран ако за сваку алтернативну функцију расподеле $F \in \mathcal{F} \backslash H_0$ важи неједнакост $P\{(X_1, ..., X_n) \in W|F\} > \alpha$.*

Центрираност теста, дакле, означава да је већа вероватноћа да ће нулта хипотеза H_0 бити одбачена у случају када је она нетачна, него у случају када је тачна. Овом особином се обезбеђује мала вероватноћа грешке прве врсте.

Дефиниција 3.1.4 *Нека је $\mathcal{F}$ класа допустивих функција расподеле обележја X и нека је $W \subset \mathbb{R}^n$ критична област за тестирање хипотезе $H_0 \subset \mathcal{F}$ са прагом значајности α на основу узорка $(X_1, ..., X_n)$. Статистички тест заснован на критичној области W је постојан ако за сваку алтернативну функцију расподеле $F \in \mathcal{F} \backslash H_0$ важи једнакост*

$$\lim_{n \to \infty} P\{(X_1, ..., X_n) \in W|F\} = 1\,.$$

Другачије речено, постојан статистички тест са вероватноћом која тежи јединици препознаје одступања стварне функције расподеле обележја од нулте хипотезе. Ова карактеристика обезбеђује велику вредност моћи теста.

3.2 Тестови случајности у временским серијама

Као што је већ поменуто претпоставка од које се полази и која се природно намеће при анализи временских серија јесте хипотеза случајности. Исправност ове хипотезе значи да се приликом изучавања серије постојећи подаци могу узимати у произвољном поретку. Одбацивање, пак, ове хипотезе значи да се подаци морају разматрати у датом временском редоследу и истраживач никако не сме нарушити тај редослед.

Избор начина тестирања нулте хипотезе зависи од алтернативних хипотеза. Међутим, за разлику од тестирања у класичној математичкој статистици, у анализи временских серија се обично не наводе алтернативне хипотезе. Оне могу бити различите и зависе од самог начина тестирања. Код неких тестова ће бити погодно изабрати за алтернативу растући тренд, док ће код других на пример бити погодно изабрати циклично или сезонско колебање у серији.

Тестови се разликују по тест статистикама, које пак могу бити засноване на различитим стварима. Па тако тестови могу бити засновани на међусобном односу чланова серије (тест тачака заокрета, тест тачака раста, Кендалов тест), могу користити узорачку медијану (Фицов тест, тест Валда Волфовица, тест серија), рангове (тест разлике рангова, Бартелсов тест) или пак преласке или кораке. Од тога како су настале тест статистике зависиће формирање алтернативне хипотезе, као и које ће мане, односно добре стране, тај тест имати.

У наставку ће бити описан низ различитих тестова случајности који се примењују у анализи временских серија.

3.3 Тест тачака заокрета

Претпоставимо да изучавамо непрекидно обележје X и да смо изучавањем добили коначну временску серију $X_1, ..., X_n$. Желимо да проверимо да ли за ову конкретну серију важи хипотеза случајности H_0. Ако је неки елемент серије X_j већи од оба своја суседа X_{j-1} и X_{j+1}, онда кажемо да је X_j пик, врх или локални максимум и то означавамо: $X_{j-1} < X_j > X_{j+1}$. Ако је пак X_j мањи од оба своја суседа, онда кажемо да је X_j долина, дно или локални минимум и то означавамо: $X_{j-1} > X_j < X_{j+1}$. Локалне максимуме и минимуме једним именом називамо **тачке заокрета**. Тест који ћемо користити да би проверили хипотезу H_0 је заснован на овим тачкама.

Дефинишемо **индикатор заокрета** на следећи начин:

$$Y_j = \begin{cases} 1 & \text{, ако је } X_{j-1} < X_j > X_{j+1} \text{ или } X_{j-1} > X_j < X_{j+1} \\ 0 & \text{, иначе} \end{cases}$$

при чему је $j = 2, 3, ..., n-1$, јер први и последњи члан серије не могу бити тачке заокрета. Поставља се питање коју расподелу има овај индикатор. Како бисмо то одредили посматраћемо бројеве 1, 2 и 3. Исписујемо све могуће пермутације ова три броја.

$$123, \boxed{132}, \boxed{213}, \boxed{231}, \boxed{312}, 321.$$

Уоквирене су оне комбинације у којима се јавља тачка заокрета. Од могућих 6 комбинација у другој и четвртој се јавља локални максимум, а у трећој и петој локални минимум. Односно постоје четири тачке заокрета. Дакле уз претпоставку да је хипотеза H_0 тачна индикатор ће имати расподелу:

$$Y_j : \begin{pmatrix} 0 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

Одавде видимо да је $E(Y_j) = \frac{2}{3}$ и $D(Y_j) = \frac{2}{9}$.

Посматрајмо сада укупан број тачака заокрета у серији дужине n. То је случајна променљива облика:

$$Z_n = Y_2 + ... + Y_{n-1}.$$

Сабирци у претходној једначини нису међусобно независни. Како је расподелу ове случајне променљиве тешко одредити, нарочито за велике вредности n, окренућемо се рачунању њеног математичког очекивања и дисперзије. Директно се може израчунати да је математичко очекивање укупног броја тачака заокрета једнако:

$$E(Z_n) = E(Y_2 + ... + Y_{n-1}) = E(Y_2) + ... + E(Y_{n-1}) = (n-2) \cdot \frac{2}{3}.$$

Рачунање дисперзије Z_n биће мало компликованије. Полазимо од рачунања вредности $E(Z_n^2)$:

$$Z_n^2 = \left(\sum_{j=2}^{n-1} Y_j\right)^2 = \sum_{j=2}^{n-1} Y_j^2 + 2\sum_{j=2}^{n-2} Y_jY_{j+1} + 2\sum_{j=2}^{n-3} Y_jY_{j+2} + 2\sum_{k\geq 3}\sum_{j=2}^{n-1-k} Y_jY_{j+k}$$

$$EZ_n^2 = \sum_{(n-2)} EY_j^2 + 2\sum_{(n-3)} E(Y_jY_{j+1}) + 2\sum_{(n-4)} E(Y_jY_{j+2}) + 2\sum_{(m)} E(Y_jY_{j+k}).$$

У претходној једначини број у загради испод знака сумирања означава број сабирака у том збиру. Можемо израчунати m на следећи начин:

$$(n-2)^2 = (n-2) + 2(n-3) + 2(n-4) + 2m \Rightarrow$$

$$n^2 - 9n + 20 = 2m \Rightarrow$$

$$2m = (n-4)(n-5).$$

Сада је потребно израчунати непозната очекивања у једначини.

Вредност Y_jY_{j+1} је одређена вредностима X_{j-1}, X_j, X_{j+1}. Посматраћемо све могуће пермутације бројева 1, 2, 3 и 4 и бележити у којим пермутацијама се јављају две тачке заокрета:

$$\begin{array}{cccccc}
1234 & 1243 & \boxed{1324} & 1342 & \boxed{1423} & 1432 \\
2134 & \boxed{2143} & \boxed{2314} & 2341 & \boxed{2413} & 2431 \\
3124 & \boxed{3142} & 3214 & \boxed{3241} & \boxed{3412} & 3421 \\
4123 & \boxed{4132} & 4213 & \boxed{4231} & 4312 & 4321
\end{array}$$

Дакле у 10 пермутација постоје две тачке заокрета па је при претпоставци H_0:

$$E(Y_jY_{j+1}) = \frac{10}{4!} = \frac{5}{12}.$$

Вредност производа Y_jY_{j+2} зависи од X_{j-1}, X_j, X_{j+1}, X_{j+2}, X_{j+3}. Понављањем претходног поступка али за бројеве 1, 2, 3, 4, 5 може се проверити да се у 54 случајева од свих $5! = 120$ пермутација јављају три тачке заокрета. Дакле закључујемо да при претпоставци H_0 важи:

$$E(Y_jY_{j+2}) = \frac{54}{120} = \frac{9}{20}.$$

Остаје још производ Y_jY_{j+k} за вредности $k \geq 3$. Проверимо шта се дешава у случају $k = 3$. На овај производ утичу вредности X_{j-1}, X_j, X_{j+1}, X_{j+2}, X_{j+3}, X_{j+4}. Међутим како на Y_j утичу само X_{j-1} и X_j, а на Y_{j+3} утичу X_{j+2} и X_{j+3}, закључујемо да су Y_j и Y_{j+3} независни. Слично важи и за све вредности $k > 3$. Одавде следи да је:

$$E(Y_jY_{j+k}) = EY_j \cdot EY_{j+k} = \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9}.$$

Користећи све претходне резултате добијамо:

$$E(Z_n^2) = (n-2)\cdot\frac{2}{3} + 2(n-3)\cdot\frac{5}{12} + 2(n-4)\cdot\frac{9}{20} + (n-4)(n-5)\cdot\frac{4}{9} = \frac{40n^2-144n+131}{90}.$$

Сада можемо да израчунамо и дисперзију уз претпоставку да важи H_0:

$$DZ_n = E(Z_n^2) - (EZ_n)^2 = \frac{40n^2-144n+131}{90} - \frac{40n^2-160n+160}{90} = \frac{16n-29}{90}.$$

После дужих израчунавања добија се да је узорачки коефицијент асиметрије $\pi_1 = E(Z_n - EZ_n)^3/\sqrt{DZ_n}^3 \approx -0.14$, док је узорачки коефицијент спљоштености $\pi_2 = E(Z_n - EZ_n)^4/\sqrt{DZ_n}^4 - 3 \approx -2.4/\sqrt{n}$. Дакле за серије великог обима ове вредности су блиске нули. Како ови коефицијенти имају вредност нула код нормалне расподеле можемо закључити да укупан број тачака заокрета има приближно нормалну расподелу при хипотези H_0, односно:

$$Z_n \approx \mathcal{N}\left(\frac{2}{3}(n-2),\, \frac{16n-29}{90}\right).$$

Добијени закључак нам омогућава да вршимо тестирање хипотезе H_0 помоћу нормалне расподеле, при чему тест статистику Z_n можемо да стандардизујемо тако да добијемо стандардну нормалну расподелу:

$$Z_n^* = \frac{Z_n - EZ_n}{\sqrt{D}Z_n} \approx \mathcal{N}(0,\, 1).$$

Из начина тестирања можемо закључити да је овај критеријум ефикасан ако се као алтернативна хипотеза узме претпоставка да у серији постоје систематске флуктуације. Такође се може приметити да ако је број тачака заокрета близак минимуму односно нули, онда аутоматски можемо одбацити нулту хипотезу и закључити да у серији постоје цикличне варијације. Ако је, пак, број тачака заокрета близак максималном броју, тј. $(n-2)$, онда нулту хипотезу одбацујемо са закључком да у серији постоји растући или опадајући тренд.

3.4 Тест тачака раста

Нека је дата коначна временска серија X_1, ..., X_n непрекидног обележја X. Ако за чланове дате серије важи $X_j < X_{j+1}$ кажемо да је X_j **тачка раста**. Дефинишемо **индикатор тачке раста** на следећи начин:

$$U_j = \begin{cases} 1 & , \text{ако је } X_j < X_{j+1} \\ 0 & , \text{ако је } X_j > X_{j+1} \end{cases}$$

где је $j = 1$, ..., $n-1$ јер све тачке осим последње могу бити тачке раста. Одмах можемо закључити да ће под претпоставком да је H_0 тачна важити:

$$U_j : \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Одавде следи:

$$EU_j = \frac{1}{2},\ DU_j = \frac{1}{4}.$$

Укупан број тачака раста биће случајна променљива облика:

$$R_n = U_1 + ... + U_{n-1}\,.$$

Можемо директно израчунати математичко очекивање ове случајне променљиве. Добијамо:

$$ER_n = E(U_1 + ... + U_{n-1}) = EU_1 + ... + EU_{n-1} = (n-1)\cdot\frac{1}{2}\,.$$

Да бисмо израчунали дисперзију искористићемо поступак аналоган поступку који смо користили за тест тачака заокрета.

$$R_n^2 = (U_1 + ... + U_n)^2 = \sum_{j=1}^{n-1} U_j^2 + 2\sum_{j=1}^{n-2} U_jU_{j+1} + 2\sum_{k\geq 2}\sum_{j=1}^{n-1-k} U_jU_{j+k}$$

$$E(R_n^2) = \sum_{(n-1)} E(U_j^2) + 2\sum_{(n-2)} E(U_jU_{j+1}) + 2\sum_{(m)} E(U_jU_{j+k})$$

Даље из:

$$(n-1)^2 = (n-1) + 2(n-2) + 2m \;\Rightarrow$$
$$n^2 - 5n + 6 = 2m$$

добијамо:

$$2m = (n-2)(n-3)\,.$$

Како вредност производа U_jU_{j+1} зависи од вредности X_j, X_{j+1}, X_{j+2} посматрамо све могуће пермутације бројева 1, 2 и 3.

$$\boxed{123}, 132, 213, 231, 312, 321\,.$$

Као што се може приметити само у првој пермутације постоје две тачке расте па закључујемо да је:

$$E(U_jU_{j+1}) = \frac{1}{6}\,.$$

Потребно је још проверити производ U_jU_{j+k} за $k \geq 2$. Када је $k = 2$ на производ утичу вредности X_j, X_{j+1}, X_{j+2}, X_{j+3}. Међутим, како на вредност U_j утичу само X_j и X_{j+1}, а на вредност U_{j+2} само X_{j+2} и X_{j+3} закључујемо да су ове две величине независне и да је:

$$E(U_jU_{j+2}) = EU_j \cdot EU_{j+2} = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}\,.$$

Исто важи и за свако $k > 2$. На основу ових резултата добијамо да је:

$$E(R_n^2) = (n-1)\cdot\frac{1}{2} + 2(n-2)\cdot\frac{1}{6} + (n-2)(n-3)\cdot\frac{1}{4} = \frac{3n^2-5n+4}{12}\,.$$

Односно:

$$DR_n = E(R_n^2) - (ER_n)^2 = \frac{3n^2 - 5n + 4}{12} - \frac{n^2 - 2n + 1}{4} = \frac{n+1}{12}\,.$$

За велику вредност n тест статистика R_n ће при нултој хипотези имати приближно нормалну расподелу

$$R_n \approx \mathcal{N}\left(\frac{n-1}{2}, \frac{n+1}{12}\right)\,.$$

Ово нам обезбеђује стандардну примену статистичког тестирања и проналажења интервала поверења за укупан број тачака раста у серији.

Овај тест је ефикасан за откривање монотоног тренда у серији. Ако је број тачака раста у серији једнак $n-1$ онда је у питању растући тренд, а ако у серији нема тачака раста онда је у питању опадајући тренд. Може се још закључити да у серији постоји цик-цак кретање ако је број тачака раста приближно $\frac{n-1}{2}$. Тестирање ће бити неефикасно ако се за алтернативну хипотезу узме симетрично варирање. Тест се на идентичан начин може применити и ако се уместо тачака раста узму тачке опадања.

3.5 Фицов тест

Фицов тест користи узорачку медијану за формирање тест статистике. Медијана се описује као број који раздваја горњу половину узорка, популације или расподеле вероватноће од доње половине. Медијана коначног низа бројева се може наћи тако што се бројеви поређају по величини и узме се средњи члан низа. Уколико постоји паран број чланова низа, медијана није јединствена, па се често узима аритметичка средина две вредности које су кандидати за медијану.

Нека је $X_1, ..., X_n$ коначна временска серија непрекидног обележја X. Дефинишемо индикаторе:

$$M_j = \begin{cases} 1 & \text{, ако је } X_j > \text{медијане} \\ 0 & \text{, ако је } X_j < \text{медијане} \end{cases}$$

При хипотези случајности H_0 важиће:

$$M_j : \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}\,,$$

па је:

$$EM_j = \frac{1}{2},\, DM_j = \frac{1}{4}.$$

Помоћу индикатора затим формирамо тест статистику:

$$S_n = M_1 + ... + M_n.$$

Како су M_j међусобно независни из претходних закључака следи:

$$ES_n = E(\sum_{(n-1)} M_j) = \sum_{(n-1)} EM_j = (n-1)\cdot\frac{1}{2}$$

$$DS_n = D(\sum_{(n-1)} M_j) = \sum_{(n-1)} DM_j = (n-1)\cdot\frac{1}{4}\,.$$

Онда, при услову да важи H_0, за ову тест статистику следи:

$$S_n \approx \mathcal{N}\left(\frac{n-1}{2}, \frac{n-1}{4}\right),$$

па је на стандардан начин можемо користити за тестирање хипотезе и налажење интервала поверења.

3.6 Тест Валда Волфовица

Овај тест је познат још као тест заснован на броју прелазака медијане. Тест је заснован на два принципа. Попут Фицовог теста користи се узорачка медијана у формирању тест статистике, али се такође користе и тзв. ранови (на енглеском: *run*), односно преласци. **Прелазак** дефинишемо као сваку промену у низу података. Наиме, ако имамо низ нула и јединица сваки пут кад се у низу појави нула иза јединице, или јединица иза нуле, то ћемо регистровати као један прелаз. Треба напоменути и да се почетак сваког низа рачуна као један прелаз. На пример, у низу 01010101 имамо 8 прелаза, док у низу 00001111 имамо 2 прелаза.

Дефинишимо следећи индикатор:

$$Y_j = \begin{cases} 1 & \text{, ако је } X_j > \text{медијане} \\ 0 & \text{, ако је } X_j < \text{медијане} \end{cases}$$

У низу $Y_1, ..., Y_n$ постоји n_1 нула и n_2 јединица, при чему важи $n_1+n_2 = n$. Означимо сада са R број прелазака у низу Y_j. Ова величина има следећу функцију расподеле:

$$F_R(r) = \begin{cases} \binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1} \Big/ \binom{n_1+n_2}{n_1} & \text{, ако је } r \text{ парно} \\ \left[\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}\right] \Big/ \binom{n_1+n_2}{n_1} & \text{, ако је } r \text{ непарно} \end{cases}$$

Када су и n_1 и n_2 велики R има асимптотски нормалну расподелу са математичким очекивањем:

$$ER = \frac{2n_1n_2}{n} + 1\,,$$

и дисперзијом:

$$DR = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}\,.$$

3.7 Голдов тест корака

Нека је дат узорак $X_1, ..., X_n$ и нека је $\tilde{X}$ узорачка медијана. Ако су $X_1, ..., X_{s_1}$ са исте стране $\tilde{X}$, а X_{s_1+1} са друге стране, онда дефинишемо **корак** дужине s_1. Ако су сада $X_{s_1+1}, ..., X_{s_1+s_2}$ са исте стране $\tilde{X}$, а $X_{s_1+s_2+1}$ са друге стране дефинишемо корак дужине s_2, и тако настављамо поступак док не дођемо до краја узорка. Означимо са s' корак највеће дужине.

Дефинишемо случајну величину $K(s)$ тако да број корака дужине s представља њену реализовану вредност. Математичко очекивање $K(s)$ при хипотези случајности ће бити:

$$EK(s) \approx \frac{n+3-s}{2^{s+1}}.$$

Сада можемо дефинисати и тест статистику:

$$Q = \sum_{s=1}^{s'} \frac{[K(s) - EK(s)]^2}{EK(s)}.$$

Ова тест статистика има хи-квадрат расподелу са $s' - 1$ степени слободе.

3.8 Тест серија

Тест серија је још један тест који је заснован на два принципа. У тесту се користи узорачка медијана за формирање тест статистике, али се осим ње користе и тзв. серије у низу података. **Серију** чине исте цифре, када се у низу налазе једна поред друге. На пример, у низу: 10011101 постоји 5 серија, док у низу 00001111 постоје само 2 серије.

Да бисмо од почетне временске серије $X_1, ..., X_n$ формирали низ нула и јединица дефинишемо следећи индикатор:

$$Y_j = \begin{cases} 1 & \text{, ако је } X_j > \text{медијане} \\ 0 & \text{, ако је } X_j < \text{медијане} \end{cases}$$

При претпоставци да је H_0 тачно важи:

$$Y_j : \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

па је:

$$EY_j = \frac{1}{2}, \; DY_j = \frac{1}{4}.$$

Када добијемо нови низ пребројавамо колико серија у низу постоји. Ту величину ћемо означити са K и она ће бити наша тест статистика.

За ову тест статистику при хипотези H_0 важи:

$$E(K) = \frac{n+2}{2}$$

$$D(K) = \frac{n(n-2)}{4(n-1)}\,.$$

Такође се може показати да ова тест статистика има приближно нормалну расподелу са математичким очекивањем $E(K)$ и дисперзијом $D(K)$ па се њене критичне вредности добијају из таблица за нормалну расподелу.

3.9 Тест разлика рангова

За почетак је потребно објаснити појам **ранга**. Нека је дат низ случајних величина $X_1, ..., X_n$. Када елементе овог низа поређамо у растућем поретку добијамо *варијациони низ*. Редни број елемента у варијационом низу представља његов ранг.

Нека су $r_1, r_2, ..., r_n$ рангови чланова непрекидне серије $X_1, X_2, ..., X_n$. Посматрајмо рангове $r_1, r_2, ..., r_n$ као реализације случајних величина $R_1, R_2, ..., R_n$. Ако посматрамо збир апсолутних разлика узастопних рангова у серији:

$$S_n = \sum_{j=1}^{n-1} |R_{j+1} - R_j|\ ,$$

онда је за велико n:

$$ES_n \approx \frac{(n-1)(n+1)}{3},\ DS_n \approx \frac{(n-2)(n+1)(4n-7)}{90}\,.$$

При хипотези H_0 важи:

$$S_n \approx \mathcal{N}\left(\frac{(n-1)(n+1)}{3}, \frac{(n-2)(n+1)(4n-7)}{90}\right)\,.$$

3.10 Фон Нојманов и Бартелсов тест

Нека је дата непрекидна временска серија $X_1, X_2, ..., X_n$ и одговарајући низ рангова $R_1, R_2, ..., R_n$. Дефинисана је Фон Нојманова статистика:

$$\xi_n = \frac{\sum_{j=1}^{n-1}(X_j - X_{j+1})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2},$$

где је $\overline{X} = \frac{X_1+...+X_n}{n}$. Када је $n < 20$ могу се наћи критичне вредности статистике ξ_n, а када је $n \geq 20$ може се користити нормална апроксимација одређена са:

$$Z_n = \frac{\xi_n - 2}{2}\sqrt{\frac{n^2-1}{n-2}} \approx \mathcal{N}(0,\, 1)\,.$$

Бартелсов тест представља модификацију Фон Нојмановог теста која користи рангове чланова дате серије. Бартелс 1982. уводи своју верзију тест статистике:

$$B_n = \frac{\sum_{j=1}^{n-1}(R_j - R_{j+1})^2}{\sum_{j=1}^{n}(R_j - \overline{R})^2} = \frac{A_n}{I_n},$$

где је $\overline{R} = \frac{R_1+...+R_n}{n} = \frac{n+1}{2}$.

Као и увек интересује нас расподела ове тест статистике. Како је:

$$I_n = \sum_{j=1}^{n}(R_j - \bar{R})^2 = \sum_{j=1}^{n}\left(j - \frac{n+1}{2}\right)^2 = \sum_{j=1}^{n} j^2 - \sum_{j=1}^{n} j\cdot(n+1) + \sum_{j=1}^{n}\left(\frac{n+1}{2}\right)^2 \Rightarrow$$

$$I_n = \frac{n(n+1)(2n+1)}{6} - (n+1)\frac{n(n+1)}{2} + n\frac{(n+1)^2}{4} = \frac{n(n^2-1)}{12} = const\,,$$

закључујемо да су A_n и B_n еквивалентне тест статистике, тј. да се особине једне од њих могу добити из особина оне друге.

Можемо видети да се вредности тест статистике A_n крећу од минимума који је једнак $m_A = n-1$ до максимума који за парно n износи $M_A = (n-1)(n^2+n-3)/3$, а за непарно $M_A - 1$. То можемо и доказати.

Како су сви R_i и R_{i+1} различити најмања њихова разлика је 1. Када се то квадрира поново добијамо 1. Сума ових квадрата је онда једнака $n-1$.

Што се тиче максимума, да бисмо добили максималну вредност потребно је да низ рангова буде распоређен на следећи начин:

- у средини се један поред другог налазе рангови n и 1
- лево од n је 2, а десно од 1 је $n-1$
- лево од 2 је $n-2$, а десно од $n-1$ је 3, итд.

На овај начин добијамо низ $...(n-2)2n1(n-1)3...$. Да је овај низ, баш онај којим се постиже максимум можемо проверити на примеру. Нека је дат низ дужине 4. Све могуће пермутације рангова су:

1234	1243	1324	1342	1423	1432
2134	2143	2314	2341	2413	2431
3124	3142	3214	3241	3412	3421
4123	4132	4213	4231	4312	4321

Проверавањем разлика рангова можемо видети да се максимум постиже у пермутацији 2413 јер су ту разлике 2, 3, 2. Онда је $A_n = 17$, што је управо

$$\frac{3 \cdot 17}{3} = \frac{(4-1)(4^2+4-3)}{3} = \frac{(n-1)(n^2+n-3)}{3} \, .$$

Слично се може видети да је за $n = 6$ најбоља пермутација 426153 када је $A_n = 65$, а за $n = 8$ то је 46281735, и тада је $A_n = 161$.

Из претходног закључка и чињенице да је $B_n = \frac{A_n}{I_n}$, а $I_n = \frac{n(n^2-1)}{12}$, видимо да ће минимум за тест статистику B_n бити $\frac{12}{(n+1)}$, а максимум $4 - \frac{12}{n(n+1)}$ за парно n, односно $4 - \frac{12}{n(n+1)} - \frac{12}{n(n^2-1)}$ за непарно n. Дакле вредности статистике B_n су у интервалу $(0, 4)$.

Бартелс је показао да за велико n важи:

$$EB_n \approx 2, \; DB_n \approx \frac{4}{n} \, ,$$

као и да су коефицијенти асиметрије и спљоштености:

$$\pi_1 \approx \frac{12}{35n\sqrt{n}}, \; \pi_2 \approx \frac{-798}{175n} \, .$$

Дакле оба коефицијента теже нули одакле закључујемо:

$$B_n \approx \mathcal{N}\left(2, \, \frac{4}{n}\right) \, ,$$

односно:

$$B_n^* = \frac{B_n - 2}{2}\sqrt{n} \approx \mathcal{N}(0, \, 1), \; n \geq 100.$$

3.11 Кендалов тест

Нека је дата непрекидна временска серија $X_1, ..., X_n$. Тестирање хипотезе случајности H_0 извршићемо упоређивањем чланова серије. Упоређивање се врши на следећи начин:

- налазимо колико је чланова у сегменту $X_2, X_3, ..., X_n$ веће од X_1,

- налазимо колико је чланова у сегменту $X_3, X_4, ..., X_n$ веће од X_2, итд,

- налазимо колико је чланова у сегменту X_{n-1}, X_n веће од X_{n-2},

- проверавамо да ли је X_n веће од X_{n-1}.

Све ово можемо формулисати математички. У том циљу дефинишемо индикатор I_{jk}, $j > k$ на следећи начин:

$$I_{jk} = \begin{cases} 1 & \text{, ако је } X_j > X_k \\ 0 & \text{, иначе} \end{cases}$$

За индикатор I_{jk} при хипотези H_0 важи:

$$I_{jk} : \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

па је:

$$EI_{jk} = \frac{1}{2},\ DI_{jk} = \frac{1}{4}.$$

Дефинишемо тест статистику која је позната као Кендалова τ статистика:

$$\tau_n = \frac{4}{n(n-1)} U_n - 1\,,$$

где је:

$$U_n = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} I_{jk}\,.$$

Расподела статистике U_n се не може одредити али се могу наћи њене нумеричке карактеристике. Из особина индикатора I_{jk} при нултој хипотези следи:

$$EU_n = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} EI_{jk} = \frac{n(n-1)}{4},\ DU_n = \frac{n(n-1)(2n+5)}{72}\,.$$

Одавде затим следи да је:

$$E\tau_n = \frac{4}{n(n-1)} \frac{n(n-1)}{4} - 1 = 0$$

$$D\tau_n = \frac{16}{n^2(n-1)^2} \frac{n(n-1)(2n+5)}{72} = \frac{2(2n+5)}{9n(n-1)}\,.$$

За тест статистику такође важи да за велике временске серије има приближно нормалну расподелу. Односно:

$$\tau_n \approx \mathcal{N}\left(0,\, \frac{2(2n+5)}{9n(n-1)}\right),$$

што нам дозвољава класичну употребу ове величине у тестирању нулте хипотезе.

Овај тест је ефикасан ако се за алтернативну хипотезу узме претпоставка да у серији постоји растући или опадајући тренд. Ово следи из чињенице да ако у сегментима $X_{k+1}, ..., X_n$ постоји доста чланова већих од X_k, онда то говори о постојању растућег тренда у серији. Можемо још приметити да ако $U_n = (n-1)+(n-2)+...+2+1 = \frac{n(n-1)}{2}$, односно ако је $\tau_n = 1$, онда је серија растућа. Ако је, пак, $U_n = 0$, односно $\tau_n = -1$, онда је серија опадајућа. Дакле у неким случајевима за конкретне вредности U_n и τ_n можемо одмах донети закључак о случајности и без даљег тестирања.

3.12 Серијска корелација и Дурбан Вотсонов тест

Нека су дате две временске серије

$$\begin{array}{ccccc} X_1, & X_2, & X_3, & ..., & X_n \\ Y_1, & Y_2, & Y_3, & ..., & Y_n \end{array}$$

Зависност међу члановима ових серија мери се њиховим *узорачким коефицијентом корелације* који се дефинише на следећи начин:

$$r_{XY} = \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)(Y_j - \overline{Y}_n)}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2\, \frac{1}{n}\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2}},$$

где је:

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \,, \overline{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j \,.$$

Посматрајмо сада серију X_n у односу на њу саму:

$$\begin{array}{cccccccc} & X_1, & X_2, & X_3, & ..., & & X_{n-1} & X_n \\ & \downarrow & \downarrow & & & & \downarrow & \\ X_1, & X_2, & X_3, & ..., & X_{n-1} & & X_n. & \end{array}$$

Ако узмемо парове по означеном правилу и применимо формулу за узорачки коефицијент корелације добијамо:

$$r_1' = \frac{\frac{1}{n-1}\sum_{j=1}^{n-1}(X_j - A_1)(X_{j+1} - B_1)}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n-1}(X_j - A_1)^2 \frac{1}{n-1}\sum_{j=1}^{n-1}(X_{j+1} - B_1)^2}},$$

где је:

$$A_1 = \frac{1}{n-1}\sum_{j=1}^{n-1} X_j,\; B_1 = \frac{1}{n-1}\sum_{j=1}^{n-1} X_{j+1}\,.$$

Овако уведен коефицијент r_1' назива се *серијски коефицијент корелације са кораком* 1.

У општем случају *серијски коефицијент корелације са кораком* k је:

$$r_k' = \frac{\frac{1}{n-k}\sum_{j=1}^{n-k}(X_j - A_k)(X_{j+1} - B_k)}{\sqrt{\frac{1}{n-k}\sum_{j=1}^{n-k}(X_j - A_k)^2 \frac{1}{n-k}\sum_{j=1}^{n-k}(X_{j+1} - B_k)^2}},$$

где је:

$$A_k = \frac{1}{n-k}\sum_{j=1}^{n-k} X_j,\; B_k = \frac{1}{n-k}\sum_{j=1}^{n-k} X_{j+1}\,.$$

За велике серије и мало k може се користити једноставнији израз:

$$r_k'' = \frac{\sum_{j=1}^{n-k}(X_j - \overline{X}_n)(X_{j+k} - \overline{X}_n)}{\sqrt{\sum_{j=1}^{n-k}(X_j - \overline{X}_n)^2 \sum_{j=1}^{n-k}(X_{j+k} - \overline{X}_n)^2}},$$

или још једноставније:

$$r_k = \frac{\sum_{j=1}^{n-k}(X_j - \overline{X}_n)(X_{j+k} - \overline{X}_n)}{\sum_{j=1}^{n-k}(X_j - \overline{X}_n)^2}\,.$$

Може се показати да је:

$$E(r_1) \approx \frac{-1}{n},\ D(r_1) \approx \frac{n^3 - 2n^2 + 2}{n^2(n^2 - 1)}\,,$$

па је при хипотези случајности за велико n расподела r_1 асимптотски $\mathcal{N}(0,\,1)$. За коефицијент корелације r_k, пак, важи:

$$r_k \approx \mathcal{N}\left(\frac{-1}{n-k-1},\ \frac{n-k-2}{(n-k-1)^2}\right)\,,$$

па и он за велико n и мало k има асимптотски $\mathcal{N}(0,\,1)$ расподелу.

Због њихових особина узорачки коефицијенти корелације $r_1, r_2, r_3, ...$ могу се користити за тестирање да ли у серији постоје јаке корелационе везе са кораком $1, 2, 3...$.

За проверу да ли у серији постоји аутокорелација са кораком 1 користи се Дурбан Вотсонов тест. У тесту се користи следећа статистика:

$$D = \frac{\sum_{j=1}^{n-1}(X_j - X_{j+1})^2}{\sum_{j=1}^{n} X_j^2}\,.$$

Ако реализацију статистике D означимо са d, а реализацију коефицијента корелације са кораком 1 са r, онда важи:

$$d \approx 2(1-r)\,,$$

одакле следи да је $0 \leq d \leq 4$. Па ако је $d \approx 2$ то значи да у серији нема аутокорелације, ако је $d \approx 4$ постоји негативна аутокорелација, а ако је $d \approx 0$ постоји позитивна аутокорелација.

4 Практична примена тестова случајности

4.1 Месечни индекс индустријске производње у Србији у периоду 1972-1989.

Демонстрираћемо примену описаних тестова случајности на конкретној временској серији. Серија је преузета из [4] и представља месечни индекс индустријске производње у Србији у периоду од 1972. до 1989. Сви подаци су дати у табели 4.1.1.

	Јан	Феб	Мар	Апр	Мај	Јун	Јул	Авг	Сеп	Окт	Нов	Дец
1972	38.5	41.8	46.6	44.7	46.3	48.0	39.6	39.5	47.3	49.9	48.8	50.7
1973	43.3	44.6	48.8	48.0	50.4	49.7	43.8	46.5	50.4	53.5	51.6	52.4
1974	47.1	48.8	55.0	54.5	55.4	55.9	48.3	50.8	55.5	60.0	56.8	62.4
1975	52.3	54.6	59.4	59.9	58.4	58.5	48.8	51.3	60.5	63.7	59.5	64.8
1976	52.9	56.6	62.1	58.2	58.5	62.1	52.0	56.7	63.2	65.8	64.6	71.6
1977	59.3	62.1	69.0	66.1	66.6	67.0	55.6	63.0	69.6	70.1	68.3	74.5
1978	63.2	67.0	75.9	70.7	72.7	72.7	58.6	68.4	76.0	78.5	75.4	79.6
1979	71.6	73.0	82.6	78.4	79.2	78.1	63.6	72.6	77.7	84.3	81.2	88.2
1980	74.7	80.9	85.3	83.1	81.1	80.3	63.6	73.7	82.5	93.4	88.7	91.2
1981	80.0	88.3	88.7	87.6	84.7	86.2	69.7	77.9	87.5	91.8	90.1	95.9
1982	79.7	81.6	92.6	87.9	83.3	85.8	67.5	79.1	88.0	91.3	86.4	93.7
1983	79.0	81.3	91.4	85.2	82.0	88.4	67.3	79.8	90.6	95.5	92.6	94.4
1984	82.4	83.5	95.0	90.2	87.1	93.3	76.6	84.3	94.4	104.4	100.3	101.5
1985	81.4	86.4	98.7	96.1	95.1	93.8	79.1	86.7	96.4	104.9	99.9	107.5
1986	93.2	94.0	102.8	98.0	95.8	98.2	80.3	86.3	98.4	106.1	104.2	106.0
1987	90.2	98.1	108.0	101.0	97.2	98.9	80.1	85.8	99.9	105.7	102.3	105.6
1988	91.6	100.1	108.9	99.5	97.1	100.0	81.3	88.5	101.6	105.5	100.5	111.2
1989	96.6	99.1	111.2	102.4	102.9	102.9	79.3	89.8	101.2	108.1	103.3	102.7

Табела 4.1.1: Месечни индекс индустријске производње у Србији

Податке ћемо прво нацртати како бисмо видели изглед серије. За цртање, као и осталу анализу и тестирања користићемо статистички програм *R*.

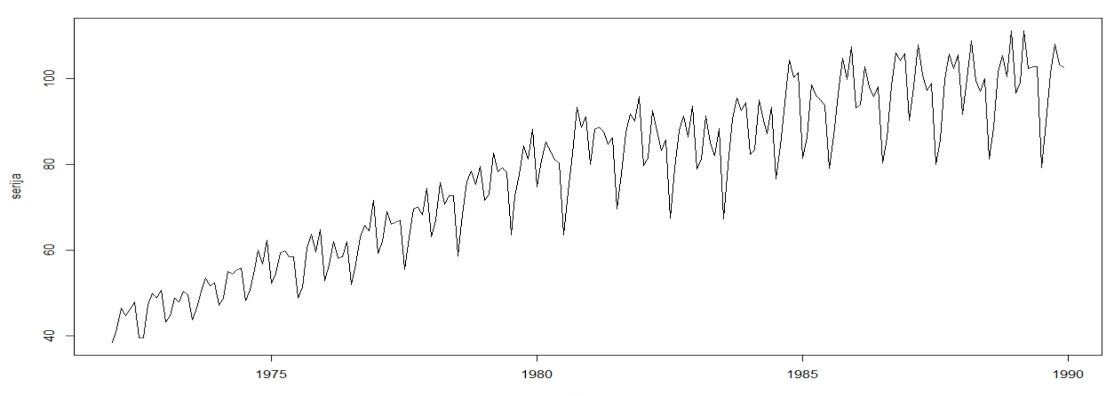

График 4.1.1: Месечни индекс индустријске производње у Србији

На графику 4.1.1 видимо да серија има растући тренд, као и то да постоје сезонске флуктуације. Међутим, утицајима ових компоненти на проверу случајности серије бавићемо се нешто касније. Сада ћемо применити тестове из претходног поглавља на серију у овом, непромењеном, облику.

Прво ћемо примениити тест тачака заокрета који је претходно испрограмиран у облику следеће функције:

```
testzaokreta<-function(X,alfa)
{
n<-length(X)
Zn<-0
for(j in 1:(n-2))
{
    Yj<-(X[j+1]-X[j])*(X[j+2]-X[j+1])
    if(Yj<0) Zn<-Zn+1
}

m<-(2*(n-2))/3
d<-(16*n-29)/90
cat("standardizovana vrednost test statistike je: ",((Zn-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))

cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je: (",-c,",",c,")\n")
}
```

У овој функцији рачунамо вредност тест статистике, а затим за њену стандардизовану вредност налазимо интервал поверења. Као аргументи функцији се прослеђују серија и ниво значајности `alfa`. Функција исписује вредност стандардизоване тест статистике и сам интервал поверења. Ако се вредност стандардизоване тест статистике налази у интервалу поверења онда прихватамо хипотезу случајности H_0, а у супротном хипотезу одбацујемо. Када тест применимо на нашу серију са нивоом значајности 0.05 добијамо следећи излаз:

```
standardizovana vrednost test statistike je:  -1.242427
95 %-tni interval poverenja za standardizovanu test statistiku je:
( -1.959964 , 1.959964 )
```

Дакле према овом тесту хипотезу H_0 би требало прихватити. Међутим, као што је раније поменуто, сваки тест се изводи са вероватноћом да дође до грешке па је у сваком тестирању пожељно користити што више тестова и тек онда донети коначни закључак.

Ми ћемо, како због демонстрације примене, тако и због доношења самог закључка о случајности, применити све раније поменуте тестове.

Следећи тест који ћемо применити је тест тачака раста. Користићемо следећу функцију:

```
testrasta<-function(X,alfa)
{
n<-length(X)
Rn<-0
for(j in 1:(n-1))
{
    Yj<-(X[j+1]-X[j])
    if(Yj>0) Rn<-Rn+1
}

m<-(n-1)/2
d<-(n+1)/12
cat("standardizovana vrednost test statistike je: ",((Rn-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))

cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je: (",-c,",",c,")\n")
}
```

Функција ради на истом принципу као и претходна и користи тврђење да тест статистика R_n има приближно нормалну расподелу. Ово тврђење можемо проверити у R-у. Генерисаћемо 1000 узорака обима 100 за које ћемо рачунати реализовану вредност статистике R_n, а затим ћемо добијене вредности тестирати на нормалност. За тестирање ћемо користити графичке методе, као и Жарк Бера тест.

У наставку ћемо генерисати потребне узорке и израчунати реализоване вредности тест статистике. За генерисање узорака користи се функција `sample()`.

```
stat<-c()
for(k in 1:1000){
    #izvlacimo prost slucajan uzorak od 100 jedinica
    n<-100
    uzorak<-sample(serija,n)
    #racunamo realizovanu vrednost test statistike
    Rn<-0
    for(j in 1:(n-1))
    {
        Yj<-(uzorak[j+1]-uzorak[j])
        if(Yj>0) Rn<-Rn+1
    }
    stat<-c(stat,Rn)}
```

Сада када смо добили низ података обима 1000 можемо проверити да ли он има нормалну расподелу. Најлакши начин провере је графичка провера. За то се користе хистограм и `q-q plot`, који исцртава податке у функцији теоретских квантила нормалне расподеле. За то користимо следећи код:

```
par(mfrow=c(1,2))
hist(stat,main="Histogram test statistike",
xlab="realizovana vrednost test statistike",ylab="frekvencija")

qqnorm(stat,main="Q-Q Plot normalnosti",xlab="teoretski kvantili",
ylab="uzoracki kvantili")
qqline(stat,col=2)
```

Добијамо график 4.1.2. Како хистограм има облик звона карактеристичан за нормалну расподелу он потврђује нашу претпоставку. Такође, како узорачки квантили прате линију теоретских квантила, и на основу `q-q plot`-а можемо закључити да тест статистика има нормалну расподелу.

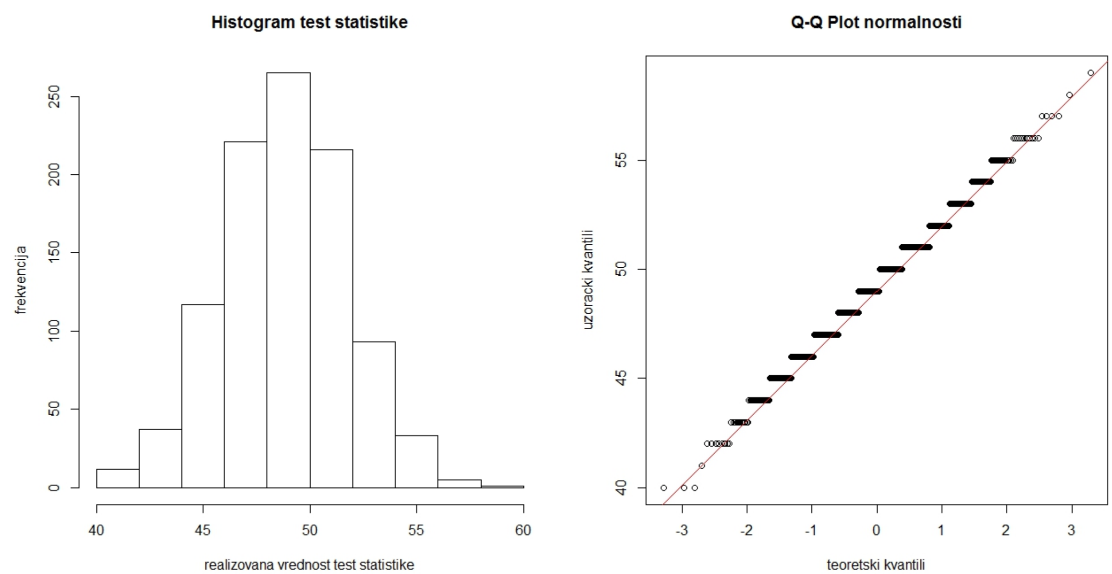

График 4.1.2: Графичка провера нормалности тест статистике

Осим графичке провере можемо користити и конкретнију проверу у виду Жарк Бера теста.

Жарк Бера тест је тест нормалности. Овим тестом испитује се, на основу анализе трећег и четвртог момента, колико емпиријска расподела дате серије одступа од нормалне расподеле. Дакле хипотезе овог теста су :

$$H_0 : \text{серија има нормалну расподелу,}$$

$$H_1 : \text{серија нема нормалну расподелу.}$$

Тест статистика се дефинише на основу тога што важи следеће:

$$z_3 = \sqrt{\frac{n}{6}} \frac{\hat{m}_3}{\hat{\sigma}^3} : \mathcal{N}(0,1)$$

$$z_4 = \sqrt{\frac{n}{24}} \left(\frac{\hat{m}_4}{\hat{\sigma}^4} - 3 \right) : \mathcal{N}(0,1),$$

где су $\hat{m}_3$ и $\hat{m}_4$ оцене за трећи и четврти централни моменат па је:

$$JB = z_3^2 + z_4^2 = \frac{n}{6} \left[\left(\frac{\hat{m}_3}{\hat{\sigma}^3} \right)^2 + \frac{\left(\frac{\hat{m}_4}{\hat{\sigma}^4} - 3 \right)^2}{4} \right] : \chi_2^2$$

Ова апроксимација важи асимптотски, што значи да се статистика JB понаша као случајна променљива са χ_2^2 расподелом само за узорке већег обима. Уколико је реализована вредност статистике мања од одговарајуће критичне вредности онда можемо сматрати да је серија нормално расподељена.

За Жарк Бера тест у R-у постоји уграђена функција `jarque.bera.test()`, која се налази у пакету `nortest`. Извршићемо тестирање низа `stat` који представља низ реализованих вредности статистике R_n помоћу следећег кода:

```
library(nortest)
jarque.bera.test(stat)

        Jarque Bera Test

data:  stat
X-squared = 0.3409, df = 2, p-value = 0.8433
```

Као и друге уграђене функције које представљају неке тестове и ова функција избацује p-вредност теста. Следећа дефиниција објашњава овај појам.

Дефиниција 4.1.1 *Нека је s реализована вредност тест статистике S. Најмањи ниво значајности на коме бисмо хипотезу H_0 одбацили при $S = s$ је p-вредност теста.*

Како смо добили велику p-вредност закључујемо да низ пролази овај тест и да тест статистика заиста има нормалну расподелу.

Пошто смо закључили да је тврђење о расподели тест статистике тачно можемо се вратити на тест тачака раста. Када извршимо тестирање серије добијамо следеће резултате:

```
standardizovana vrednost test statistike je:  4.585592
95 %-tni interval poverenja za standardizovanu test statistiku je:
( -1.959964 , 1.959964 )
```

Дакле овај тест показује да нулту хипотезу не треба прихватити, односно да серија није случајна. Ово такође потврђује нашу претпоставку да серија има растући тренд, јер је овај тест ефикасан када је алтернативна хипотеза управо да у серији постоји монотони тренд.

Сада ћемо искористити следећу функцију:

```
Ftest<-function(X,alfa)
{
n<-length(X)
Sn<-0
m<-median(X)
for(j in 1:n)
{
    Mj<-(X[j]-m)
    if(Mj>0) Sn<-Sn+1
}

m<-(n-1)/2
d<-(n-1)/4
cat("standardizovana vrednost test statistike je: ",((Sn-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je (",-c,",",c,")\n")
}
```

како бисмо помоћу Фицовог теста тестирали хипотезу случајности. Као што се може видети и ова функција ради исто што и претходне. За израчунавање узорачке медијане користи се уграђена функција `median()`. Када извршимо тестирање добијамо:

```
standardizovana vrednost test statistike je:  0.06819943
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Овај тест потврђује резултате теста тачака заокрета. Стандардизована тест статистика упада у интервал поверења па закључујемо да треба прихватити хипотезу H_0.

За следеће тестирање користићемо тест заснован на преласцима медијане. Користимо следећи код:

```
medijanatest<-function(X,alfa)
{
n<-length(X)
m<-median(X)
n1<-0
n2<-0
Yj<-c()
```

```
R<-1
for(j in 1:n)
{
    if(X[j]>=m)
    {
        pom<-1
        n1<-n1+1
    }
    else
    {
        pom<--1
        n2<-n2+1
    }
    Yj<-c(Yj,pom)
}

for(i in 1:(n-1))
{
    if(Yj[i]!=Yj[i+1]) R<-R+1
}

m<-(2*n1*n2)/(n1+n2)+1
d<-(2*n1*n2*(2*n1*n2-n))/((n1+n2)^2*(n1+n2-1))
cat("standardizovana vrednost test statistike je: ",((R-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je (",-c,",",c,")\n")
}
```

У овој функцији прво правимо низ 1 и -1 у зависности да ли је елемент серије већи или мањи од узорачке медијане, а затим у другој петљи проверавамо колико прелазака постоји у том новом низу. Функција затим рачуна стандардизовану вредност тест статистике и интервал поверења за њу. Када применимо тест на нашу серију за `alfa=0.05` добијамо:

```
standardizovana vrednost test statistike je:  -11.04843
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Међутим овај тест потврђује резултате теста тачака раста, који су супротни резултатима остала два теста. Односно према овом тесту хипотезу случајности треба одбацити.

Следећи тест који користимо је Голдов тест корака. Он је програмиран у облику следеће функције у R-у:

```
testkoraka<-function(X,alfa)
{
n<-length(X)
s<-c()
br<-1
m<-median(X)
y<-c()
for(j in 1:n)
{
    if(X[j]>=m) nova=1
    else nova=-1
    y<-c(y,nova)
}
for(i in 1:(n-1))
{
    if(y[i]==y[i+1]) br<-br+1
    else{
        s<-c(s,br)
        br<-1}
}
s<-c(s,br)
pom<-0
Ks<-c()
for(l in 1:max(s)){
    for(k in 1:length(s))
    {
        if(s[k]==l) pom<-pom+1
    }
    Ks<-c(Ks,pom)
    pom<-0
}
EKs<-c()
for(j in 1:max(s))
{
    pom<-(n+3-j)/2^(j+1)
    EKs<-c(EKs,pom)
}

Q<-sum((Ks-EKs)^2/EKs)
cat("vrednost test statistike je: ",Q,"\n")
c<-qchisq(1-alfa,max(s)-1)
cat("Kriticna oblast testa je (",c,",inf)\n")
}
```

У овој функцији прво правимо исти низ као у претходној, а затим рачунамо дужине корака. У дуплој петљи тражимо број корака дужине 1, а у последњој петљи рачунамо математичко очекивање $EK(s)$. Функција затим исписује вредност тест статистике, али, за разлику од претходних тестова, ова тест статистика нема нормалну расподелу па не правимо интервал поверења, већ исписујемо критичну област. Када функцију покренемо за нашу серију и ниво значајности 0.05 добијамо излаз:

```
vrednost test statistike je:  1.163477e+24
Kriticna oblast testa je ( 107.5217 ,inf)
```

Како вредност реализоване тест статистике не упада у критичну област закључујемо да нулту хипотезу треба прихватити.

Следећи тест који ћемо користити је тест серија. Тест је испрограмиран на следећи начин:

```
testserija<-function(X,alfa)
{
n<-length(X)
m<-median(X)
Y<-c()
for(j in 1:n)
{
    if(X[j]<m)
        Y[j]=0
    else if(X[j]>m)
        Y[j]=1
}

K<-0
for(j in 1:(length(Y)-1))
{
    if(Y[j]==Y[j+1])
        K<-K+1
}

m<-(n+2)/2
d<-(n*(n-2))/(4*(n-1))
cat("standardizovana vrednost test statistike je: ",((K-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je (",-c,",",c,")\n")
}
```

У тесту прво правимо низ нула и јединица, а затим у другој петљи бројимо серије. Када тест покренемо за дату серију и `alfa=0.05` добијамо следеће:

```
standardizovana vrednost test statistike je:  -11.18483
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Дакле, како вредност тест статистике не упада у интервал поверења закључујемо да нулту хипотезу треба одбацити, односно да серија није случајна.

Следеће тестирање извршићемо помоћу теста разлике рангова. За то ћемо користити следећи код:

```
testrangova<-function(X,alfa)
{
n<-length(X)
R<-rank(X)
Sn<-0
for(j in 1:(n-1))
{
    Sn<-Sn+abs(R[j+1]-R[j])
}

m<-((n-1)*(n+1))/3
d<-((n-2)*(n+1)*(4*n-7))/90
cat("standardizovana vrednost test statistike je: ",((Sn-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je (",-c,",",c,")\n")
}
```

Овај тест има једноставан код. За одређивање рангова чланова серије користи се уграђена функција `rank()`. Пошто тест статистика има нормалну расподелу враћамо се стандардном исписивању стандардизоване вредности тест статистике и њеном интервалу поверења. За `alfa=0.05` добијамо испис:

```
standardizovana vrednost test statistike je:  -16.53092
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Дакле још једном долазимо до закључка да нулту хипотезу треба одбацити.

Сада ћемо на серију применити Фон Нојманов и Бартелсов тест. Док смо први сами програмирали као функцију:

```
Ntest<-function(X,alfa)
{
n<-length(X)
a<-0
b<-0
for(j in 1:(n-1))
{
    pom1<-(X[j]-X[j+1])^2
    a<-a+pom1
}
for(j in 1:n)
{
    pom2<-(X[j]-mean(X))^2
    b<-b+pom2
}

Sn<-a/b
Zn<-((Sn-2)/2)*sqrt((n^2-1)/(n-2))
cat("vrednost test statistike je: ",Zn,"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za test
statistiku je (",-c,",",c,")\n")
}
```

за други у пакету `lawstat` постоји уграђена функција `bartels.test()` која за прослеђену серију даје p-вредност теста. Вршимо тестирање помоћу ова два теста:

```
Ntest(serija,0.05)

vrednost test statistike je:  -13.49557
95 %-tni interval poverenja za test statistiku je ( -1.959964 , 1.959964 )

bartels.test(serija)

        Bartels Test - Two sided

data:  serija
Standardized Bartels Statistic = -13.2367, RVN Ratio = 0.199,
p-value <2.2e-16
```

У првом тест стандардизована вредност тест статистике не упада у интервал поверења, па закључујемо да хипотезу треба одбацити. У другом тесту је добијена изузетно мала p-вредност теста, па долазимо до истог закључка.

Следећи тест који ћемо користити је Кендалов тест:

```
Ktest<-function(X,alfa)
{
n<-length(X)
Un<-0
for(k in 1:(n-1))
{
    for(j in (k+1):n)
    if(X[j]>X[k]) Un<-Un+1
}
Tn<-(4*Un)/(n*(n-1))-1

m<-0
d<-(2*(2*n+5))/(9*n*(n-1))
cat("standardizovana vrednost test statistike je: ",((Tn-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je (",-c,",",c,")\n")
}
```

За `alfa=0.05` добијамо:

```
standardizovana vrednost test statistike je:  16.84844
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Дакле и овај тест нам говори да треба одбацити хипотезу случајности.

Остаје нам још тест корелације и Дурбан Вотсонов тест, који ћемо овом приликом посматрати као тест корелације за $k = 1$. Наиме, за оба теста ћемо користити једну функцију, која поред уобичајених аргумената, серије и нивоа значајности, има још и вредност корака корелације k. Функција је облика:

```
testkorelacije<-function(X,k,alfa)
{
n<-length(X)
a<-0
b<-0
for(j in 1:(n-k))
{
    a<-a+(X[j]-mean(X))*(X[j+k]-mean(X))
    b<-b+(X[j]-mean(X))^2
}

rk<-a/b
m<--1/(n-k-1)
d<-(n-k-2)/(n-k-1)^2
```

```
cat("standardizovana vrednost test statistike je: ",((rk-m)/sqrt(d)),"\n")
c<-qnorm(1-(alfa/2))
cat((1-alfa)*100,"%-tni interval poverenja za standardizovanu test
statistiku je (",-c,",",c,")\n")
}
```

Тест ћемо покренути два пута за вредности $k = 1$ и $k = 12$. Добијамо:

```
> testkorelacije(serija,1,0.05)
standardizovana vrednost test statistike je:  13.3749
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkorelacije(serija,12,0.05)
standardizovana vrednost test statistike je:  13.33483
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Дакле у оба случаја закључујемо да корелација постоји и да треба одбацити нулту хипотезу. Овим смо показали да постоји корелација између суседних месеци, као и корелација између истих месеци различитих година.

Да резимирамо. У циљу тестирања случајности серије која представља месечни индекс индустријске производње у Србији у периоду од 1972. до 1989. извршили смо 12 тестова. Као резултат 9 тестова добили смо да хипотезу случајности треба одбацити, док смо код 3 теста дошли до закључка да исту хипотезу треба прихватити. Већина ових тестова има своју слабу тачку, у смислу да за те тестове можемо направити низ који ће бити детерминистички, а за који ће тест ипак показати да је случајан. Управо због ове особине тестова пожељно је извршити више тестирања што смо ми и урадили. Узевши у обзир резултате и мане тестова можемо, са великом вероватноћом да смо у праву, закључити да наша серија нема особину случајности.

Како бисмо истраживање проширили из серије ћемо одстранити тренд и сезонску компоненту, које су саме по себи неслучајне, и исте ове тестове применити на новонасталу серију.

Пре свега можемо проверити да ли су наше претпоставке тачне и да ли ова серија заиста има тренд. За то можемо користити Ман Кендалов тест. Нулта хипотеза овог теста је да дата временска серија x нема тренд, насупрот хипотези да тренд постоји. Ман Кендалова статистика која се користи за добијање тест статистике је:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_j - x_i),$$

где је

$$sgn(x_j - x_i) = \begin{cases} -1, & (x_j - x_i) < 0 \\ 0, & (x_j - x_i) = 0 \\ 1, & (x_j - x_i) > 0 \end{cases}$$

Ова статистика за $n > 8$ има приближно нормалну расподелу са:

$$E(S) = 0,\ D(S) = \frac{n(n-1)(2n+5)}{18}$$

Из ове статистике се даље формира тест статистика:

$$Z_c = \begin{cases} \frac{S-1}{\sqrt{D(S)}}, & S < 0 \\ 0, & S = 0 \\ \frac{S+1}{\sqrt{D(S)}}, & S > 0 \end{cases}$$

Z_c има стандардну нормалну расподелу. Позитивна вредност тест статистике указује на растући, а негативна вредност на опадајући тренд.

У пакету `Kendall` постоји функција `MannKendall` која обавља тест уместо нас. Када тест применимо на нашу серију добијамо:

```
MannKendall(serija)
tau = 0.772, 2-sided pvalue =< 2.22e-16
```

Због изузетно мале p-вредности можемо закључити да се почетна хипотеза одбацује, односно да тренд постоји. Како је тест статистика, у излазу теста дата као `tau`, позитивна можемо још закључити и да је тренд растући.

Сада када смо сигурни да тренд постоји прећићемо на његово уклањање из серије. Помоћу уграђене функције `decompose()` серија се може раздвојити на компоненте. Када на декомпоновану серију применимо функцију `plot` добијамо график 4.1.3.

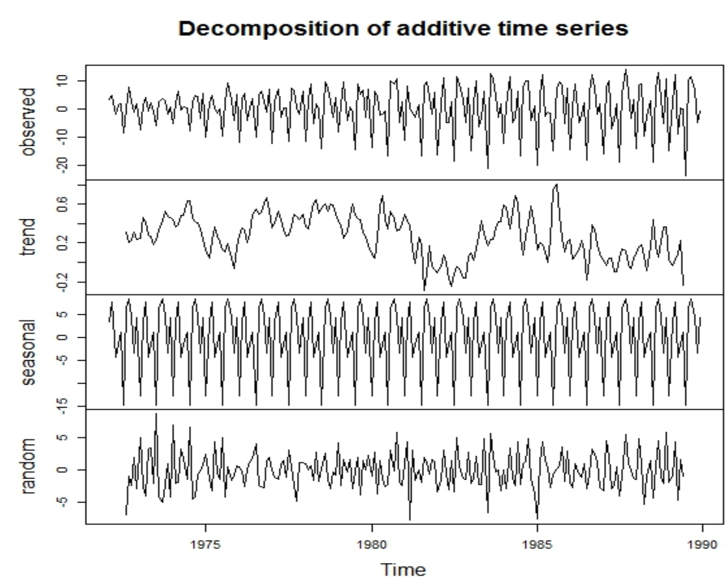

График 4.1.3: Серија раздвојена на своје компоненте

На графику се јасно види изглед тренда (други график) и сезонске компоненте (трећи график), као и како серија изгледа када се ове компоненте уклоне (четврти график). Серију без неслучајних компонената добијамо наредбом

```
serija2<-coredata(decompose(serija)$random)
```

На ову серију ћемо затим применити све претходно описане тестове како бисмо видели да ли се нешто изменило. Применом тестова добијамо следећи испис:

```
> testzaokreta(serija2,0.05)
standardizovana vrednost test statistike je:  -2.112742
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testrasta(serija2,0.05)
standardizovana vrednost test statistike je:  -1.088745
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> Ftest(serija2,0.05)
standardizovana vrednost test statistike je:  0.07018624
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> medijanatest(serija2,0.05)
standardizovana vrednost test statistike je:  -2.807484
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkoraka(serija2,0.05)
vrednost test statistike je:  23.12133
Kriticna oblast testa je ( 14.06714 ,inf)

> testserija(serija2,0.05)
standardizovana vrednost test statistike je:  -2.947858
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testrangova(serija2,0.05)
standardizovana vrednost test statistike je:  -3.221842
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

```
> Ntest(serija2,0.05)
vrednost test statistike je:  -4.3284
95 %-tni interval poverenja za test statistiku je
( -1.959964 , 1.959964 )

> bartels.test(serija2)
        Bartels Test - Two sided

data:  serija2
Standardized Bartels Statistic = -3.9159, RVN Ratio = 1.452,
p-value = 9.008e-05

> Ktest(serija2,0.05)
standardizovana vrednost test statistike je:  1.202336
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkorelacije(serija2,1,0.05)
standardizovana vrednost test statistike je:  4.035916
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkorelacije(serija2,12,0.05)
standardizovana vrednost test statistike je:  8.337033
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Као што видимо постоје неке разлике. У прва два теста добијамо резултате супротне оним које смо добили код прве серије. Како смо уклонили тренд из серије резултат другог теста, да хипотезу треба прихватити, је донекле очекиван, јер је алтернативна хипотеза овог теста да постоји тренд. Резултати се разликују још код Кендаловог теста и Голдовог теста корака. Узрок различитог резултата код Кендаловог теста је поново алтернативна хипотеза која је иста као код теста тачака раста.

Закључак до кога долазимо је да, како 9 од 12 тестова показује да хипотезу треба одбацити, ни ова серија није случајна. Дакле, иако смо избацили компоненте за које знамо да нису случајне, серија и даље нема особину случајности.

Следећи корак који бисмо могли да предузмемо како бисмо добили случајну серију је да нађемо одговарајући модел у који бисмо уклопили почетну серију. Да бисмо дошли до модела могуће је да ће пре тога бити потребне још неке трансформације, попут стабилизације варијансе. Уколико је варијанса промењливе пропорционална њеној очекиваној вредности погодна трансформација јесте логаритамска. Након избора модела бисмо оценили параметре модела, а затим оцењену случајну компоненту тестирали на случајност.

4.2 Дневна депрецијација девизног курса евра према долару у периоду 2010-2012.

Посматраћемо дневну депрецијацију девизног курса евра према долару у периоду од 1. јануара 2010. до 31. децембра 2012. Дневна депрецијација је пад цена неке валуте у односу на другу валуту или у односу на све друге валуте. Податке о средњем курсу евра и америчког долара у односу на динар за дати период преузели смо са интернет презентације Народне банке Србије (*http : //www.nbs.rs/export/sites/default/internet/cirilica/scripts/klperiod.html*).

Дневна депрецијација добија се као прва диференца логаритмованог нивоа девизног курса. Диференцирање серије представља начин за добијање стационарне од нестационарне серије. Процеси код којих се стационарност постиже диференцирањем се називају диференцно стационарни процеси. У *R*-у постоји уграђена функција `diff()` која врши диференцирање.

За разлику од серије коју смо користили у претходној анализи ова серија је стационарна и нема ни тренд, ни сезонску компоненту. Серија се у *R*-у може добити на следећи начин:

```
dolar<-read.table("dolar.txt")
euro<-read.table("euro.txt")
razlika<-ts(euro$V2-dolar$V2)
razlika.l<-log(razlika)
razlika.d<-diff(razlika.l)
```

Њен изглед се може видети на графику 4.2.1.

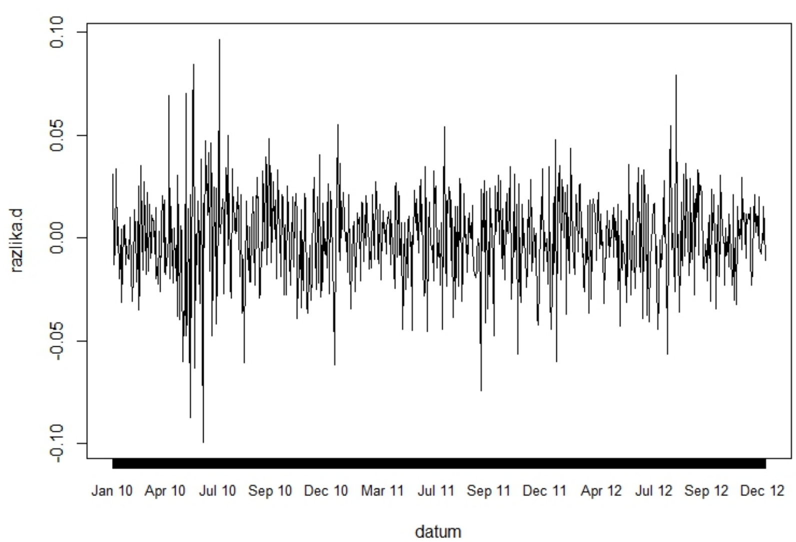

График 4.2.1: Дневна депрецијација девизног курса евра према долару у периоду од 1. јануара 2010. до 31. децембра 2012.

Применићемо сада на ову серију све тестове. Добијамо следеће:

```
> testzaokreta(razlika.d,0.05)
standardizovana vrednost test statistike je:  -0.4022225
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testrasta(razlika.d,0.05)
standardizovana vrednost test statistike je:  -1.005249
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> Ftest(razlika.d,0.05)
standardizovana vrednost test statistike je:  0
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> medijanatest(razlika.d,0.05)
standardizovana vrednost test statistike je:  -0.4721346
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkoraka(razlika.d,0.05)
vrednost test statistike je:  10.98981
Kriticna oblast testa je ( 16.91898 ,inf)

> testserija(razlika.d,0.05)
standardizovana vrednost test statistike je:  -0.6901117
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testrangova(razlika.d,0.05)
standardizovana vrednost test statistike je:  -0.225363
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> Ntest(razlika.d,0.05)
vrednost test statistike je:  -0.3955503
95 %-tni interval poverenja za test statistiku je
( -1.959964 , 1.959964 )
```

```
> bartels.test(razlika.d)

        Bartels Test - Two sided

data:  razlika.d
Standardized Bartels Statistic = -0.1687, RVN Ratio = 1.988, p-value =
0.866

> Ktest(razlika.d,0.05)
standardizovana vrednost test statistike je:  0.3951551
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkorelacije(razlika.d,1,0.05)
standardizovana vrednost test statistike je:  0.3902899
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )

> testkorelacije(razlika.d,12,0.05)
standardizovana vrednost test statistike je:  0.7063327
95 %-tni interval poverenja za standardizovanu test statistiku je
( -1.959964 , 1.959964 )
```

Као што видимо свих 12 тестова показује да нулту хипотезу треба прихватити. Одавде закључујемо да је ова серија заиста случајна. Овакав исход је био и за очекивати, јер серија нема тренд ни сезонску компоненту, али и из других разлога. Наиме, временска серија реални девизни курс треба да осцилира правилно током времена да би теорија о паритету куповне снаге у анализи девизног курса (по којој скуп добара треба да кошта приближно исто у различитим економијама) била валидна. Обратно, одступање девизног курса од препознатљиве путање, у смислу нестационарности временске серије, значи да ова теорија није прихватљива за дату економију у изабраном периоду.

5 Закључак

Временске серије су распрострањене у великом броју различитих области живота. Управо због тога расте значај њиховог истраживања.

Први задатак који се намеће при анализи временских серија јесте управо тестирање случајности. Као што је већ речено, под појмом случајности подразумевамо да се ради о реализацији низа случајних величина $X_1, ..., X_n$ које су међусобно независне и све имају исту расподелу вероватноћа. У овом раду је представљено 12 тестова случајности од којих су неки засновани на међусобном односу чланова серије, на ранговима, а неки могу користити узорачке медијане или пак преласке (кораке). Исправност нулте хипотезе често може да значи да би се тада конкретни подаци могли узимати у произвољном поретку при даљем изучавању тј. да се могу користити методе математичке статистике. Ако нулта хипотеза не важи, онда се мора поштовати редослед података и за даља истраживања можемо користити грану статистике која се назива анализа временских серија.

Тестови објашњени у раду су обрађени у програмском пакету R и примењени су на две временске серије, од којих се једна односи на *Месечни индекс индустријске производње у Србији у периоду од 1972-1989. године*, а друга се односи на *Дневну депрецијацију девизног курса евра према долару у периоду од 2010-2012. године.* У првом примеру нисмо прихватили хипотезу о случајности временске серије, док у другом примеру можемо на основу тестирања закључити да је временска серија случајна. Уколико временска серија не прође тестове случајности, као што је случај са првом серијом у овом раду, даља истраживања извршавамо применом анализе временских серија. Уколико, пак, временска серија прође тестове случајности, као што је случај са другом серијом за даља истраживања можемо користити и неке статистичке методе.

Литература

[1] Ј. Малишић и В. Јевремовић: *Случајни процеси и временске серије*, Математички факултет, Београд, 2008.

[2] З. Младеновић и А. Нојковић: *Примењена анализа временских серија*, Центар за издавачку делатност Економског факултета у Београду

[3] Павле Младеновић: *Вероватноћа и статистика*, Математички факултет, Београд, 2008.

[4] Златко Ковачић: *Анализа временских серија*, Економски факултет, Београд, 1995.

[5] Јован Малишић: *Временске серије*, Математички факултет, Београд, 2002.

[6] В. Јевремовић и Ј. Малишић: *Статистичке методе у метеорологији и инжењерству*, Београд, 2002.

[7] `http://a-little-book-of-r-for-time-series.readthedocs.org/-en/latest/src/timeseries.html`

[8] `http://cran.r-project.org/`

[9] Ying Wang: Nonparametric tests for randomness, 2003.

Биографија

Рођена сам 17.7.1989. године у Крушевцу. Тамо сам завршила основну школу, као и средњу Економско-трговинску школу.

Уписала сам Математички факултет, смер статистика, актуарска и финансијска математика, у Београду 2008. године. Дипломирала сам 2012. године са просечном оценом 8.02.

Мастер студије уписала сам октобра 2012. године на Математичком факултету у Београду.