VILNIAUS UNIVERSITETO
MATEMATIKOS IR INFORMATIKOS FAKULTETAS

**Vilijandas Bagdonavičius**

**Julius Jonas Kruopis**

# MATEMATINĖ STATISTIKA

*Vadovėlis*

## III DALIS

## NEPARAMETRINĖ STATISTIKA

Vilniaus universiteto leidykla
2015

Apsvarstė ir rekomendavo spausdinti Vilniaus universiteto Matematikos ir informatikos fakulteto taryba (2015 m. vasario 17 d.; protokolas Nr 3); vadovėlio statusą suteikė Vilniaus universiteto Senatas (2015 m. balandžio 21 d. nutarimas Nr. S – 2015 – 4 –12).

Recenzavo:
prof. habil. dr. Algimantas Bikelis (Vytauto Didžiojo universitetas),
prof. habil. dr. Kęstutis Dučinskas (Klaipėdos universitetas)

ISBN 978–609–459–517–2

# Turinys

# Pratarmė

Ši vadovėlio dalis skiriama neparametrinių modelių hipotezių tikrinimo uždaviniams spręsti. Statistinis modelis vadinamas neparametriniu, jeigu imties skirstinys negali būti nusakytas naudojant baigtinės dimensijos parametrą. Konstruojami kriterijai suderinamumo, nepriklausomumo, atsitiktinumo, homogeniškumo hipotezėms tikrinti. Šioje vadovėlio dalyje apsiribojama pilnomis imtimis. Didesnė dalis pateikiamos medžiagos išspausdinta anglų kalba [2]. Analogiškų uždavinių sprendimas cenzūruotų imčių atveju nagrinėjamas knygoje [3]. Norintiems plačiau studijuoti hipotezių tikrinimo kriterijų sudarymo metodus neparametriniuose pilnų imčių modeliuose rekomenduojame monografijas [8], [13], [14], [15].

Pirmame skyriuje pateikiamos pagrindinės sąvokos, susijusios su kriterijų sudarymu ir jų palyginimu, ir suformuluotos dažniausiai tikrinamos neparametrinės hipotezės.

Sprendžiant kiekvieną matematinės statistikos uždavinį pirmiausia yra parenkamas statistinis modelis, kurio rėmuose bus analizuojami turimi duomenys. Nuo tinkamo modelio parinkimo daug priklauso gaunamų išvadų korektiškumas. Jeigu parenkant statistinį modelį nepakanka turimos apriorinės informacijos, šiam tikslui galima pasitelkti suderinamumo kriterijus, kuriais tikrinamos prielaidos, kad turimi duomenys suderinami su vienu ar kitu tikimybiniu modeliu.

Viena iš pagrindinių suderinamumo hipotezių tikrinimo kriterijų klasių yra chi kvadrato tipo kriterijai, kurie sudaromi remiantis sugrupuota į tam tikrus intervalus imtimi. Reikia pasakyti, kad daugelyje matematinės statistikos knygų ir programų paketų šie kriterijai taikomi nekorektiškai. Teoriniai rezultatai, kuriais grindžiami chi kvadrato kriterijai, kai suderinamumo hipotezės sudėtinės, yra gauti tariant, kad grupavimo intervalai nepriklauso nuo imties, o parametrų įvertiniai gauti naudojant grupuotąją imtį. Dažnai praktiškai taikant kriterijus abi šios sąlygos yra pažeidžiamos: grupavimo intervalų galai priklauso nuo imties, o parametrai vertinami pagal pradinius negrupuotus duomenis. Antrame skyriuje pateikiamas modifikuotasis chi kvadrato kriterijus neturi minėtų trūkumų. Teoriniai rezultatai apie naudojamos statistikos skirstinį gauti tariant, kad parametrai vertinami pagal pradinius negrupuotus duomenis, o grupavimo intervalų galai priklauso nuo imties.

Trečiame skyriuje nagrinėjami vadinamieji glodūs Neimano ir Bartono tipo kriterijai, kurių statistikos sudaromos naudojant pradinius negrupuotus duomenis.

Trečia suderinamumo hipotezių tikrinimo kriterijų klasė grindžiama funkcionalais nuo teorinės ir empirinės pasiskirstymo funkcijų skirtumo (Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo kriterijai). Reikia pažymėti, kad taikant šiuos kriterijus sudėtinei suderinamumo hipotezei tikrinti kartais naudojami teoriniai rezultatai, kurie gauti paprastosios hipotezės atveju.

Tikrinant sudėtines suderinamumo hipotezes reikėtų naudoti ketvirtame skyriuje pateikiamus modifikuotuosius kriterijus. Kai kuriuose programų paketuose į šią aplinkybę yra atsižvelgiama (pvz., SAS programų paketas).

Keletas specializuotų kriterijų hipotezėms dėl imties skirstinio priklausymo dažniausiai taikomų skirstinių (normaliojo, eksponentinio, Veibulo, Puasono) šeimoms tikrinti pateikiama 6.5 skyrelyje.

Kriterijus dėl dviejų ar daugiau priklausomų ar nepriklausomų imčių tikimybinių skirstinių sutapimo (homogeniškumo hipotezė) galima rasti 2–6 skyriuose. 2.5 skyrelyje pateiktas chi kvadrato kriterijus; 4.5 skyrelyje – Kolmogorovo ir Smirnovo bei Kramero ir Mizeso dviejų imčių kriterijai. Tikrinant homogeniškumo hipotezes daugeliu atvejų pasirodo efektyvesni ranginiai kriterijai, jie pateikiami 5.5–5.9 ir 6.3–6.4 skyreliuose.

Nepriklausomumo hipotezei tikrinti pateikiame chi kvadrato kriterijų (2.4 skyrelis), Spirmeno ir Kendalo ranginius kriterijus (5.3, 5.10 skyreliai). Ranginiai atsitikinumo hipotezės tikrinimo kriterijai pateikiami 5.4 skyrelyje.

Pateikdami kiekvieną kriterijų stengėmės prisilaikyti tokių etapų: 1) hipotezės ir alternatyvos formulavimas; 2) kriterijaus konstravimo idėjos aptarimas; 3) kriterijaus statistikos apibrėžimas; 4) statistikos tikslaus ar asimptotinio skirstinio radimas; 5) kriterijaus ir jo modifikacijų (tolydumo pataisa, sutampantys duomenys) formulavimas; 6) kriterijaus taikymo iliustravimas konkrečiu pavyzdžiu; 7) gauto rezultato interpretavimas.

Šioje vadovėlio dalyje pateikiama medžiaga apima standartinį matematinės pakraipos studentų vieno semestro kursą. Pateikiamai medžiagai suprasti studentai turėtų būti išklausę universitetinių programų apimties bendruosius matematikos kursus, tikimybių teorijos kursą ir parametrinės matematinės statistikos kursą (šio vadovėlio I dalis). Pagrindinius naudojamus tikimybių teorijos ir parametrinės matematinės statistikos faktus pateikiame A ir B prieduose (7 ir 8 skyriai).

Medžiaga suskaidyta į 6 skyrius ir smulkesnius skyrelius. Kiekviename skyrelyje teoremos, apibrėžimai, pavyzdžiai, formulės numeruojamos trimis indeksais: skyriaus, skyrelio ir eilės numeriu skyrelyje.

# Trumpiniai ir žymenys

A. d. — atsitiktinis dydis
n. a. d. — nepriklausomi atsitiktiniai dydžiai
a. v. — atsitiktinis vektorius
n. a. v. — nepriklausomi atsitiktiniai vektoriai
TG — tolygiai galingiausias (kriterijus)
TGN — tolygiai galingiausias nepaslinktasis (kriterijus)
DT — didžiausiojo tikėtinumo (funkcija, metodas, įvertinys)
MK — mažiausiųjų kvadratų (metodas, įvertinys)
ASE — asimptotinis santykinis efektyvumas (įvertinių, kriterijų)
TPP — taikomieji programų paketai
$X,\ Y,\ Z, \ldots$ — atsitiktiniai dydžiai
$\mathbf{X},\ \boldsymbol{Y},\ \boldsymbol{Z}, \ldots$ — atsitiktiniai vektoriai
$\mathbf{X}^T$ — transponuotas vektorius, t. y. vektorius – eilutė
$x(P)$ — $P$-asis kvantilis
$x_P$ — $P$-oji kritinė reikšmė
$\mathbf{P}\{A\}$ — įvykio $A$ tikimybė
$\mathbf{P}\{A|B\}$ — įvykio $A$ sąlyginė tikimybė
$\mathbf{P}_\theta\{A\},\ \mathbf{P}\{A|\theta\}$ — tikimybė, priklausanti nuo parametro $\theta$
$F_\theta(x),\ F(x;\ \theta),\ F(x|\theta)$ — pasiskirstymo funkcija, priklausanti nuo parametro $\theta$ (analogiškai tankio funkcijai)
$\mathbf{E}X$ — a. d. $X$ vidurkis
$\boldsymbol{V}X$ — a. d. $X$ dispersija
$\mathbf{E}_\theta(X),\ \mathbf{E}(X|\theta),\ \boldsymbol{V}_\theta(X),\ \boldsymbol{V}(X|\theta)$ — a. d. $X$ vidurkis ir dispersija, priklausantys nuo parametro $\theta$
$\mathbf{E}(\mathbf{X})$ — a. v. $\mathbf{X}$ vidurkių vektorius
$\boldsymbol{V}(\mathbf{X})$ — a. v. $\mathbf{X}$ kovariacijų matrica
$\mathbf{Cov}\,(X,\ Y)$ — a. d. $X$ ir $Y$ kovariacija
$\mathbf{Cov}\,(\mathbf{X},\ \boldsymbol{Y})$ — a. v. $\mathbf{X}$ ir $\boldsymbol{Y}$ kovariacijų matrica
$B(n,\ p)$ — binominis skirstinys su parametrais $n$ ir $p$
$\mathcal{P}(\lambda)$ — Puasono skirstinys su parametru $\lambda$;
$N(0,\ 1)$ — standartinis normalusis skirstinys
$N(\mu,\ \sigma^2)$ — normalusis skirstinys su parametrais $\mu$ ir $\sigma^2$

$LN(\mu,\ \sigma)$ — lognormalusis skirstinys su parametrais $\mu$ ir $\sigma$
$\mathcal{E}(\lambda)$ — eksponentinis skirstinys su parametru $\lambda$
$G(\lambda,\ \eta)$ — gama skirstinys su parametrais $\lambda$ ir $\eta$
$W(\theta,\ \nu)$ — Veibulo skirstinys su parametrais $\theta$ ir $\nu$
$AW(\theta,\ \nu,\ \gamma)$ — apibendrintasis Veibulo skirstinys su parametrais $\theta$, $\nu$ ir $\gamma$
$U(\alpha,\ \beta)$ — tolygusis skirstinys intervale $(\alpha,\ \beta)$
$\chi^2(n)$ — chi kvadrato skirstinys su $n$ laisvės laipsnių
$\chi^2(n;\ \delta)$ — necentrinis chi kvadrato skirstinys su $n$ laisvės laipsnių ir necentriškumo parametru $\delta$
$S(n)$ — Stjudento skirstinys su $n$ laisvės laipsnių
$S(n;\ \delta)$ — necentrinis Stjudento skirstinys su $n$ laisvės laipsnių ir necentriškumo parametru $\delta$
$F(m,\ n)$ — Fišerio skirstinys su $m$ ir $n$ laisvės laipsnių
$F(m,\ n;\ \delta)$ — necentrinis Fišerio skirstinys su $m$ ir $n$ laisvės laipsnių ir necentriškumo parametru $\delta$
$z_\alpha$ — standartinio normaliojo skirstinio $\alpha$ kritinė reikšmė
$t_\alpha(n)$ — Stjudento skirstinio su $n$ laisvės laipsnių $\alpha$ kritinė reikšmė
$\chi^2_\alpha(n)$ — chi kvadrato skirstinio su $n$ laisvės laipsnių $\alpha$ kritinė reikšmė
$F_\alpha(\mathrm{m, n})$ — Fišerio skirstinio su $m$ ir $n$ laisvės laipsnių $\alpha$ kritinė reikšmė
$\mathcal{P}_k(n,\ \boldsymbol{\pi})$ — $k$-matis polinominis skirstinys su parametrais $n$ ir $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)^T$, $\pi_1 + ... + \pi_k = 1$
$N_k(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ — $k$-matis normalusis skirstinys su vidurkių vektoriumi $\boldsymbol{\mu}$ ir kovariacijų matrica $\boldsymbol{\Sigma}$
$X \sim N(\mu,\ \sigma^2)$ — a. d. $X$ pasiskirstęs pagal normalųjį dėsnį su parametrais $\mu$ ir $\sigma^2$ (analogiškai kitų skirstinių atveju)
$X_n \xrightarrow{P} X$ — konvergavimas pagal tikimybę $(n \to \infty)$
$X_n \xrightarrow{b.t.} X$ — konvergavimas su tikimybe 1 arba beveik tikrai $(n \to \infty)$
$X_n \xrightarrow{d} X$, $F_n(x) \xrightarrow{d} F(x)$ — konvergavimas pagal pasiskirstymą (silpnasis; $n \to \infty$)
$X_n \xrightarrow{d} X \sim N(\mu,\ \sigma^2)$ — a. d. $X_n$ asimptotiškai $(n \to \infty)$ turi normalųjį skirstinį su parametrais $\mu$ ir $\sigma^2$;
$X_n \sim Y_n$ — a. d. $X_n$ ir $Y_n$ asimptotiškai $(n \to \infty)$ ekvivalentūs $(X_n - Y_n \xrightarrow{P} 0)$
$||\boldsymbol{x}||$ — kai $\boldsymbol{x} = (x_1, ..., x_k)^T$ yra vektorius, reiškia atstumą $(\boldsymbol{x}^T\boldsymbol{x})^{1/2} = (\sum_i x_i^2)^{1/2}$
$||\boldsymbol{A}||$ — kai $\boldsymbol{A} = [a_{ij}]$ yra matrica, reiškia $(\sum_i \sum_j a_{ij}^2)^{1/2}$
$\boldsymbol{A} > \boldsymbol{B}$ $(\boldsymbol{A} \geq \boldsymbol{B})$ — kai $\boldsymbol{A}$ ir $\boldsymbol{B}$ yra vienodos dimensijos kvadratinės matricos, reiškia, kad matrica $\boldsymbol{A} - \boldsymbol{B}$ yra teigiamai (neneigiamai) apibrėžta.

# 1 skyrius

# Pradinės sąvokos

## 1.1. Statistinės hipotezės

Atsitiktinis vektorius $\mathbf{X} = (X_1, ..., X_n)^T$ vadinamas didumo $n$ *paprastąja imtimi*, jeigu jo koordinatės yra vienodai pasiskirstę nepriklausomi a. d. Realiame eksperimente vektoriaus $\mathbf{X}$ įgytoji reikšmė $\boldsymbol{x} = (x_1, ..., x_n)^T$ vadinama *paprastosios imties realizacija*, o realizacijos vektoriaus elementai vadinami *stebiniais*.

Bendresniu atveju vektoriaus $\mathbf{X}$ elementai $X_i$ gali būti priklausomi arba nevienodai pasiskirstę. Tada vektorius $\mathbf{X}$ vadinamas *imtimi*, o jo įgyta reikšmė $\boldsymbol{x}$ *imties realizacija*.

Tarkime, kad a. v. $\mathbf{X}$ (arba atskiro imties elemento $X_i$ paprastosios imties atveju) pasiskirstymo funkcija $F$ priklauso pasiskirstymo funkcijų aibei $\mathcal{F}$. Pavyzdžiui, paprastosios imties atveju $\mathcal{F}$ gali būti tolydžiųjų, diskrečiųjų, normaliųjų, Puasono skirstinių pasiskirstymo funkcijų aibės. Aibė F nusako *statistinį modelį*.

Tegu $\mathcal{F}_0$ yra aibės $\mathcal{F}$ poaibis.

*Statistine hipoteze* $H_0$ suprasime tvirtinimą: pasiskirstymo funkcija $F$ priklauso aibei $\mathcal{F}_0$. Žymėsime $H_0 : F \in \mathcal{F}_0$. Hipotezę $H_1 : F \in \mathcal{F}_1$, kai $\mathcal{F}_1 \cup \mathcal{F}_0 = \mathcal{F}$, vadiname *alternatyviąja hipoteze*, arba, trumpiau, *alternatyva*.

Jeigu pasiskirstymo funkcijų aibė $\mathcal{F} = \{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \in \mathbf{R}^m\}$ nusakoma baigtinės dimensijos parametru $\boldsymbol{\theta}$, tai statistinis modelis vadinamas *parametriniu*. Tokiu atveju statistinės hipotezės yra *parametrinės*, t. y. jos gali būti suformuluotos baigtinės dimensijos parametro $\boldsymbol{\theta}$ terminais.

Jeigu pasiskirstymo funkcijų aibė $\mathcal{F}$ negali būti nusakyta baigtinės dimensijos parametru, tai tokia aibė ir ją atitinkantis statistinis modelis vadinami *neparametriniais*.

Jei poaibis $\mathcal{F}_0$ susideda tik iš vieno aibės $\mathcal{F}$ elemento, tai hipotezė vadinama *paprastąja*, priešingu atveju – *sudėtine*.

## 1.2. Neparametrinių modelių hipotezių pavyzdžiai

Suformuluosime dažniausiai tikrinamas neparametrinių statistinių modelių hipotezes. Alternatyviųjų hipotezių nepateikiame, nes jas nusakyti neparametriniuose modeliuose paprastai būna problemiška. Alternatyvos bus suformuluotos nagrinėjant konkrečius statistinius kriterijus.

### 1.2.1. Suderinamumo hipotezės

Tarkime, $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis a. d. $X$, kurio pasiskirstymo funkcija $F$ priklauso šeimai $\mathcal{F}$. *Suderinamumo hipoteze* vadiname paprastąją hipotezę $H : F(x) \equiv F_0(x)$; čia $F_0(x)$ – visiškai nusakyta aibės $\mathcal{F}$ pasiskirstymo funkcija, t. y. aibė $\mathcal{F}_0$ susideda iš vienintelio elemento $F_0(x)$. Tokią hipotezę tikriname, pavyzdžiui, jeigu norime įsitikinti, kad kompiuteris sugeneravo skirstinių $N(1,\ 4)$, $U(0,\ 1)$, $\mathcal{P}(3)$ ir pan. paprastųjų imčių realizacijas.

*Suderinamumo hipoteze* taip pat vadiname sudėtinę hipotezę $H : F \in \mathcal{F}_0$, kai $\mathcal{F}_0 = \{F(x; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \in \mathbf{R}^m \subset \mathcal{F}\}$, o $F(x;\ \boldsymbol{\theta})$ yra žinomos analizinės išraiškos pasiskirstymo funkcija, priklausanti nuo baigtinės dimensijos parametro $\boldsymbol{\theta}$. Pavyzdžiui, $\mathcal{F}_0$ gali būti normaliųjų, binominių, Puasono ir pan. pasiskirstymo funkcijų aibė.

Bendresniu atveju sudėtinės suderinamumo hipotezės negali būti nusakytos baigtinės dimensijos parametro terminais. Pavyzdžiui, išgyvenamumo analizėje gali būti tikrinama hipotezė, kad $i$-ojo objekto gyvenimo trukmės pasiskirstymo funkcija turi tokį pavidalą:

$$F_i(x, \boldsymbol{\beta}) = 1 - \{1 - F_0(x)\}^{\exp\{\boldsymbol{\beta}^T \boldsymbol{z}_i\}},$$

čia $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)^T$ nežinomų parametrų vektorius, $\boldsymbol{z} = (z_{1i}, ..., z_{mi})^T$ fiksuota $i$-ojo objekto kovariančių vektoriaus, nuo kurio gali priklausyti gyvenimo trukmė, reikšmė, o $F_0(x)$ nežinoma bazinė pasiskirstymo funkcija.

Paprastosioms suderinamumo hipotezėms tikrinti pateikiame chi kvadrato ir tikėtinumų santykio kriterijus (2.1 skyrelis), Neimano ir Bartono tipo kriterijus (3.1, 3.2 skyreliai), kriterijus, grindžiamus empirinės ir teorinės pasiskirstymo funkcijų skirtumu (4.2, 4.3 skyreliai).

Sudėtinėms suderinamumo hipotezėms tikrinti pateikiame chi kvadrato kriterijus (2.2 ir 2.3 skyreliai), Neimano ir Bartono tipo modifikuotuosius kriterijus (3.4 ir 3.5 skyreliai), kriterijus, grindžiamus empirinės ir teorinės pasiskirstymo funkcijų skirtumu (4.4 skyrelis), bei specialius kriterijus, sukonstruotus konkrečioms tikimybinių skirstinių šeimoms (6.5 skyrelis).

### 1.2.2. Nepriklausomumo hipotezė

Sakykime, kad $(X_i,\ Y_i)^T,\ i = 1, 2, ..., n$, yra paprastoji imtis a. v. $(X,\ Y)^T$, kurio pasiskirstymo funkcija $F = F(x,\ y) \in \mathcal{F}$ priklauso dvimačių pasiskirstymo funkcijų aibei $\mathcal{F}$. Reikia patikrinti hipotezę, kad a. d. $X$ ir $Y$ yra nepriklausomi, t. y. $H : F(x_1,\ x_2) \in \mathcal{F}_0$; čia $\mathcal{F}_0 \subset \mathcal{F}$ yra aibė tokių dvimačių pasiskirstymo funkcijų $\widetilde{F}(x_1,\ x_2)$, kurioms teisinga tapatybė $\widetilde{F}(x_1,\ x_2) \equiv F_1(x_1)F_2(x_2)$ su visais $(x_1,\ x_2) \in \mathbf{R}^2$.

Analogiškai formuluojamos hipotezės dėl didesnio skaičiaus a. d. nepriklausomumo.

Nepriklausomumo hipotezėms tikrinti pateikiame chi kvadrato kriterijų (2.4 skyrelis) ir ranginius kriterijus (5.3 ir 5.10 skyreliai).

### 1.2.3. Atsitiktinumo hipotezė

Sakykime, kad a. v. $\mathbf{X} = (X_1, ..., X_n)^T$ koordinatės yra n. a. d. ir jo pasiskirstymo funkcija $F(x_1, ..., x_n) = F_1(x_1)\cdots F_n(x_n)$ priklauso šeimai $\mathcal{P} = \{F,\ F \in \mathcal{F}\}$; čia $\mathcal{F}$ – tam tikra $n$-mačių pasiskirstymo funkcijų, lygių marginaliųjų pasiskirstymo funkcijų sandaugai, aibė. Reikia patikrinti hipotezę $F(x_1, ..., x_n) \in \mathcal{F}_0$; čia $\mathcal{F}_0 \subset \mathcal{F}$ – aibė tokių $n$ mačių pasiskirstymo funkcijų $\widetilde{F}$, kurių marginaliosios pasiskirstymo funkcijos yra vienodos, t. y. $\widetilde{F}(x_1, ..., x_n) = F(x_1)\cdots F(x_n)$ su visais $(x_1, ..., x_n)^T \in \mathbf{R}^n$. Kitaip sakant, tikriname hipotezę, kad $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji atsitiktinė didumo $n$ imtis.

Atsitiktinumo hipotezei tikrinti pateikiame ranginius kriterijus (5.4 skyrelis) ir serijų kriterijų (6.2 skyrelis).

### 1.2.4. Homogeniškumo hipotezė

Dviejų nepriklausomų paprastųjų imčių $\mathbf{X} = (X_1, ..., X_n)^T$ ir $\boldsymbol{Y} = (Y_1, ..., Y_m)^T$ homogeniškumo hipotezė $H : F_1(x) \equiv F_2(x)$; čia $F_1(x)$ ir $F_2(x)$ yra imčių elementų $X_i$ ir $Y_j$ pasiskirstymo funkcijos. Homogeniškumo hipotezė tuo atveju, kai nepriklausomų imčių skaičius $k > 2$, formuluojama analogiškai.

Homogeniškumo hipotezei tikrinti paprastųjų nepriklausomų imčių atveju pateikiame chi kvadrato kriterijų (2.5 skyrelis), kriterijus, grindžiamus empirinių pasiskirstymo funkcijų skirtumu (4.5 skyrelis), ranginius kriterijus (5.5 ir 5.8 skyreliai) ir keletą specialių kriterijų (6.1, 6.2.3 skyreliai).

Tarkime, kad turime atsitiktinio vektoriaus $\mathbf{X} = (X_1, ..., X_k)^T$ paprastąją didumo $n$ imtį $\mathbf{X}_i = (X_{1i}, ..., X_{ki})^T$, $i = 1, 2, ..., n$. Tada atskirų vektoriaus $\mathbf{X}$ koordinačių paprastosios imtys $(X_{j1}, ..., X_{jn})$, $j = 1, ..., k$, gali būti tarpusavyje priklausomos. Priklausomų imčių homogeniškumo hipotezė $H : F_1(x) \equiv ... \equiv F_k(x)$ tvirtina, kad a. v. $\mathbf{X}$ koordinačių marginaliosios pasiskirstymo funkcijos $F_1(x), ..., F_k(x)$ sutampa.

Priklausomų imčių homogeniškumo hipotezėms tikrinti pateikiame ranginius kriterijus (5.7 ir 5.9 skyreliai) ir keletą specialių kriterijų (6.1, 6.3 ir 6.4 skyreliai).

### 1.2.5. Hipotezė dėl medianos reikšmės

Tarkime, $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis, gauta stebint absoliučiai tolydųjį a. d. $X$. Pažymėkime $M$ atsitiktinio dydžio $X$ medianą. Tikriname hipotezę $H : M = M_0$, kad medianos reikšmė lygi skaičiui $M_0$.

Hipotezėms dėl medianos reikšmės pateikiame ranginius kriterijus (5.6 skyrelis) ir ženklų kriterijų (6.1 skyrelis).

## 1.3. Statistinis kriterijus

*Statistinis kriterijus* arba tiesiog *kriterijus* yra taisyklė, pagal kurią remiantis imties realizacija daromas sprendimas apie hipotezės $H_0$ teisingumą ar klaidingumą. Paprastai sprendimas grindžiamas tam tikros statistikos $T = T(\mathbf{X}) = T(X_1, ..., X_n)$, vadinamos *kriterijaus statistika*, realizacija. Natūralu parinkti statistiką $T$ taip, kad jos skirstinys esant teisingai ir klaidingai tikrinamai hipotezei skirtųsi kuo labiau.

Jeigu statistika $T$, kai hipotezė teisinga, turi tendenciją įgyti mažesnes (didesnes) reikšmes, negu esant teisingai alternatyvai $H_1$, tai hipotezė $H_0$ atmetama, kai $T > c$ ($T < c$), čia $c$ yra specialiai parenkamas realus skaičius. Jeigu esant teisingai hipotezei $H_0$ statistikos $T$ reikšmės turi tendenciją įgyti reikšmes iš tam tikro intervalo, o esant teisingai alternatyvai – už intervalo ribų, tai hipotezė $H_0$ atmetama, kai $T < c_1$ arba $T > c_2$, čia $c_1$ ir $c_2$ yra specialiai parinkti realūs skaičiai.

Tarkime, hipotezė $H_0$ atmetama, kai $T > c$ (kiti du atvejai aptariami analogiškai).

Tikimybė

$$\beta(F) = \mathbf{P}_F\{T > c\}, \quad F \in \mathcal{F},$$

atmesti hipotezę $H_0$, kai imties pasiskirstymo funkcija yra $F \in \mathcal{F}$, vadinama kriterijaus *galios funkcija*. Naudodami bet kurį kriterijų galime padaryti dviejų rūšių klaidas:

1. Galima atmesti hipotezę $H_0$, kai ji yra teisinga, t. y. $F \in \mathcal{F}_0$. Tokia klaida vadinama *pirmosios rūšies klaida*. Šios klaidos padarymo tikimybė yra $\beta(F), F \in \mathcal{F}_0$.

2. Galima priimti hipotezę $H_0$, kai ji yra klaidinga, t. y. $F \in \mathcal{F}_1$. Tokia klaida vadinama *antrosios rūšies klaida*. Šios klaidos padarymo tikimybė yra $1 - \beta(F), F \in \mathcal{F}_1$.

Skaičius

$$\sup_{F \in \mathcal{F}_0} \beta(F) \tag{1.3.1}$$

vadinamas kriterijaus *reikšmingumo lygmeniu*.

Fiksuokime $\alpha \in (0, 1)$. Statistinis kriterijus vadinamas *reikšmingumo lygmens $\alpha$ kriterijumi*, jeigu su visais $F \in \mathcal{F}_0$ pirmosios rūšies klaidos tikimybė neviršija $\alpha$. Paprastai reikšmingumo lygmeniu parenkamas artimas nuliui skaičius: $\alpha = 0,1$; $0,05$; $0,01$ ir pan.

Jeigu statistikos $T$ skirstinys yra absoliučiai tolydus, tai su bet kuriuo $\alpha \in (0, 1)$ reikšmingumo lygmuo yra pasiekiamas, t. y. atsiras toks $F \in \mathcal{F}_0$, kad $\beta(F) = \alpha$.

Reikšmingumo lygmens $\alpha$ kriterijus vadinamas *nepaslinktuoju*, jeigu

$$\inf_{F \in \mathcal{F}_1} \beta(F) \geq \alpha. \tag{1.3.2}$$

Tai reiškia, kad tikimybė atmesti hipotezę $H_0$, kai ji neteisinga, yra ne mažesnė, negu tada, kai ji teisinga.

Tarkime, $\mathcal{T}$ yra aibė statistikų, kurių pagrindu sukonstruoti nepaslinktieji reikšmingumo lygmens $\alpha$ kriterijai. Sakysime, kad statistika $T$ apibrėžia *tolygiai galingiausiąjį reikšmingumo lygmens $\alpha$ kriterijų*, jeigu bet kuriai kitai statistikai $T^* \in \mathcal{T}$ galioja nelygybė

$$\beta_T(F) \geq \beta_{T^*}(F), \quad \forall F \in \mathcal{F}_1. \tag{1.3.3}$$

Statistinis kriterijus vadinamas *pagrįstuoju*, jeigu su visais $F \in \mathcal{F}_1$

$$\beta(F) \to 1, \quad n \to \infty. \tag{1.3.4}$$

## 1.4. $P$ reikšmė

Praktiškai statistiniai kriterijai dažnai formuluojami vadinamųjų $P$ reikšmių terminais (žr. I dalį, 4.1.2 skyrelį). Priminsime $P$ reikšmių apibrėžimą ir statistinių kriterijų formulavimą jų terminais atsižvelgdami į neparametrinių hipotezių specifiką.

Tegu kriterijus grindžiamas vienamate statistika $T = T(\mathbf{X})$ ir jo kritinė sritis (hipotezės atmetimo sritis) turi vieną iš tokių trijų pavidalų

$$1)\ K_1 = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq c_1\}; \quad 2)\ K_2 = \{\boldsymbol{x} : T(\boldsymbol{x}) \leq c_2\};$$
$$3)\ K_3 = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq d_1 \ \text{arba}\ T(\boldsymbol{x}) \leq d_2\}. \tag{1.4.1}$$

Nagrinėjant reikšmingumo lygmens $\alpha$ hipotezės $H : F \in \mathcal{F}_0$ tikrinimo kriterijus, konstantos $c_1, c_2, d_1, d_2$ turėtų tenkinti sąlygas

$$1)\ \alpha = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq c_1\}; \quad 2)\ \alpha = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq c_2\};$$
$$\frac{\alpha}{2} = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq d_1\} = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq d_2\}. \tag{1.4.2}$$

Pažymėkime $t = T(\boldsymbol{x})$ statistikos $T$ realizaciją, kuri žinoma, jei žinoma imties $\mathbf{X}$ realizacija $\boldsymbol{x}$.

Apibrėžkime $P$ reikšmes tokio tipo kritinėms sritims lygybėmis:

$$1)\ pv = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq t\}; \quad 2)\ pv = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq t\};$$
$$3)\ pv = 2\min(\sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq t\}, \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq t\}. \tag{1.4.3}$$

**4.1.1 pastaba.** Dažniausiai

$$\sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq t\} = \mathbf{P}_{F_0}\{T \geq t\}, \quad \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq t\} = \mathbf{P}_{F_0}\{T \leq t\},$$

čia $F_0$ yra aibių $\mathcal{F}_0$ ir $\mathcal{F}_1$ uždarinių sankirta.

**1.4.1 teorema.** *Tarkime, kad kriterijaus kritinė sritis turi vieną iš trijų (1.4.1) pavidalų. Eksperimente, kuriame statistika $T$ įgijo reikšmę $t$, hipotezė $H$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi tada ir tik tada, kai $pv \leq \alpha$.*

**Įrodymas.** Remdamiesi (1.4.1), (1.4.2) ir $pv$ apibrėžimais (1.4.3) gauname

$$1) t \geq c_1 \Leftrightarrow pv = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq t\} \leq \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq c_1\} = \alpha;$$

$$2) t \leq c_2 \Leftrightarrow pv = \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq t\} \leq \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq c_2\} = \alpha;$$

$$3) t \leq d_1 \text{ arba } t \geq d_2 \Leftrightarrow \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq t\} \leq \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq d_1\} = \alpha/2; \text{ arba}$$

$$\sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq t\} \leq \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq d_2\} = \alpha/2; \Leftrightarrow$$

$$pv = 2\min(\sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq t\}, \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq t\}) \leq$$

$$2\min(\sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \geq d_1\}, \sup_{F \in \mathcal{F}_0} \mathbf{P}_F\{T \leq d_2\}) = \alpha.$$

▲

Jeigu kritinė sritis apibrėžiama naudojant asimptotinį statistikos $T$ skirstinį (paprastai normalųjį ar chi kvadrato), tada $P$ reikšmė $pv_a$, randama iš asimptotinio statistikos $T$ skirstinio, vadinama *asimptotine $P$ reikšme*.

## 1.5. Tolydumo pataisa

Jeigu statistikos $T$ skirstinys yra diskretusis ir aproksimuojamas absoliučiai tolydžiuoju (paprastai normaliuoju) skirstiniu, tai aproksimavimo tikslumas padidėja įvedus vadinamąją Jeitso tolydumo pataisą [31].

Tolydumo pataisos idėją pailiustruosime konkrečiu pavyzdžiu.

**1.5.1 pavyzdys.** Tikrinama parametrinė hipotezė $H : p \leq 0,5$, kai alternatyva yra $H_1 : p > 0,5$, pagal didumo $n$ paprastąją imtį, gautą stebint a.d. $X \sim B(1, p)$. Tarkime, atlikus $n = 20$ Bernulio eksperimentų, nagrinėjamas įvykis pasirodė $T = 13$ kartų. Hipotezė $H$ atmetama, kai statistika $T$ įgyja dideles reikšmes, t.y. turi pavidalą $T \geq c$. Kai hipotezė teisinga, statistika $T \sim B(20,\, 0,5)$. Gauname $P$ reikšmę

$$pv = \mathbf{P}\{T \geq 13\} = \sum_{i=13}^{20} C_{20}^i (1/2)^{20} = I_{1/2}(13,\, 8) = 0,131588.$$

Pagal normaliąją aproksimaciją

$$Z_n = (T - 0,5n)/\sqrt{0,25n} = (T-10)/\sqrt{5} \xrightarrow{d} Z \sim N(0,\, 1)$$

gauname

$$pv_a = \mathbf{P}\{T \geq 13\} = \mathbf{P}\left\{\frac{T-10}{\sqrt{5}} \geq \frac{13-10}{\sqrt{5}}\right\} \approx$$

$$1 - \Phi\left(\frac{13-10}{\sqrt{5}}\right) = 0,089856.$$

Gautoji reikšmė žymiai mažesnė už tikrąją $P$ reikšmę.

Reikia pažymėti, kad $\mathbf{P}\{T \geq 13\} = \mathbf{P}\{T > 12\}$. Todėl galima sudaryti dvi šios tikimybės aproksimacijas normaliuoju skirstiniu:

$$1 - \Phi(13 - 10/\sqrt{5}) = 0,089856$$

arba

$$1 - \Phi(12 - 10/\sqrt{5}) = 0,185547.$$

Tolydumo pataisa reiškia, kad normaliąją aproksimaciją taikome imdami intervalo (12, 13] vidurį. Taigi asimptotinė $P$ reikšmė su tolydumo pataisa yra

$$pv_{ap} = 1 - \Phi((13 - 0,5 - 10)/\sqrt{5}) = 0,131776.$$

Gautoji reikšmė yra artima tikrajai $P$ reikšmei.

Jeigu tikrinama parametrinė hipotezė $H$, kai alternatyva yra $H_2 : p < 0,5$, tai hipotezė $H$ atmetama, kai statistika $T$ įgyja mažas reikšmes, t. y. turi pavidalą $T \leq d$. Gauname $P$ reikšmę

$$pv = \mathbf{P}\{T \leq 13\} = \sum_{i=0}^{13} C_{20}^{i}(1/2)^{20} = I_{1/2}(7,\, 14) = 0,942341.$$

Pagal normaliąją aproksimaciją

$$pv_a = \Phi(13 - 10/\sqrt{5}) = 0,910144.$$

Gautoji reikšmė daug mažesnė už tikrąją $P$ reikšmę.

Reikia pažymėti, kad $\mathbf{P}\{T \leq 13\} = \mathbf{P}\{T < 14\}$. Todėl galima sudaryti dvi šios tikimybės aproksimacijas normaliuoju skirstiniu:

$$\Phi(13 - 10/\sqrt{5}) = 0,910144, \quad \Phi(14 - 10/\sqrt{5}) = 0,963181.$$

Taikydami normaliąją aproksimaciją intervalo (13, 14] viduryje, gauname

$$pv_{ap} = \Phi((13 + 0,5 - 10)/\sqrt{5}) = 0,941238,$$

kuri yra artima tikrajai $P$ reikšmei.

Dvipusės alternatyvos $H_3 : p \neq 0,5$ atveju $P$ reikšmė

$$pv = 2\min\{F_T(13), 1 - F_T(13-)\} =$$
$$2\min\{0,942341,\, 0,131588\} = 0,263176,$$

o

$$pv_a = 2\min\{0,910144,\, 0,089856\} = 0,179712.$$

Asimptotinė $P$ reikšmė su tolydumo pataisa

$$pv_{ap} = 2\min\{\Phi((13 + 0,5 - 10)/\sqrt{5}), 1 - \Phi((13 - 0,5 - 10)/\sqrt{5})\} =$$
$$2\min\{0,941238,\, 0,131776\} = 0,263452.$$

Matome, kad visais atvejais aproksimacija su tolydumo pataisa yra gerokai tikslesnė.

Bendru atveju tarkime, kad sveikaskaitinė statistika $T$ esant teisingai tikrinamajai hipotezei asimptotiškai turi normalųjį skirstinį, t. y. centruotos ir normuotos statistikos

$$Z = \frac{T - \mathbf{E}T}{\sqrt{\boldsymbol{V}T}}$$

skirstinys asimptotiškai yra standartinis normalusis. Tarkime, tikrinamos hipotezės kriterijus, atsižvelgiant į alternatyvas, apibrėžiamas nelygybėmis

$$a)\ T \geq c; \quad b)\ T \leq d; \quad c)\ T \leq c_1 \ \text{ arba } \ T \geq c_2.$$

Tada $P$ reikšmė su tolydumo pataisa yra

$$a)\ pv_{ap} = 1 - \Phi((t - 0,5 - \mathbf{E}T)/\sqrt{\boldsymbol{V}T});$$
$$b)\ pv_{ap} = \Phi((t + 0,5 - \mathbf{E}T)/\sqrt{\boldsymbol{V}T});$$
$$c)\ pv_{ap} = 2\min\left[\Phi\left(\frac{t + 0,5 - \mathbf{E}T}{\sqrt{\boldsymbol{V}T}}\right), 1 - \Phi\left(\frac{t - 0,5 - \mathbf{E}T}{\sqrt{\boldsymbol{V}T}}\right)\right], \qquad (1.5.1)$$

čia $t$ yra statistikos $T$ realizacija.

## 1.6. Kriterijų asimptotinis santykinis efektyvumas

Tarkime, kad esant teisingai tikrinamajai hipotezei arba alternatyvai imties skirstinys priklauso neparametrinei skirstinių šeimai, priklausančiai nuo skaliarinio parametro $\theta$ ir kito parametro $\boldsymbol{\vartheta}$. Tikrinsime hipotezę $H_0 : \theta = \theta_0$, kai alternatyva yra vienpusė $H_1 : \theta > \theta_0$ arba $H_2 : \theta < \theta_0$ bei dvipusė $H_3 : \theta \neq \theta_0$.

**1.6.1 pavyzdys.** Tegu $\mathbf{X} = (X_1, ..., X_n)^T$ ir $\boldsymbol{Y} = (Y_1, ..., Y_m)^T$ yra dvi nepriklausomos paprastosios imtys; $X_i \sim F(x)$ ir $Y_i \sim F(x - \theta)$; kur $F(x)$ yra nežinoma absoliučiai tolydi pasiskirstymo funkcija (parametras $\boldsymbol{\vartheta}$), o $\theta$ yra poslinkio parametras. Homogeniškumo hipotezę galima suformuluoti parametro $\theta$ terminais: $H_0 : \theta = 0$, kai alternatyvos yra $H_1 : \theta > 0, H_2 : \theta < 0, H_3 : \theta \neq 0$.

Nagrinėkime vienpusę alternatyvą $H_1$. Fiksuokime reikšmingumo lygmenį $\alpha \in (0, 1)$. Tarkime, hipotezė atmetama, kai

$$T_n > c_{n,\alpha},$$

čia $n$ yra imties didumas, o $T_n$ – kriterijaus statistika. Kriterijaus galios funkcija

$$\beta_n(\theta) = \mathbf{P}\{T_n > c_{n,\alpha}\}.$$

Jeigu kriterijus yra pagrįstas, tai jo galios funkcija artėja į 1 su bet kuria alternatyviąja parametro $\theta$ reikšme. Taigi kriterijaus galios riba su bet kuria fiksuota alternatyva nėra tinkamas rodiklis kriterijams palyginti. Kriterijų galias galima palyginti imant alternatyvų seką

$$H_n : \theta = \theta_n = \theta_0 + \frac{h}{n^\delta}, \quad \delta > 0, \quad h > 0,$$

kuri augant imties didumui $n$ artėja prie hipotetinės parametro reikšmės $\theta_0$.

Tarkime, kad kriterijams, kurių statistikos yra $T_{1n}$ ir $T_{2n}$, galioja lygybės

$$\theta_m = \theta_0 + \frac{h_1}{n_{1m}^\delta} = \theta_0 + \frac{h_2}{n_{2m}^\delta},$$

čia $n_{im} \to \infty$, kai $m \to \infty$, ir

$$\lim_{m\to\infty} \beta_{n_{1m}}(\theta_m) = \lim_{m\to\infty} \beta_{n_{2m}}(\theta_m).$$

Tada riba (jeigu ji egzistuoja su kuria nors seka $\theta_m$)

$$e(T_{1n}, T_{2n}) = \lim_{m\to\infty} \frac{n_{2m}}{n_{1m}}$$

yra vadinama pirmojo kriterijaus *asimptotiniu santykiniu efektyvumu* (ASE) antrojo kriterijaus atžvilgiu [26].

Kai įvykdytos tam tikros reguliarumo sąlygos ASE egzistuoja ir turi palyginti paprastą pavidalą.

**Reguliarumo sąlygos:**

$$1)\mathbf{P}_{\theta_0}\{T_{in} \geq c_{n,\alpha}\} \to \alpha.$$

2) Taško $\theta_0$ aplinkoje egzistuoja

$$\mu_{in}(\theta) = \mathbf{E}_\theta T_{in}, \quad \sigma_{in}^2 = \boldsymbol{V}_\theta T_{in};$$

funkcija $\mu_{in}(\theta)$ turi baigtinę išvestinę $\dot\mu_{in}(\theta_0)$ taške $\theta_0$, tarkime, $\dot\mu_{in}(\theta_0) > 0$.

3) Egzistuoja ribos

$$\lim_{n\to\infty} \mu_{in}(\theta) = \mu_i(\theta), \quad \lim_{n\to\infty} n^\delta \sigma_{in}(\theta) = \sigma_i(\theta), \quad \mu_i(\theta_0)/\sigma_i(\theta_0) > 0,$$

čia $\delta > 0$.

4) Su visais $h > 0$

$$\dot\mu_{in}(\theta_n) \to \dot\mu_i(\theta_0), \quad \sigma_{in}(\theta_n) \to \sigma_i(\theta_0).$$

5) Kriterijų statistikos asimptotiškai normaliosios:

$$\mathbf{P}_{\theta_n}\{(T_{in} - \mu_{in}(\theta_n))/\sigma_{in}(\theta_n) \leq z\} \to \Phi(z).$$

**1.6.1 teorema.** *Jeigu išpildytos pateiktos reguliarumo sąlygos, tai kriterijų asimptotinis santykinis efektyvumas gali būti apskaičiuotas pagal tokią formulę:*

$$e(T_{1n}, T_{2n}) = \left(\frac{\dot\mu_1(\theta_0)/\sigma_1(\theta_0)}{\dot\mu_2(\theta_0)/\sigma_2(\theta_0)}\right)^{1/\delta}. \tag{1.6.1}$$

**Įrodymas.** Pradžioje praleisime indeksą $i$. Nagrinėsime kriterijaus galios ribą $\lim_{n\to\infty} \beta_n(\theta_n)$. Remdamiesi 1) sąlyga gauname

$$\mathbf{P}_{\theta_0}\left\{\frac{T_n - \mu_n(\theta_0)}{\sigma_n(\theta_0)} > z_{n,\alpha}\right\} \to \alpha,$$

$$z_{n,\alpha} = \frac{c_{n,\alpha} - \mu_n(\theta_0)}{\sigma_n(\theta_0)} \to z_\alpha.$$

Remdamiesi 2) – 4) sąlygomis

$$\frac{\mu_n(\theta_n) - \mu_n(\theta_0)}{\sigma_n(\theta_0)} = \frac{\dot\mu_n(\theta_0)hn^{-\delta} + o(1)}{\sigma_n(\theta_0)n^{-\delta} + o(1)} \to \frac{\dot\mu(\theta_0)}{\sigma(\theta_0)}h.$$

Panaudoję 5) sąlyga randame

$$\beta_n(\theta_n) = \mathbf{P}_{\theta_n}\{T_n > c_{n,\alpha}\} = \mathbf{P}_{\theta_n}\left\{\frac{T_n - \mu_n(\theta_n)}{\sigma_n(\theta_n)} > \frac{c_{n,\alpha} - \mu_n(\theta_n)}{\sigma_n(\theta_n)}\right\}$$

$$= \mathbf{P}_{\theta_n}\left\{\frac{T_n - \mu_n(\theta_n)}{\sigma_n(\theta_n)} > z_{n,\alpha}\frac{\sigma_n(\theta_0)}{\sigma_n(\theta_n)} - \frac{\mu_n(\theta_n) - \mu_n(\theta_0)}{\sigma_n(\theta_0)}\frac{\sigma_n(\theta_0)}{\sigma_n(\theta_n)}\right\}$$

$$\to 1 - \Phi\left(z_\alpha - h\frac{\dot{\mu}_n(\theta_0)}{\sigma(\theta_0)}\right).$$

Tegu $T_{1n}$ ir $T_{2n}$ dvi statistikos, kurioms patenkintos teoremos sąlygos, ir

$$\theta_m = \theta_0 + \frac{h_1}{n_{1m}^\delta} = \theta_0 + \frac{h_2}{n_{2m}^\delta}$$

artėjančių alternatyvų seka. Iš šios lygybės gauname

$$\frac{n_{2m}}{n_{1m}} = \left(\frac{h_2}{h_1}\right)^{1/\delta}$$

Turime

$$\beta_{n_{im}}(\theta_m) \to 1 - \Phi\left(z_\alpha - h_i\frac{\dot{\mu}_i(\theta_0)}{\sigma_i(\theta_0)}\right), \quad i = 1,\, 2.$$

Parenkame $n_{1m}$ ir $n_{2m}$ sulygindami kriterijų galių ribas. Gauname

$$\frac{n_{2m}}{n_{1m}} = \left(\frac{h_2}{h_1}\right)^{1/\delta} = \left(\frac{\dot{\mu}_1(\theta_0)/\sigma_1(\theta_0)}{\dot{\mu}_2(\theta_0)/\sigma_2(\theta_0)}\right)^{1/\delta}.$$

▲

# 2 skyrius

# Chi kvadrato kriterijus

Vienas iš būdų neparametriniams kriterijams sudaryti yra toks: vietoje gautosios imties yra naudojami stebėjimų patekimo į tam tikras nesikertančias sritis dažniai. Tada, kad ir koks būtų pradinis skirstinys, gauname polinominį skirstinį, aprašantį minėtų dažnių pasiskirstymą. Tokiu būdu sukonstruoti kriterijai tampa tiesiogiai nepriklausomi nuo pradinio skirstinio.

## 2.1. Paprastosios suderinamumo hipotezės tikrinimas

Tarkime, paprastoji imtis $\mathbf{X} = (X_1, ..., X_n)^T$ gauta stebint a. d. $X$, kurio pasiskirstymo funkcija $F$ priklauso aibei $\mathcal{F}$.

**Paprastoji suderinamumo hipotezė**:

$$H_0 : F(x) = F_0(x), \quad \forall\ x \in \mathbf{R}, \tag{2.1.1}$$

čia $F_0(x)$ yra visiškai nusakyta (žinoma) aibės $\mathcal{F}$ pasiskirstymo funkcija.

Pavyzdžiui,

$$H_0 : X \sim U(0,\,1), \quad H_0 : X \sim B(10,\,0{,}5), \quad X \sim N(3,\,4),$$

yra paprastosios suderinamumo hipotezės. Tokias hipotezes tikriname, pavyzdžiui, tada, kai norime įsitikinti, kad kompiuteriu sugeneruotą skaičių rinkinį galime interpretuoti kaip realizaciją paprastosios imties, gautos stebint a. d. $X \sim N(0,\,1)$, $X \sim \mathcal{P}(3)$ ir pan.

Sudalinkime abscisių ašį į intervalus: $-\infty = a_0 < a_1 < ... < a_k = \infty$. Pažymėkime $U_j$ stebėjimų, patekusių į intervalą $(a_{j-1},\ a_j]$, skaičių

$$U_j = \sum_{i=1}^{n} \mathbf{1}_{(a_{j-1},a_j]}(X_i), \quad j = 1, ..., k.$$

Vietoje pradinės imties $\mathbf{X}$ gauname mažiau informatyvią grupuotąją imtį

$$\boldsymbol{U} = (U_1, ..., U_k)^T.$$

Atsitiktinis vektorius $(U_1, ..., U_k)^T$ turi polinominį skirstinį $\mathcal{P}(n, \ \boldsymbol{\pi})$: kai $0 \leq m_i \leq n, \sum_i m_i = n$,

$$\mathbf{P}\{U_1 = m_1, ..., U_k = m_k\} = \frac{n!}{m_1!...m_k!}\pi_1^{m_1}...\pi_k^{m_k}, \tag{2.1.2}$$

čia $\pi_i = \mathbf{P}\{X \in (a_{i-1}, a_i]\} = F(a_i) - F(a_{i-1})$ yra tikimybė, kad a. d. $X$ įgis reikšmę iš intervalo $(a_{i-1}, a_i]$, $\boldsymbol{\pi} = (\pi_1, ..., \pi_k)^T$, $\pi_1 + ... + \pi_k = 1$.

Vietoje hipotezės (2.1.1) tikrinsime hipotezę apie polinominio skirstinio parametrų reikšmes.

**Hipotezė apie polinominio skirstinio parametrų reikšmes:**

$$H_0' : \pi = \pi_{i0} = F_0(a_i) - F_0(a_{i-1}), \ i = 1, 2, ..., k. \tag{2.1.3}$$

Kai hipotezė $H_0'$ teisinga, tai

$$\boldsymbol{U} \sim \mathcal{P}_k(n, \boldsymbol{\pi}_0),$$

čia $\boldsymbol{\pi}_0 = (\pi_{10}, ..., \pi_{k0})$, $\pi_{10} + ..., +\pi_{k0} = 1$.

Atmetus hipotezę $H_0'$, natūralu atmesti ir hipotezę $H_0$.

Pirsono chi kvadrato kriterijus hipotezei $H_0'$ tikrinti yra grindžiamas skirtumais tarp tikimybių $\pi_j$ DT įvertinių $\hat{\pi}_j$, gautų pagal grupuotąją imtį $\boldsymbol{U}$, ir hipotetinių šių tikimybių reikšmių $\pi_{j0}$.

Iš sąlygos $\pi_1 + ... + \pi_k = 1$ gauname, kad polinominis skirstinys $\mathcal{P}_k(n, \boldsymbol{\pi})$ faktiškai priklauso nuo $(k-1)$-mačio parametro $(\pi_1, ..., \pi_{k-1})^T$.

Pagal (2.1.2) atsitiktinio vektoriaus $(U_1, ..., U_k)^T$ tikėtinumo funkcija

$$L(\boldsymbol{\pi}) = \frac{n!}{U_1!...U_k!}\pi_1^{U_1}...\pi_k^{U_k}, \tag{2.1.4}$$

o logtikėtinumo funkcija

$$\ell(\pi_1, ..., \pi_{k-1}) = \sum_{j=1}^{k-1} U_j \ln \pi_j + U_k \ln(1 - \sum_{j-1}^{k-1} \pi_j) + \ln C.$$

Iš čia

$$\dot{\ell}_j = \frac{U_j}{\pi_j} - \frac{U_k}{1 - (\pi_1 + ... + \pi_{k-1})} = \frac{U_j}{\pi_j} - \frac{U_k}{\pi_k},$$

ir su visais $j, l = 1, ..., k$

$$U_j \pi_l = U_l \pi_j.$$

Sumuodami pagal $l$ ir atsižvelgę į tai, kad $\pi_1 + ... + \pi_k = 1$, $U_1 + ... + U_k = n$, gauname $U_j = n\pi_j$. Taigi parametrų $\pi_j$ DT įvertiniai yra

$$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, ..., \hat{\pi}_k)^T, \quad \hat{\pi}_j = U_j/n, \quad j = 1, ..., k.$$

*Pirsono statistika* turi tokį pavidalą

$$X_n^2 = \sum_{i=1}^{k} \frac{(\sqrt{n}(\hat{\pi}_i - \pi_{i0}))^2}{\pi_{i0}} = \sum_{i=1}^{k} \frac{(U_i - n\pi_{i0})^2}{n\pi_{i0}} = \frac{1}{n}\sum_{i=1}^{k} \frac{U_i^2}{\pi_{i0}} - n. \qquad (2.1.5)$$

Jeigu hipotezė $H_0'$ teisinga, tai skirtumo $\hat{\pi}_i - \pi_{i0}$ realizacijos turi tendenciją koncentruotis apie nulį. Priešingu atveju atsiras tokios indeksų $i$ reikšmės, kad šių skirtumų realizacijos grupuosis apie reikšmę, nutolusią nuo nulio, taigi statistika $X_n^2$ turės tendenciją įgyti didesnes reikšmes. Vadinasi, hipotezė $H_0'$ atmestina, kai statistika $X_n^2$ įgyja dideles reikšmes.

Pirsono kriterijus yra asimptotinis ir grindžiamas toliau pateikiama statistikos $X_n^2$ aproksimacija chi kvadrato skirstiniu.

**2.1.1 teorema.** *Jeigu* $0 < \pi_{i0} < 1$, $\pi_{10} + ... + \pi_{k0} = 1$, *tai esant teisingai hipotezei*

$$X_n^2 \xrightarrow{d} \chi_{k-1}^2, \quad \text{kai} \quad n \to \infty.$$

**Įrodymas.** Kai hipotezė $H_0'$ teisinga, a. v. $\boldsymbol{U} = (U_1, ..., U_k)^T$ yra suma vienodai pasiskirsčiusių nepriklausomų a. v. $\mathbf{X}_j \sim \mathcal{P}_k(1, \boldsymbol{\pi}_0)$ su vidurkiu $\boldsymbol{\pi}_0$ ir kovariacine matrica $\boldsymbol{D} = [d_{ij}]_{k\times k}$, $d_{ii} = \pi_{i0}(1 - \pi_{i0})$, $d_{ij} = -\pi_{i0}\pi_{j0}$, $i \neq j$.

Jeigu $0 < \pi_{i0} < 1$, $\pi_{10} + ... + \pi_{k0} = 1$, tai sumai galioja daugiamatė CRT

$$\frac{(\boldsymbol{U} - n\boldsymbol{\pi}_0)^T}{\sqrt{n}} = \sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)^T \xrightarrow{d} \boldsymbol{Y} \sim N_k(\mathbf{0}, \boldsymbol{D}), \qquad (2.1.6)$$

kai $n \to \infty$. Matricą $\boldsymbol{D}$ galima užrašyti tokiu pavidalu:

$$\boldsymbol{D} = \boldsymbol{p}_0 - \boldsymbol{p}_0\boldsymbol{p}_0^T,$$

čia $\boldsymbol{p}_0$ yra diagonalioji matrica su elementais $\pi_{10}, ..., \pi_{k0}$ ant pagrindinės įstrižainės. Atsitiktiniam vektoriui

$$\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{p}_0^{-1/2}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) = \left(\frac{\sqrt{n}(\hat{\pi}_1 - \pi_{10})}{\sqrt{\pi_{10}}}, ..., \frac{\sqrt{n}(\hat{\pi}_k - \pi_{k0})}{\sqrt{\pi_{k0}}}\right)^T$$

taip pat galioja daugiamatė CRT

$$\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma}),$$

čia

$$\boldsymbol{\Sigma} = \boldsymbol{p}_0^{-1/2}\boldsymbol{D}\boldsymbol{p}_0^{-1/2} = \boldsymbol{I}_k - \boldsymbol{q}\boldsymbol{q}^T,$$

čia $\boldsymbol{q} = (\sqrt{\pi_{10}}, ..., \sqrt{\pi_{k0}})^T$, $\boldsymbol{q}^T\boldsymbol{q} = 1$, o $\boldsymbol{I}_k$ yra vienetinė $k \times k$ matrica. Matrica $\boldsymbol{\Sigma}$ yra idempotentinė ($\boldsymbol{\Sigma}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$), jos rangas $Rang\boldsymbol{\Sigma} = k - 1$, o apibendrintoji atvirkštinė $\boldsymbol{\Sigma}^- = \boldsymbol{I}_k + \boldsymbol{q}\boldsymbol{q}^T$ (žr. 2.4 pratimą). Gauname

$$X_n^2 = \boldsymbol{Z}_n^T\boldsymbol{\Sigma}^-\boldsymbol{Z}_n = ||\boldsymbol{Z}_n||^2 \xrightarrow{d} ||\boldsymbol{Z}||^2 \sim \chi^2(k-1). \qquad (2.1.7)$$

▲

Remdamiesi teorema gauname:

**Pirsono chi-kvadrato kriterijus:** hipotezė $H_0'$ atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai teisinga nelygybė

$$X_n^2 > \chi_\alpha^2(k-1). \tag{2.1.8}$$

Hipotezei $H_0'$ tikrinti galime sudaryti tikėtinumų santykio kriterijų, grindžiamą statistika

$$\Lambda = \frac{L(\boldsymbol{\pi}_0)}{\sup_{\boldsymbol{\pi}} L(\boldsymbol{\pi})} = \frac{L(\boldsymbol{\pi}_0)}{L(\hat{\boldsymbol{\pi}})} = n^n \prod_{i=1}^{k} \left(\frac{\pi_{i0}}{U_i}\right)^{U_i} = \prod_{i=1}^{k} \left(\frac{n\pi_{i0}}{U_i}\right)^{U_i}.$$

Kai teisinga hipotezė (žr. A priedą, 6.1.2 pastabą), asimptotiškai ($n \to \infty$)

$$R_n = -2\ln\Lambda = 2\sum_{i=1}^{k} U_i \ln\frac{U_i}{n\pi_{i0}} \xrightarrow{d} V \sim \chi^2(k-1). \tag{2.1.9}$$

Taigi statistikos $R_n$ ir $X_n^2$ asimptotiškai ekvivalenčios.

**Tikėtinumų santykio kriterijus:** hipotezė $H_0'$ atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai teisinga nelygybė

$$R_n > \chi_\alpha^2(k-1). \tag{2.1.10}$$

**2.1.1 pastaba.** Hipotezės $H_0$ ir $H_0'$ bendru atveju nėra ekvivalenčios. Hipotezėje $H_0'$ tvirtinama tik tiek, kad pasiskirstymo funkcijos pokytis $j$-ame intervale yra $\pi_{j0}$, tačiau nereglamentuojamas pasiskirstymo funkcijos elgesys intervalo viduje. Jeigu $n$ didelis, tai galima padidinti grupavimo intervalų skaičių ir šitaip šias hipotezes suartinti.

**2.1.2 pastaba.** Reikia turėti omenyje, kad kriterijai (2.1.8), (2.1.10) yra apytiksliai, gauti su sąlyga, kad imties didumas $n \to \infty$. Todėl išvadų tikslumas priklauso nuo to, kaip gerai galioja aproksimacijos (2.1.7), (2.1.9). Jeigu parinksime per didelį grupavimo intervalų skaičių, tai kiekviename intervale dažniai įgis tik reikšmę 0 arba 1, ir aproksimacija bus netiksli. Todėl intervalų skaičius $k$ neturėtų būti per daug didelis. Praktinė taisyklė: grupavimo intervalus reikia parinkti taip, kad $n\pi_{i0} \geq 5$.

**2.1.3 pastaba.** Jeigu kyla abejonių dėl statistikos $X_n^2$ (arba $R_n$) aproksimavimo chi kvadrato skirstiniu tikslumo, tai kriterijų galima patikslinti naudojant kompiuterinį modeliavimą.

Tarkime, statistikos $X_n^2$ realizacija yra $x_n^2$. Modeliuokime $N$ kartų atsitiktinį vektorių $\boldsymbol{U} \sim \mathcal{P}_k(n, \boldsymbol{\pi}_0)$ ir kiekvieną kartą apskaičiuokime statistikos $X_n^2$ reikšmę. Tegu $M$ žymi, kiek kartų gautosios reikšmės viršija turimą realizaciją $x_n^2$. Tada $P$ reikšmės įverčiu galime imti $\hat{pv} = M/N$. Hipotezė $H_0'$ atmetama apytiksliu $\alpha$ lygmens kriterijumi, kai $\hat{pv} < \alpha$. Pateikto kriterijaus tikslumas priklauso tik nuo realizacijų skaičiaus $N$.

**2.1.4 pastaba.** Jeigu stebimo a. d. $X$ skirstinys yra diskretusis, sukoncentruotas taškuose $x_1, ..., x_k$, tai grupavimas nereikalingas. Imtyje gauname tik galimas reikšmes, o $U_i$ šiuo atveju reiškia reikšmės $x_i$ pasirodymo dažnį.

**2.1.5 pastaba.** Jeigu hipotezė $H_0'$ neteisinga ir $\boldsymbol{U} \sim \mathcal{P}_k(n, \boldsymbol{\pi})$, tai statistikų $R_n$ arba $X_n^2$ skirstiniai aproksimuojami necentriniu chi kvadrato skirstiniu su $k-1$ laisvės laipsniu ir necentriškumo parametru

$$\Delta = 2n \sum_{i=1}^{k} \pi_i \ln \frac{\pi_i}{\pi_{i0}} \approx \delta = n \sum_{i=1}^{k} \frac{(\pi_i - \pi_{i0})^2}{\pi_{i0}}. \qquad (2.1.11)$$

**2.1.1 pavyzdys.** Kompiuteriu sugeneruota $n = 80$ atsitiktinių skaičių. Gautieji rezultatai:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0,0100 | 0,0150 | 0,0155 | 0,0310 | 0,0419 | 0,0456 | 0,0880 | 0,1200 | 0,1229 |
| 0,1279 | 0,1444 | 0,1456 | 0,1621 | 0,1672 | 0,1809 | 0,1855 | 0,1882 | 0,1917 |
| 0,2277 | 0,2442 | 0,2456 | 0,2476 | 0,2538 | 0,2552 | 0,2681 | 0,3041 | 0,3128 |
| 0,3810 | 0,3832 | 0,3969 | 0,4050 | 0,4182 | 0,4259 | 0,4365 | 0,4378 | 0,4434 |
| 0,4482 | 0,4515 | 0,4628 | 0,4637 | 0,4668 | 0,4773 | 0,4799 | 0,5100 | 0,5309 |
| 0,5391 | 0,6033 | 0,6283 | 0,6468 | 0,6519 | 0,6686 | 0,6689 | 0,6865 | 0,6961 |
| 0,7058 | 0,7305 | 0,7337 | 0,7339 | 0,7440 | 0,7485 | 0,7516 | 0,7607 | 0,7679 |
| 0,7765 | 0,7846 | 0,8153 | 0,8445 | 0,8654 | 0,8700 | 0,8732 | 0,8847 | 0,8935 |
| 0,8987 | 0,9070 | 0,9284 | 0,9308 | 0,9464 | 0,9658 | 0,9728 | 0,9872 | |

Ar šie duomenys neprieštarauja prielaidai, kad tai yra paprastosios imties, gautos stebint a. d. $X \sim U(0,\ 1)$, realizacija?

Sudalinkime intervalą $(0,\ 1)$ į 5 vienodo ilgio intervalus: $[0;\ 0,2]$, $(0,2;\ 0,4]$, $(0,4;\ 0,6]$, $(0,6;\ 0,8]$, $(0,8\,;1]$.

Gauname vektoriaus $\boldsymbol{U}$ realizaciją: 18; 12; 16; 19; 15. Tikriname hipotezę $H_0' : \pi_i = 0,2$, $i = 1, ..., 5$. Gauname

$$X_n^2 = \frac{1}{n} \sum_{i=1}^{k} \frac{U_i^2}{\pi_{i0}} - n = \frac{18^2 + 12^2 + 16^2 + 19^2 + 15^2}{80 \cdot 0,2} - 80 = 1,875.$$

Asimptotinė P reikšmė $pv_a = \mathbf{P}\{\chi_4^2 > 1,875\} = 0,7587$. Atmesti hipotezę $H_0'$ nėra pagrindo. Tikėtinumų santykio kriterijus leidžia gauti tą patį atsakymą, nes statistikos $R_n$ realizacija yra 1,93 ir $pv_a = \mathbf{P}\{\chi_4^2 > 1,93\} = 0,7486$.

Paprastąją suderinamumo hipotezę dažnai tenka tikrinti, norint įsitikinti, ar atsitiktiniai kampai (arba, ekvivalenčiai, taškai ant apskritimo) yra pasiskirstę tolygiai. Apie atsitiktinių kampų skirstinius ir jų taikymą žr. I dalies 3.7.15 ir 4.7.12 skyrelius ir monografiją [21].

**2.1.2 pavyzdys.** Lentelėje pateikti duomenys apie užregistruotus susirgimo leukemija atvejus Anglijoje per 1946 – 60 metų laikotarpį sugrupuoti mėnesiniais intervalais.(žr.[21])

| Mėnuo | Susirgo | Mėnuo | Susirgo | Mėnuo | Susirgo |
|---|---|---|---|---|---|
| Sausis | 39 | Gegužė | 38 | Rugsėjis | 37 |
| Vasaris | 37 | Birželis | 59 | Spalis | 47 |
| Kovas | 29 | Liepa | 50 | Lapkritis | 34 |
| Balandis | 45 | Rugpjūtis | 54 | Gruodis | 37 |

Perveskime duomenis į kampų stebėjimus sutapatindami metų intervalą su intervalu $(0,\ 2\pi]$, t. y. sausis atitinka sektorių nuo $0°$ iki $30°$; vasaris – sektorių nuo $30°$ iki $60°$ ir t. t. Patikrinkime prielaidą, kad kampai pasiskirstę tolygiai su tankiu $1/2\pi$. Duomenys jau sugrupuoti į vienodo ilgio intervalus. Taigi tikrinsime hipotezę $H_0' : \pi_{10} = \pi_{20} = ... = \pi_{k0} = 1/12$. Randame statistikų $R_n$ ir $X_n^2$ realizacijas: $R_n = 20,0797$, $X_n^2 = 20,4822$. Asimptotinės $P$ reikšmės: $pv_a = \mathbf{P}\{\chi^2_{11} > 20,0797\} = 0,0443$; $pv_a = \mathbf{P}\{\chi^2_{11} > 20,4822\} = 0,0391$. Abu kriterijai atmeta tolygumo hipotezę, kai kriterijaus reikšmingumo lygmuo viršija 0,0443.

**2.1.6 pastaba.** Kriterijai (2.1.8),(2.1.10) nebūtinai susiję su paprastosios suderinamumo hipotezės $H_0$ tikrinimu. Gali reikėti tiesiog patikrinti hipotezę $H_0'$ apie polinominio skirstinio parametrų reikšmes (žr. pateikiamą pavyzdį).

**2.1.3 pavyzdys.** Nustatyta, kad gamyklos tam tikro ilgo laikotarpio produkcijos 0,35 dalį sudaro pirmosios rūšies gaminiai; 0,6 dalį – antrosios rūšies, o likusią 0,05 dalį sudaro brokas. Patikrinus 300 gaminių partiją surasta 115 gaminių pirmosios rūšies, 165 – antrosios ir 20 – su defektais. Ar galima daryti išvadą, kad gaminių kokybė nepakito?

Šiame pavyzdyje $U_1 = 115$, $U_2 = 165$, $U_3 = 20$, $n = 300$ ir reikia patikrinti hipotezę $H_0' : \pi_1 = 0,35$, $\pi_2 = 0,60$, $\pi_3 = 0,05$. Randame statistikų (2.1.5) ir (2.1.9) reikšmes

$$X_n^2 = 3,869, \quad R_n = 3,717.$$

Laisvės laipsnių skaičius yra $k - 1 = 3 - 1 = 2$.

Kadangi $\mathbf{P}\{\chi^2_2 > 3,869\} = 0,1445$ ir $\mathbf{P}\{\chi^2_2 > 3,717\} = 0,1559$, tai atmesti hipotezę $H_0'$ nėra pagrindo.

## 2.2. Pirsono suderinamumo kriterijus: sudėtinė hipotezė

Sakykime, $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis, gauta stebint a. d. $X$, kurio pasiskirstymo funkcija $F(x)$ priklauso šeimai $\mathcal{F}$.

**Sudėtinė hipotezė**

$$H_0 : F(x) \in \mathcal{F}_0 = \{F(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \subset \mathcal{F}, \tag{2.2.1}$$

kad stebimojo a. d. $X$ pasiskirstymo funkcija priklauso aibei $\mathcal{F}_0$, kuri sudaryta iš žinomos funkcinės išraiškos pasiskirstymo funkcijų $F(x; \boldsymbol{\theta})$, priklausančių nuo nežinomo $s$-mačio parametro $\boldsymbol{\theta} = (\theta_1, ..., \theta_s)^T \in \boldsymbol{\Theta} \subset \mathbf{R}^s$.

Pavyzdžiui, tikriname hipotezę, kad stebimojo a. d. $X$ skirstinys priklauso normaliųjų, eksponentinių, Puasono, binominių ar kitų skirstinių šeimai.

Kaip ir pirmesniame poskyryje, sudalinkime abscisių ašį į $k > s+1$ intervalų ir tegu $U_j$ reiškia stebėjimų, patekusių į $j$-ąjį intervalą, skaičių, $j = 1, 2, ..., k$.

Grupuotoji imtis $\boldsymbol{U} = (U_1, ..., U_k)^T$ turi $k$-matį polinominį skirstinį $\mathcal{P}_k(n, \boldsymbol{\pi})$, čia

$$\boldsymbol{\pi} = (\pi_1, ..., \pi_k)^T,$$
$$\pi_i = \mathbf{P}\{X \in (a_{i-1}, a_i]\} = F(a_i) - F(a_{i-1}), \quad F \in \mathcal{F}.$$

Jeigu hipotezė $H_0$ teisinga, tai teisinga ir hipotezė

$$H_0' : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

čia

$$\boldsymbol{\pi}(\boldsymbol{\theta}) = (\pi_1(\boldsymbol{\theta}), ..., \pi_k(\boldsymbol{\theta}))^T, \quad \pi_i(\boldsymbol{\theta}) = F_0(a_i; \boldsymbol{\theta}) - F_0(a_{i-1}; \boldsymbol{\theta}). \tag{2.2.2}$$

Taigi hipotezėje $H_0'$ tvirtinama, kad vektoriaus $\boldsymbol{U}$ polinominio skirstinio tikimybes galima išreikšti pavidalo (2.2.2) funkcijomis nuo parametrų $\theta_1, ..., \theta_s, \quad s + 1 < k$.

Kadangi parametras $\boldsymbol{\theta}$ nežinomas, tai negalima apskaičiuoti Pirsono statistikos (2.1.5)

$$X_n^2(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{(U_i - n\pi_i(\boldsymbol{\theta}))^2}{n\pi_i(\boldsymbol{\theta})} = \frac{1}{n}\sum_{i=1}^{k} \frac{U_i^2}{\pi_i(\boldsymbol{\theta})} - n. \tag{2.2.3}$$

Natūralu šioje išraiškoje pakeisti nežinomus parametrus tam tikrais įvertiniais ir išnagrinėti gautų tokiu būdu statistikų savybes. Pasirodo, kad įrašius į (2.2.3) parametro $\boldsymbol{\theta}$ DT įvertinį, sudarytą iš *negrupuotų duomenų*, gautos statistikos skirstinys priklauso nuo $F_0(x; \boldsymbol{\theta})$ (ir nuo parametro $\boldsymbol{\theta}$). Jeigu parametro $\boldsymbol{\theta}$ įvertinys randamas remiantis mažiau informatyvia *grupuotąja* imtimi $\boldsymbol{U} = (U_1, ..., U_k)^T$, tai gautosios statistikos asimptotinis skirstinys yra chi kvadrato skirstinys su $k - 1 - s$ laisvės laipsnių ir nepriklauso nuo nežinomo parametro $\boldsymbol{\theta}$.

Pateiksime keletą tokio tipo įvertinių.

1) Kai hipotezė $H_0$ teisinga, imties $\boldsymbol{U}$ tikėtinumo funkcija ir jos logaritmas yra

$$\tilde{L}(\boldsymbol{\theta}) = \frac{n!}{U_1!...U_k!}\prod_{i=1}^{k} \pi_i^{U_i}(\boldsymbol{\theta}), \quad \tilde{\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{k} U_i \ln \pi_i(\boldsymbol{\theta}) + C. \tag{2.2.4}$$

Parametro $\boldsymbol{\theta}$ grupuotosios imties DT įvertinį $\boldsymbol{\theta}_n^*$ gauname spręsdami lygčių sistemą

$$\frac{\partial \tilde{\ell}}{\partial \theta_j} = \sum_{i=1}^{k} \frac{U_i}{\pi_i(\boldsymbol{\theta})}\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, 2, ..., s. \tag{2.2.5}$$

Pakeitę išraiškoje (2.2.3) nežinomą parametrą $\boldsymbol{\theta}$ įvertiniu $\boldsymbol{\theta}_n^*$, gausime statistiką

$$X_n^2(\boldsymbol{\theta}_n^*) = \sum_{j=1}^{k} \frac{(U_j - n\pi_j(\boldsymbol{\theta}_n^*))^2}{n\pi_j(\boldsymbol{\theta}_n^*)}. \tag{2.2.6}$$

2) Kitas būdas rasti įvertinį – minimizuoti kvadratinę formą (2.2.3), t. y. rasti įvertinį $\tilde{\boldsymbol{\theta}}_n$ iš sąlygos

$$X_n^2(\tilde{\boldsymbol{\theta}}_n) = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} X_n^2(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{k} \frac{(U_i - n\pi_i(\boldsymbol{\theta}))^2}{n\pi_i(\boldsymbol{\theta})}. \tag{2.2.7}$$

Šis įvertinių radimo būdas vadinamas *chi kvadrato minimumo metodu.*

3) Ieškant įvertinio $\tilde{\boldsymbol{\theta}}_n$ gaunamos gana sudėtingos lygčių sistemos, todėl kartais išraiška (2.2.3) supaprastinama pakeičiant vardiklį į $U_i$ ir paskui ją minimizuojant. Tai vadinamasis *modifikuotas chi kvadrato minimumo* metodas

įvertiniams rasti. Pažymėję šiuo metodu gautą įvertinį $\bar{\boldsymbol{\theta}}_n$, gauname statistiką

$$X_n^2(\bar{\boldsymbol{\theta}}_n) = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^k \frac{(U_i - n\pi_i(\boldsymbol{\theta}))^2}{U_i}. \tag{2.2.8}$$

Gavome tris chi kvadrato tipo statistikas (2.2.6)–(2.2.8).

4) Apibrėžkime *tikėtinumų santykio statistiką* pagal grupuotąją imtį

$$R_n = -2\ln\frac{\sup_{\boldsymbol{\theta}\in\Theta}\tilde{L}}{\sup_{\boldsymbol{\pi}} L(\boldsymbol{\pi})} = -2\ln\frac{\sup_{\boldsymbol{\theta}\in\Theta}\prod_{i=1}^k \pi_i^{U_i}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\pi}}\prod_{i=1}^k \pi_i^{U_i}}$$

$$= 2\sum_{i=1}^k U_i \ln\frac{U_i}{n\pi_i(\boldsymbol{\theta}_n^*)}.$$

Šią statistiką galima užrašyti tokiu pavidalu:

$$R_n = R_n(\boldsymbol{\theta}_n^*) = \inf_{\boldsymbol{\theta}\in\Theta} R_n(\boldsymbol{\theta}), \quad R_n(\boldsymbol{\theta}) = 2\sum_{i=1}^k U_i \ln\frac{U_i}{n\pi_i(\boldsymbol{\theta})}. \tag{2.2.9}$$

Įrodysime, kad statistikos $X_n^2(\boldsymbol{\theta}_n^*)$, $X_n^2(\tilde{\boldsymbol{\theta}}_n)$, $X_n^2(\bar{\boldsymbol{\theta}}_n)$ ir $R_n(\boldsymbol{\theta}_n^*)$ yra asimptotiškai ekvivalenčios, kai $n \to \infty$.

Tarkime, $\{Y_n\}$ yra atsitiktinių dydžių seka. Žymėsime $Y_n = o_P(1)$, jeigu $Y_n \xrightarrow{P} 0$, ir žymėsime $Y_n = O_P(1)$, jeigu

$$\forall\varepsilon > 0 \quad \exists c > 0 : \sup_n \mathbf{P}\{|Y_n| > c\} < \varepsilon.$$

Prochorovo teorema (žr. [30]) duoda pakankamą sąlygą: jeigu egzistuoja a. d. $Y$, kad $Y_n \xrightarrow{d} Y$, kai $n \to \infty$, tai $Y_n = O_P(1)$.

**Sąlygos A:**

**1)** su visais $i = 1, \ldots, k$ ir visais $\boldsymbol{\theta} \in \Theta$

$$0 < \pi_i(\boldsymbol{\theta}) < 1, \quad \pi_1(\boldsymbol{\theta}) + \cdots + \pi_k(\boldsymbol{\theta}) = 1.$$

**2)** funkcijos $\pi_i(\boldsymbol{\theta})$ turi tolydžias pirmos ir antros eilės dalines išvestines aibėje $\Theta$;

**3)** matricos

$$\mathbf{B} = \left[\frac{\partial\pi_i(\boldsymbol{\theta})}{\partial\theta_j}\right]_{k\times s}, \quad i = 1, ..., k, \quad j = 1, ..., s,$$

rangas lygus $s$.

**2.2.1 Lema.** *Tarkime, hipotezė $H_0'$ teisinga ir įvykdytos sąlygos* **A**. *Tada įvertiniai $\tilde{\pi}_{in} = \pi_i(\tilde{\boldsymbol{\theta}}_n)$, $\pi_{in}^* = \pi_i(\boldsymbol{\theta}_n^*)$ ir $\bar{\pi}_{in} = \pi_i(\bar{\boldsymbol{\theta}}_n)$ yra $\sqrt{n}$ pagrįsti, t. y.*

*$\sqrt{n}(\tilde{\pi}_{in} - \pi_i) = O_P(1)$, $\sqrt{n}(\pi^*_{in} - \pi_i) = O_p(1)$ ir $\sqrt{n}(\bar{\pi}_{in} - \pi_i) = O_P(1)$, čia $\pi_i = \pi_i(\boldsymbol{\theta})$ yra tikroji tikimybės reikšmė.*

**Įrodymas.** Nagrinėkime įvertinį $\tilde{\pi}_{in}$. Kadangi $0 \leq \tilde{\pi}_{in} \leq 1$, tai $\tilde{\pi}_{in} = O_P(1)$. Reikia pažymėti, kad $U_i/n - \pi_i = o_P(1)$. Todėl iš nelygybių (naudojamės chi kvadrato minimumo įvertinio apibrėžimu)

$$\sum_{i=1}^{k} \frac{(U_i/n - \tilde{\pi}_{in})^2}{\tilde{\pi}_{in}} \leq \sum_{i=1}^{k} \frac{(U_i/n - \pi_i)^2}{\pi_i} = o_P(1)$$

išplaukia, kad su visais $i$: $U_i/n - \tilde{\pi}_{in} = o_P(1)$, taigi ir

$$\tilde{\pi}_{in} - \pi_i = (\tilde{\pi}_{in} - U_i/n) + (U_i/n - \pi_i) = o_P(1).$$

Kadangi $\sqrt{n}(U_i/n - \pi_i) \xrightarrow{d} Z_i \sim N(0, \pi_i(1-\pi_i))$, tai $\sqrt{n}(U_i/n - \pi_i) = O_P(1)$. Todėl iš nelygybių

$$\sum_{i=1}^{k} \frac{(U_i - n\tilde{\pi}_{in})^2}{n\tilde{\pi}_{in}} \leq \sum_{i=1}^{k} \frac{(U_i - n\pi_i)^2}{n\pi_i} = O_P(1)$$

išeina, kad su visais $i$: $(U_i - n\tilde{\pi}_{in})/\sqrt{n} = O_P(1)$, ir

$$\sqrt{n}(\tilde{\pi}_{in} - \pi_i) = \frac{n\tilde{\pi}_{in} - n\pi_i}{\sqrt{n}} = \frac{n\tilde{\pi}_{in} - U_i}{\sqrt{n}} + \frac{U_i - n\pi_i}{\sqrt{n}} = O_P(1).$$

Analogiškai gauname

$$\sum_{i=1}^{k} \frac{(U_i - n\bar{\pi}_{in})^2}{U_i} \leq \sum_{i=1}^{k} \frac{(U_i - n\pi_i)^2}{U_i} = O_P(1)$$

ir

$$\sqrt{n}(\bar{\pi}_{in} - \pi_i) = \frac{n\bar{\pi}_{in} - U_i}{\sqrt{n}} + \frac{U_i - n\pi_i}{\sqrt{n}} = O_P(1).$$

Nagrinėkime įvertinį $\pi^*_{in} = \pi_i(\boldsymbol{\theta}^*_n)$. Kai tenkinamos sąlygos **A**, a. v. $\sqrt{n}(\boldsymbol{\theta}^*_n - \boldsymbol{\theta})$ asimptotiškai turi normalųjį skirstinį (žr. A priedą). Remiantis delta metodu [30]

$$\begin{aligned}\sqrt{n}(\pi^*_{in} - \pi_i) &= \sqrt{n}(\pi_i(\boldsymbol{\theta}^*_n) - \pi_i(\boldsymbol{\theta})) \\ &= \sqrt{n}\dot{\pi}_i^T(\boldsymbol{\theta})(\boldsymbol{\theta}^*_n - \boldsymbol{\theta}) + o_P(1) = O_P(1).\end{aligned}$$

▲

**2.2.1 teorema.** *Jeigu hipotezė $H_0'$ teisinga ir įvykdytos sąlygos* **A**, *tai statistikos $X_n^2(\tilde{\boldsymbol{\theta}}_n)$, $X_n^2(\bar{\boldsymbol{\theta}}_n)$, $X_n^2(\boldsymbol{\theta}^*_n)$ ir $R_n(\boldsymbol{\theta}^*_n)$ asimptotiškai ekvivalenčios ($n \to \infty$):*

$$X_n^2(\tilde{\boldsymbol{\theta}}_n) = X_n^2(\bar{\boldsymbol{\theta}}_n) + o_P(1) = X_n^2(\boldsymbol{\theta}^*_n) + o_P(1) = R_n(\boldsymbol{\theta}^*_n) + o_P(1).$$

*Kiekvienos iš šių statistikų skirstinys konverguoja į chi kvadrato skirstinį su $k - s - 1$ laisvės laipsnių.*

**Įrodymas**. Imkime bet kokį $\boldsymbol{\theta}$ įvertinį $\hat{\boldsymbol{\theta}}$, kuriam

$$\hat{\boldsymbol{\pi}}_n = (\hat{\pi}_{1n}, \dots, \hat{\pi}_{kn}) = (\pi_1(\hat{\boldsymbol{\theta}}), \dots, \pi_k(\hat{\boldsymbol{\theta}}))$$

būtų $\sqrt{n}$ pagrįstas parametro $\boldsymbol{\pi}$ įvertinys. Remdamiesi $\sqrt{n}$ pagrįstumo apibrėžimu ir konvergavimu $U_i/n \xrightarrow{P} \pi_i$ gauname, kad su visais $i$

$$\hat{\pi}_{in} - \frac{U_i}{n} = o_P(1), \quad \sqrt{n}\left(\hat{\pi}_{in} - \frac{U_i}{n}\right) = O_P(1), \quad \frac{U_i}{n} = O_P(1).$$

Naudodami paskutines lygybes, Teiloro skleidinį

$$\ln(1+x) = x - x^2/2 + o(x^2), \quad x \to 0,$$

ir lygybę $U_1 + ... + U_k = n$, gauname

$$\frac{1}{2}R_n(\hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^k U_i \log\frac{U_i}{n\hat{\pi}_i} = -\sum_{i=1}^k U_i \log\left(1 + \frac{n\hat{\pi}_i}{U_i} - 1\right)$$

$$= -\sum_{i=1}^k U_i \log\left(1 + \frac{\hat{\pi}_i - U_i/n}{U_i/n}\right) = -\sum_{i=1}^k U_i \left(\frac{\hat{\pi}_i - U_i/n}{U_i/n}\right)$$

$$+\frac{1}{2}\sum_{i=1}^k U_i \left(\frac{\hat{\pi}_i - U_i/n}{U_i/n}\right)^2 + \sum_{i=1}^k U_i o_P\left(\left(\frac{\hat{\pi}_i - U_i/n}{U_i/n}\right)^2\right)$$

$$= -n\sum_{i=1}^k \hat{\pi}_i + \sum_{i=1}^k U_i + \frac{1}{2}\sum_{i=1}^k \frac{(U_i - n\hat{\pi}_i)^2}{U_i} + o_P(1)$$

$$= \frac{1}{2}\sum_{i=1}^k \frac{(U_i - n\hat{\pi}_i)^2}{U_i} + o_P(1)$$

$$= \frac{1}{2}\sum_{i=1}^k \frac{(U_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} - \frac{1}{2}\sum_{i=1}^k \frac{(U_i - n\hat{\pi}_i)^3}{U_i n\hat{\pi}_i} + o_P(1)$$

$$= \frac{1}{2}\sum_{i=1}^k \frac{(U_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} + o_P(1) = \frac{1}{2}X_n^2(\hat{\boldsymbol{\theta}}_n) + o_P(1).$$

Imant $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ ir $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_n$, gaunama

$$X_n^2(\boldsymbol{\theta}_n^*) = R_n(\boldsymbol{\theta}_n^*) + o_P(1), \quad X_n^2(\tilde{\boldsymbol{\theta}}_n) = R_n(\tilde{\boldsymbol{\theta}}_n) + o_P(1).$$

Remiantis $\tilde{\boldsymbol{\theta}}_n$ apibrėžimu gaunama $X_n^2(\tilde{\boldsymbol{\theta}}_n) \leq X_n^2(\boldsymbol{\theta}_n^*)$, o remiantis $R_n$ apibrėžimu (2.2.9) gaunama $R_n(\boldsymbol{\theta}_n^*) \leq R_n(\tilde{\boldsymbol{\theta}}_n)$. Taigi

$$X_n^2(\tilde{\boldsymbol{\theta}}_n) \leq X^2(\boldsymbol{\theta}_n^*) = R_n(\boldsymbol{\theta}_n^*) + o_P(1) \leq R_n(\tilde{\boldsymbol{\theta}}_n) + o_P(1) = X_n^2(\tilde{\boldsymbol{\theta}}_n) + o_P(1).$$

Tokiu būdu $X^2(\tilde{\boldsymbol{\theta}}_n) = R_n(\boldsymbol{\theta}_n^*) + o_P(1)$. Analogiškai gauname $X_n^2(\bar{\boldsymbol{\theta}}_n) = R_n(\boldsymbol{\theta}_n^*) + o_P(1)$.

Kadangi $k-1$-matis vektorius $(\pi_1, ..., \pi_{k-1})^T$ yra $s$-mačio parametro $\boldsymbol{\theta}$ funkcija, tai tikėtinumų santykio statistikos $R_n(\boldsymbol{\theta}_n^*)$ ribinis dėsnis yra chi kvadrato skirstinys su $k-s-1$ laisvės laipsnių (žr. A priedą, 6.1.3 pastabą). Tokį patį ribinį dėsnį turi ir kitos nagrinėtos statistikos. ▲

**Chi kvadrato kriterijus:** Hipotezė $H_0'$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai teisinga nelygybė

$$X_n^2(\hat{\boldsymbol{\theta}}_n) > \chi_\alpha^2(k-1-s), \tag{2.2.10}$$

čia $\hat{\boldsymbol{\theta}}_n$ yra bet kuris iš įvertinių $\boldsymbol{\theta}_n^*, \tilde{\boldsymbol{\theta}}_n, \bar{\boldsymbol{\theta}}_n$.

**Tikėtinumų santykio kriterijus:** Hipotezė $H_0'$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai teisinga nelygybė

$$R_n(\boldsymbol{\theta}_n^*) > \chi_\alpha^2(k-1-s). \tag{2.2.11}$$

Jeigu hipotezė $H_0'$ atmetama, tai atmetama ir suderinamumo hipotezė $H_0$.

**2.2.1 pavyzdys.** Turimi gaminių patikimumo duomenys, sugrupuoti į intervalus $(a_{i-1}, a_i]$, $i = 1, ..., 11$, kurie pateikti lentelėje

| $i$ | $(a_{i-1}, a_i]$ | $U_i$ | $i$ | $(a_{i-1}, a_i]$ | $U_i$ |
|---|---|---|---|---|---|
| 1 | $(0, 100]$ | 8 | 7 | $(600, 700]$ | 25 |
| 2 | $(100, 200]$ | 12 | 8 | $(700, 800]$ | 18 |
| 3 | $(200, 300]$ | 19 | 9 | $(800, 900]$ | 15 |
| 4 | $(300, 400]$ | 23 | 10 | $(900, 1000]$ | 14 |
| 5 | $(400, 500]$ | 29 | 11 | $(1000, \infty)$ | 18 |
| 6 | $(500, 600]$ | 30 | | | |

Tikrinsime hipotezę, kad gaminių darbo laikas aprašomas Veibulo skirstiniu.
Pagal (2.2.9) įvertinys $(\theta_n^*, \nu_n^*)$ minimizuoja funkciją

$$R_n(\theta, \nu) = 2\sum_{i=1}^k U_i \ln \frac{U_i}{n\pi_i(\theta, \nu)}, \quad \pi_i(\theta, \nu) = e^{-(a_{i-1}/\theta)^\nu} - e^{-(a_i/\theta)^\nu}.$$

Diferencijuodami parametrų $\theta$ ir $\nu$ atžvilgiu ir prilyginę išvestines nuliui, įvertiniams $\theta_n^*$ ir $\nu^*$ rasti gauname lygčių sistemą

$$\sum_{i=1}^k U_i \frac{a_{i-1}^\nu e^{-(a_{i-1}/\theta)^\nu} - a_i^\nu e^{-(a_{i-1}/\theta)^\nu}}{e^{-(a_{i-1}/\theta)^\nu} - e^{-(a_i/\theta)^\nu}} = 0,$$

$$\sum_{i=1}^k U_i \frac{a_{i-1}^\nu e^{-(a_{i-1}/\theta)^\nu} \ln a_{i-1} - a_i^\nu e^{-(a_{i-1}/\theta)^\nu} \ln a_i}{e^{-(a_{i-1}/\theta)^\nu} - e^{-(a_i/\theta)^\nu}} = 0.$$

Išsprendę šią lygčių sistemą arba tiesiog minimizuodami funkciją $R_n(\theta, \nu)$ gauname įverčius $\theta^* = 649,516$ ir $\nu^* = 2,004$. Minimizuodami (2.2.7) ir (2.2.8) gauname įverčius $\tilde{\theta} = 647,380, \tilde{\nu} = 1,979$ ir $\bar{\theta} = 653,675, \bar{\nu} = 2,052$. Naudodami šiuos įverčius gauname statistikų realizacijas

$$R_n(\theta^*, \nu^*) = 4,047; \; X_n^2(\theta^*, \nu^*) = 4,377; \; X_n^2(\tilde{\theta}, \tilde{\nu}) = 4,324, \; X_n^2(\bar{\theta}, \bar{\nu}) = 3,479.$$

Laisvės laipsnių skaičius $k-s-1=8$. Asimptotinės $P$ reikšmės atitinkamai yra 0,853; 0,822; 0,827; 0,901. Atmesti hipotezę nėra pagrindo.

**2.2.2 pavyzdys.** Lentelėje pateikti azimutai tų horizonto taškų, kuriuos stebėtojas užfiksavo paskutinį kartą matydamas paleistą antį (prapuolimo kampas). Eksperimento metu paleista $n = 714$ ančių. Eksperimentas atliktas Anglijoje Gločesterio grafystėje (žr. [21]).

| $\varphi_i^\circ$ | $V_i$ | $\hat{V}_i$ | $\varphi_i^\circ$ | $V_i$ | $\hat{V}_i$ | $\varphi_i^\circ$ | $V_i$ | $\hat{V}_i$ |
|---|---|---|---|---|---|---|---|---|
| 10° | 40 | 45.48 | 130° | 3 | 2,09 | 250° | 24 | 48,71 |
| 30° | 22 | 23,17 | 150° | 1 | 2,40 | 270° | 58 | 83,04 |
| 50° | 20 | 11,30 | 170° | 6 | 3,51 | 290° | 136 | 116,72 |
| 70° | 9 | 5,76 | 190° | 3 | 6,18 | 310° | 138 | 130,11 |
| 90° | 6 | 3,33 | 210° | 11 | 12,23 | 330° | 143 | 113,63 |
| 110° | 23 | 2,34 | 230° | 22 | 25,06 | 350° | 69 | 78,94 |

Duomenys sugrupuoti į 20° ilgio intervalus. Lentelėje nurodytas vidurinis i-ojo intervalo kampas $\varphi_i^\circ$ ir stebėjimų, patekusių į i-tąjį intervalą, dažnis $V_i$, $i = 1, ..., 18$.

Šį pavyzdį jau nagrinėjome I dalies 3.7.15 ir 4.7.12 skyreliuose. Hipotezė dėl kampų tolygaus pasiskirstymo atmetama su labai aukštu reikšmingumo lygmeniu. Atsakymas į klausimą, ar šie duomenys gali būti aprašyti Mizeso skirstiniu (žr. I dalies 3.7.15 skyrelį), buvo atidėtas iki III dalies.

Patikrinsime hipotezę, kad stebimas atsitiktinis kampas $\varphi$ turi Mizeso skirstinį $M(\mu, \theta)$. Parametrų $\mu$ ir $\theta$ DT įverčiai remiantis pateiktos lentelės grupuotais duomenimis surasti I dalies 3.7.15 skyrelyje: $\hat{\mu} = 308,9^\circ, \hat{\theta} = 2,077$. Palyginti lentelėje pateikti tikėtini dažniai $\hat{V}_i = n\pi_i(\hat{\mu}, \hat{\theta})$. Apskaičiuojame statistikų $R_n(\hat{\mu}, \hat{\theta})$ ir $X_n^2(\hat{\mu}, \hat{\theta})$ reikšmes: 50,7324 ir 49,2718. Atitinkamos asimptotinės $P$ reikšmės yra $0,9 \cdot 10^{-5}$ ir $1,5 \cdot 10^{-5}$. Hipotezė atmetama.

**2.2.3 pavyzdys** (**2.1.2** pavyzdžio tęsinys). 2.1.2 pavyzdžio sąlygomis patikrinsime hipotezę, kad sergamumo leukemija momentų pasiskirstymą galima aprašyti Mizeso skirstiniu $M(\mu, \theta)$. Parametrų $\mu$ ir $\theta$ DT įverčiai remiantis pateiktos lentelės grupuotais duomenimis yra tokie: $\hat{\mu} = 198,34^\circ, \hat{\theta} = 0,2021$. Su šiais įverčiais tikėtini patekimo į intervalus dažniai $\hat{V}_i = n\pi_i(\hat{\mu}, \hat{\theta})$ yra tokie: 34,2; 34,9; 37,4; 41,3; 45,7; 49,3; 51,0; 49,9; 46,6; 42,2; 38,2; 35,3. Apskaičiuojame statistikų $R_n(\hat{\mu}, \hat{\theta})$ ir $X_n^2(\hat{\mu}, \hat{\theta})$ reikšmes: 9,8717 ir 9,6457. Atitinkamos asimptotinės $P$ reikšmės yra $0,3610$ ir $0,3797$. Hipotezę atmesti nėra pagrindo. Galima daryti išvadą, kad turimi duomenys prieštarauja prielaidai apie tolygų susirgimų leukemija pasiskirstymą, tačiau gerai aprašomi Mizeso skirstiniu. Tankio įvertis unimodalus su moda taške $\varphi = 198,34^\circ$ (liepos mėnesio vidurys) ir antimoda taške $\varphi = 18,34^\circ$ (sausio vidurys); tankio įvertis modos taške 1,5 karto didesnis už tankio įvertį antimodos taške.

# 2.3. Modifikuotasis chi kvadrato kriterijus

Nors chi kvadrato kriterijus yra universalus ir dažnai taikomas tikrinant suderinamumo hipotezes, tačiau jis, ypač kai skirstinys tolydusis, turi tam tikrų trūkumų.

Pirma, griežtai nenurodoma, kaip parinkti grupavimo intervalo galus, todėl gautosios išvados iš dalies priklauso nuo grupavimo intervalų parinkimo. Be to, nagrinėjant statistikų asimptotinius skirstinius 2.2.1 teoremoje buvo laikoma, kad grupavimo intervalai parinkti neatsižvelgiant į imtį. Tačiau praktiškai grupavimo intervalus parenkame atsižvelgdami į imties rezultatus. Todėl, formaliai žiūrint, naudotis teoremos 2.2.1 rezultatais negalima.

Antra, kad būtų galima apskaičiuoti 2.2.1 teoremos statistikų reikšmes, reikia išspręsti gana sudėtingas lygčių sistemas parametrui $\boldsymbol{\theta}$ įvertinti pagal grupuotus duomenis. Gautieji įvertiniai nėra optimalūs, nes naudoja mažiau informatyvią grupuotąją imtį.

Trečia, asimptotiškai optimaliais įvertiniais (kai išpildytos reguliarumo sąlygos tai DT įvertiniai, gauti pagal pradinius nesugrupuotus duomenis) naudotis negalime, nes tada chi kvadrato statistikos skirstinys nebus toks, kaip nurodyta 2.2.1 teoremoje. Asimptotinis skirstinys priklauso nuo skirstinio pavidalo ir nežinomų parametrų. Todėl, netgi jei jau apskaičiuoti parametrų įvertiniai pagal pradinius duomenis, tai $\chi^2$ kriterijumi tikrinant suderinamumo hipotezę reikia vėl perskaičiuoti įvertinius pagal grupuotus duomenis.

Pateikiamas modifikuotas chi kvadrato kriterijus neturi minėtų trūkumų. Teoriniai rezultatai apie statistikos asimptotinį skirstinį gaunami tariant, kad parametrai vertinami DT metodu pagal pradinę negrupuotą imtį, o grupavimo intervalų galai tam tikru būdu priklauso nuo imties.

### 2.3.1. Bendras atvejis

Tikriname sudėtinę suderinamumo hipotezę (2.2.1). Nagrinėsime ribinį skirstinį atsitiktinio vektoriaus $\boldsymbol{Z}_n = (Z_{1n}, ..., Z_{kn})^T$

$$Z_{jn} = \frac{U_j - n\pi_j(\hat{\boldsymbol{\theta}}_n)}{\sqrt{n\pi_j(\hat{\boldsymbol{\theta}}_n)}} = \frac{\sqrt{n}(U_j/n - \pi_j(\hat{\boldsymbol{\theta}}_n))}{\sqrt{\pi_j(\hat{\boldsymbol{\theta}}_n)}}, \tag{2.3.1}$$

čia $\hat{\boldsymbol{\theta}}_n$ yra DT įvertinys, gautas pagal pradinę imtį $\mathbf{X} = (X_1, ..., X_n)^T$. Šis įvertinys maksimizuoja logtikėtinumo funkciją

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \ln f(X_i, \boldsymbol{\theta}).$$

Jeigu modelis $\{F_0(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ yra reguliarus (žr. A priedą), tai imties elemento $X_1$ Fišerio informacinė matrica yra $i(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}} \dot{\ell}_1(\boldsymbol{\theta}) \dot{\ell}_1^T(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}} \ddot{\ell}_1(\boldsymbol{\theta})$; čia $\dot{\ell}_1(\boldsymbol{\theta})$ yra $\ell_1(\boldsymbol{\theta})$ pirmųjų išvestinių pagal parametrus $\theta_1, ..., \theta_s$ vektorius, o $\ddot{\ell}_1(\boldsymbol{\theta})$ – antrųjų išvestinių matrica.

**2.3.1 teorema.** *Tarkime, modelis $\{F_0(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ yra reguliarus (A priedas, 6.1.1 teorema) ir įvykdytos* **A** *sąlygos. Tada*

$$\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})),$$

$$Y_n^2 = \boldsymbol{Z}_n^T \boldsymbol{\Sigma}^- \boldsymbol{Z}_n \xrightarrow{d} Y^2 \sim \chi^2(k-1), \tag{2.3.2}$$

*čia*

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = (\mathbf{E}_k - \boldsymbol{q}(\boldsymbol{\theta})\boldsymbol{q}^T(\boldsymbol{\theta}))(\mathbf{E}_k - \boldsymbol{C}^T(\boldsymbol{\theta})\boldsymbol{i}^{-1}(\boldsymbol{\theta})\boldsymbol{C}(\boldsymbol{\theta})),$$
$$\boldsymbol{\Sigma}^-(\boldsymbol{\theta}) = (\mathbf{E}_k - \boldsymbol{C}^T(\boldsymbol{\theta})\boldsymbol{i}^{-1}(\boldsymbol{\theta})\boldsymbol{C}(\boldsymbol{\theta}))^{-1};$$

*čia $\boldsymbol{\Sigma}^-$ yra matricos $\boldsymbol{\Sigma}$ apibendrintoji atvirkštinė matrica, t. y. matrica, tenkinanti sąlygą $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^-\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$; $\mathbf{E}_k$ – $k \times k$ vienetinė matrica; $\boldsymbol{i}(\boldsymbol{\theta})$ imties elemento*

$X_1$ *Fišerio informacinė matrica;*

$$\boldsymbol{C}(\boldsymbol{\theta}) = [c_{ij}(\boldsymbol{\theta})]_{s\times k} = \left[\frac{1}{\sqrt{\pi_j(\boldsymbol{\theta})}}\frac{\partial \pi_j(\boldsymbol{\theta})}{\partial \theta_i}\right]_{s\times k},$$

$$\boldsymbol{q}(\boldsymbol{\theta}) = (\sqrt{\pi_1(\boldsymbol{\theta})}, ..., \sqrt{\pi_k(\boldsymbol{\theta})})^T.$$

**Įrodymas.** Reguliariuose modeliuose DT įvertiniai tenkina sąryšį (A priedas, 7.1.1 teorema):

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \boldsymbol{i}^{-1}(\boldsymbol{\theta})\frac{1}{\sqrt{n}}\dot{\ell}(\boldsymbol{\theta}) + o_P(1). \tag{2.3.3}$$

Pagal delta metodą [30]

$$\sqrt{n}(\pi_j(\hat{\boldsymbol{\theta}}_n) - \pi_j(\boldsymbol{\theta})) = \dot{\pi}_j^T(\boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o_P(1).$$

Remdamiesi sąryšiu (2.3.1), formule (2.3.3) ir lygybe $\pi_j(\hat{\boldsymbol{\theta}}_n) = \pi_j(\boldsymbol{\theta}) + o_P(1)$ apibrėžiame

$$Z_{jn} = \left\{\sqrt{n}\left(\frac{U_j}{n} - \pi_j(\hat{\boldsymbol{\theta}}_n)\right) - \frac{\dot{\pi}_j^T(\boldsymbol{\theta})\boldsymbol{i}^{-1}(\boldsymbol{\theta}\dot{\ell}(\boldsymbol{\theta}))}{\sqrt{n}}\right\}/\sqrt{\pi_j(\boldsymbol{\theta})} + o_P(1) =:$$

$$Y_{jn} + o_P(1).$$

Paqal A priedo 7.1.1 teoremą atsitiktiniai vektoriai

$$(\sqrt{n}(\frac{U_j}{n} - \pi_1(\boldsymbol{\theta})), ..., \sqrt{n}(\frac{U_j}{n} - \pi_1(\boldsymbol{\theta}))), \quad \dot{\ell}(\boldsymbol{\theta}), \quad \boldsymbol{Z}_n$$

yra asimptotiškai normalieji. Rasime a.v. $\boldsymbol{Y}_n = (Y_{1n}, ..., Y_{kn})^T$ kovariacijų matricą. Kadangi

$$\mathbf{E}\left(\sqrt{n}(\frac{U_j}{n} - \pi_1(\boldsymbol{\theta}))\right) = \mathbf{E}(\dot{\ell}(\boldsymbol{\theta})) = 0,$$

$$\boldsymbol{V}\left(\frac{U_j}{\sqrt{n}}\right) = \pi_j(\boldsymbol{\theta})(1 - \pi_j(\boldsymbol{\theta})),$$

$$\mathbf{Cov}\left(\frac{U_j}{\sqrt{n}}, \frac{U_l}{\sqrt{n}}\right) = -\pi_j(\boldsymbol{\theta})\pi_l(\boldsymbol{\theta}), \quad \boldsymbol{V}\left(\frac{\dot{\ell}(\boldsymbol{\theta})}{\sqrt{n}}\right) = \boldsymbol{i}(\boldsymbol{\theta}),$$

$$\mathbf{Cov}\left(\frac{U_j}{\sqrt{n}}, \frac{\dot{\ell}(\boldsymbol{\theta})}{\sqrt{n}}\right) = \mathbf{E}(\mathbf{1}_{(a_{j-1},a_j]}(X_1)\dot{\ell}(\boldsymbol{\theta}))$$

$$= \int_{a_{j-1}}^{a_j} \dot{f}(x;\boldsymbol{\theta})dx = \dot{\pi}_j(\boldsymbol{\theta}),$$

gauname

$$\mathbf{E}(Y_{jn}) = 0, \quad \boldsymbol{V}(Y_{jn}) = 1 - \pi_j(\boldsymbol{\theta}) - \dot{\pi}_j^T(\boldsymbol{\theta})\boldsymbol{i}^{-1}(\boldsymbol{\theta})\dot{\pi}_j(\boldsymbol{\theta})/\pi_j(\boldsymbol{\theta}),$$

ir su visais $j \neq l$

$$\mathbf{Cov}\,(Y_{jn}, Y_{ln}) = -\sqrt{\pi_j(\boldsymbol{\theta})\pi_l(\boldsymbol{\theta})} - \dot{\pi}_j^T(\boldsymbol{\theta})\boldsymbol{i}^{-1}(\boldsymbol{\theta})\dot{\pi}_l(\boldsymbol{\theta})/\sqrt{\pi_j(\boldsymbol{\theta})\pi_l(\boldsymbol{\theta})}.$$

Gauname, kad $\boldsymbol{Z}_n \xrightarrow{d} \boldsymbol{Z} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, kai kovariacijų matrica (argumentą $\boldsymbol{\theta}$ praleidžiame)

$$\boldsymbol{\Sigma} = \mathbf{E}_k - \boldsymbol{q}\boldsymbol{q}^T - \boldsymbol{C}^T\boldsymbol{i}^{-1}\boldsymbol{C}.$$

Kadangi

$$\boldsymbol{C}\boldsymbol{q} = \left(\frac{\partial}{\partial\theta_1}\sum_{j=1}^k \pi_j(\boldsymbol{\theta}), ..., \frac{\partial}{\partial\theta_s}\sum_{j=1}^k \pi_j(\boldsymbol{\theta})\right)^T = (0, ..., 0)^T = \mathbf{0}_s$$

ir $\boldsymbol{q}^T\boldsymbol{q} = 1$, tai

$$\boldsymbol{\Sigma} = (\mathbf{E}_k - \boldsymbol{q}\boldsymbol{q}^T)(\mathbf{E}_k - \boldsymbol{C}^T\boldsymbol{i}^{-1}\boldsymbol{C}).$$

Matricos $\boldsymbol{\Sigma}$ rangas yra $k-1$ (žr. 2.4 pratimą), o jos apibendrintoji atvirkštinė yra

$$\boldsymbol{\Sigma}^- = (\mathbf{E}_k - \boldsymbol{C}^T\boldsymbol{i}^{-1}\boldsymbol{C})^{-1},$$

nes iš lygybės

$$(\mathbf{E}_k - \boldsymbol{q}\boldsymbol{q}^T)(\mathbf{E}_k - \boldsymbol{q}\boldsymbol{q}^T) = (\mathbf{E}_k - \boldsymbol{q}\boldsymbol{q}^T)$$

išplaukia lygybė $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^-\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$.

Kiti teoremos tvirtinimai gaunami remiantis šia daugiamačio normaliojo skirstinio savybe: jeigu $\boldsymbol{Y} \sim N_k(\mathbf{0}, \boldsymbol{\Sigma})$, tai $\boldsymbol{Y}^T\boldsymbol{\Sigma}^-\boldsymbol{Y} \sim \chi^2(r)$, čia $r$ yra matricos $\boldsymbol{\Sigma}$ rangas. ▲

**2.3.1 pastaba.** Jeigu matrica $\boldsymbol{G} = \boldsymbol{i} - \boldsymbol{C}\boldsymbol{C}^T$ nėra išsigimusi (ši sąlyga dažniausiai taikomiems skirstiniams yra išpildyta), tai matricos $\boldsymbol{\Sigma}$ apibendrintoji atvirkštinė

$$\boldsymbol{\Sigma}^- = \mathbf{E}_k + \boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C}.$$

Matome, kad ieškant $\boldsymbol{\Sigma}^-$ nėra reikalo apvertinėti dimensijos $k \times k$ matricos, o pakanka rasti atvirkštinę matricai $\boldsymbol{G}$, kurios dimensija yra $s \times s$ (paprastai $s = 1$ arba $s = 2$).

**Įrodymas**

$$(\mathbf{E}_k - \boldsymbol{C}^T\boldsymbol{i}^{-1}\boldsymbol{C})(\mathbf{E}_k + \boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C}) = \mathbf{E}_k + \boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C} - \boldsymbol{C}^T\boldsymbol{i}^{-1}\boldsymbol{C}$$

$$-\boldsymbol{C}^T\boldsymbol{i}^{-1}\boldsymbol{C}\boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C} = \mathbf{E}_k + \boldsymbol{C}^T\boldsymbol{i}^{-1}(\boldsymbol{i}\boldsymbol{G}^{-1} - \mathbf{E}_s - \boldsymbol{C}\boldsymbol{C}^T\boldsymbol{G}^{-1})\boldsymbol{C}$$

$$= \mathbf{E}_k + \boldsymbol{C}^T\boldsymbol{i}^{-1}(\boldsymbol{G}\boldsymbol{G}^{-1} - \mathbf{E}_s)\boldsymbol{C} = \mathbf{E}_k. \blacktriangle$$

Kriterijaus statistika turi pavidalą

$$Y_n^2 = \boldsymbol{Z}_n^T\boldsymbol{\Sigma}^-(\hat{\boldsymbol{\theta}})\boldsymbol{Z}_n.$$

**2.3.2 pastaba.** Remiantis 2.3.1 pastaba šią statistiką galima suvesti į tokią formą:

$$Y_n^2 = X_n^2 + \frac{1}{n}\boldsymbol{v}^T\boldsymbol{G}^{-1}\boldsymbol{v}\,; \tag{2.3.4}$$

čia

$$X_n^2 = \sum_{j=1}^{k}\frac{(U_j - n\pi_j(\hat{\boldsymbol{\theta}}_n))^2}{n\pi_j(\hat{\boldsymbol{\theta}}_n)} = \sum_{j=1}^{k}\frac{U_j^2}{n\pi_j(\hat{\boldsymbol{\theta}}_n)} - n, \tag{2.3.5}$$

$$\boldsymbol{v} = (v_1, ..., v_s)^T, \quad v_i = \sum_{j=1}^{k}\frac{U_j}{\pi_j(\hat{\boldsymbol{\theta}}_n)}\frac{\partial\pi_j(\hat{\boldsymbol{\theta}}_n)}{\partial\theta_i},$$

$$\boldsymbol{G} = [g_{rr'}]_{s\times s}, \quad g_{rr'} = i_{rr'} - \sum_{l=1}^{k}\frac{1}{\pi_l(\hat{\boldsymbol{\theta}}_n)}\frac{\partial\pi_l(\hat{\boldsymbol{\theta}}_n)}{\partial\theta_r}\frac{\partial\pi_l(\hat{\boldsymbol{\theta}}_n)}{\partial\theta_{r'}};$$

$i_{rr'}$ yra matricos $\boldsymbol{i}(\hat{\boldsymbol{\theta}}_n)$ elementas.

**Įrodymas**

$$Y_n^2 = \boldsymbol{Z}_n^T(\mathbf{E}_k + \boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C})\boldsymbol{Z}_n = \boldsymbol{Z}_n^T\boldsymbol{Z}_n$$
$$+\tilde{\boldsymbol{U}}^T\boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C}\tilde{\boldsymbol{U}} + n\boldsymbol{q}^T\boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C}\boldsymbol{q},$$

čia

$$\tilde{\boldsymbol{U}} = \left(U_1/\sqrt{n\pi_1(\hat{\boldsymbol{\theta}}_n)}, ..., U_k/\sqrt{n\pi_k(\hat{\boldsymbol{\theta}}_n)}\right)^T.$$

Pirmasis dėmuo $\boldsymbol{Z}_n^T\boldsymbol{Z}_n = X_n^2$, antrasis dėmuo lygus 0, o paskutinis dėmuo yra

$$\tilde{\boldsymbol{U}}^T\boldsymbol{C}^T\boldsymbol{G}^{-1}\boldsymbol{C}\tilde{\boldsymbol{U}} = \frac{1}{n}\boldsymbol{v}^T\boldsymbol{G}^{-1}\boldsymbol{v}.$$

▲

**2.3.3 pastaba.** Kai įvykdytos reguliarumo sąlygos, įrodyta (žr. [24], [25]), kad kriterijaus statistikos asimptotinis skirstinys nepakis, jeigu fiksuotus grupavimo intervalų galus $a_i$ pakeisime tam tikromis imties funkcijomis (statistikomis).

Parinkime $k$ fiksuotų teigiamų skaičių $p_1, ..., p_k$, tenkinančių sąlygą $p_1 + \cdots + p_k = 1$, paprastai $p_i = 1/k$. Apibrėžkime

$$F_0^{-1}(x, \boldsymbol{\theta}) = \inf\{y : F_0(y, \boldsymbol{\theta}) \geq x\}$$

atvirkštinę pasiskirstymo funkcijai $F_0$.

Grupavimo intervalų galais imkime statistikas

$$a_i = a_i(\hat{\boldsymbol{\theta}}_n) = F_0^{-1}(P_i, \hat{\boldsymbol{\theta}}_n), \quad P_i = p_1 + \cdots + p_i,$$

$$i = 1, 2...., k, \quad z_0 = -\infty.$$

Tada formulėje (2.3.2) tikimybės $\pi_j(\boldsymbol{\theta})$ yra

$$\pi_j(\boldsymbol{\theta}) = F_0(a_j(\boldsymbol{\theta}); \hat{\boldsymbol{\theta}}_n) - F_0(a_{j-1}(\boldsymbol{\theta}); \hat{\boldsymbol{\theta}}_n) = \int_{a_{j-1}(\boldsymbol{\theta})}^{a_j(\boldsymbol{\theta})} f(x; \hat{\boldsymbol{\theta}}_n)dx$$

ir

$$c_{ij} = \frac{\partial \pi_j(\hat{\boldsymbol{\theta}}_n)}{\partial \theta_i} = f(a_j(\hat{\boldsymbol{\theta}}_n); \hat{\boldsymbol{\theta}}_n) \frac{\partial a_j(\hat{\boldsymbol{\theta}}_n)}{\partial \theta_i} - f(a_{j-1}(\hat{\boldsymbol{\theta}}_n); \hat{\boldsymbol{\theta}}_n) \frac{\partial a_{j-1}(\hat{\boldsymbol{\theta}}_n)}{\partial \theta_i}.$$

Kriterijaus statistika (2.3.4) įgauna tokį pavidalą:

$$Y_n^2 = X_n^2 + \frac{1}{n} \boldsymbol{v}^T \boldsymbol{G}^{-1} \boldsymbol{v}\,; \tag{2.3.6}$$

čia

$$X_n^2 = \sum_{i=1}^{k} \frac{(U_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{U_i^2}{np_i} - n, \tag{2.3.7}$$

$$\boldsymbol{v} = (v_1, ..., v_s)^T, \quad v_j = \frac{c_{1j} U_1}{p_1} + ... + \frac{c_{kj} U_k}{p_k},$$

$$\boldsymbol{G} = [g_{rr'}]_{s \times s}, \quad g_{rr'} = i_{rr'} - \sum_{l=1}^{k} \frac{c_{lr} c_{lr'}}{p_l}.$$

**Nikulino, Rao ir Robsono kriterijus:** hipotezė $H_0'$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$Y_n^2 > \chi_\alpha^2(k-1). \tag{2.3.8}$$

**2.3.4 pastaba.** Taikant modifikuotąjį chi kvadrato kriterijų kartais pakanka apsiriboti statistikos $X_n^2$ radimu. Iš tikrųjų, jeigu $X_n^2 > \chi_\alpha^2(k-1)$, tai hipotezę reikia atmesti, nes $Y_n^2 \geq X_n^2 > \chi_\alpha^2(k-1)$.

## 2.3.2. Eksponentiškumo tikrinimas

Tikrinsime hipotezę

$$H : F \in \{G : G(x; \theta) = 1 - e^{-x/\theta},\ x \geq 0;\ \theta > 0\},$$

kad paprastoji imtis gauta stebint eksponentinį a. d.

Pažymėkime $\hat{\theta} = \bar{X}$ parametro $\theta$ DT įvertinį. Pagal formulę (2.3.6) gauname:

$$a_i = \hat{\theta} z_i, \quad z_i = -\ln(1 - P_i), \quad \frac{\partial a_i}{\partial \theta} = z_i, \tag{2.3.9}$$

$$c_i := c_{i1} = \frac{z_i e^{-z_i} - z_{i-1} e^{-z_{i-1}}}{\hat{\theta}} = \frac{b_i}{\hat{\theta}}, \quad i_{11} = \frac{1}{\hat{\theta}^2},$$

Taigi kriterijaus statistika turi tokį pavidalą

$$Y_n^2 = X_n^2 + Q_n, \quad X_n^2 = \sum_{i=1}^{k} \frac{U_i^2}{np_i} - n, \quad Q_n = \frac{v^2}{n\lambda}; \tag{2.3.10}$$

čia

$$v = \sum_{i=1}^{k} \frac{b_i U_i}{p_i}, \quad \lambda = 1 - \sum_{i=1}^{k} \frac{b_i^2}{p_i}.$$

**2.3.1 pavyzdys.** Stebint $n = 69$ elektros lempučių degimo laiką gauti tokie rezultatai (sąlyginiais vienetais):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5,017 | 0,146 | 6,474 | 13,291 | 5,126 | 8,934 | 10,971 | 7,863 | 5,492 | 13,930 |
| 12,708 | 7,329 | 5,408 | 6,808 | 0,923 | 4,679 | 2,242 | 4,120 | 12,080 | 2,502 |
| 16,182 | 6,592 | 2,653 | 4,252 | 8,609 | 10,419 | 2,173 | 3,321 | 4,086 | 11,667 |
| 19,474 | 11,067 | 11,503 | 2,284 | 0,926 | 2,065 | 4,703 | 3,744 | 5,286 | 5,497 |
| 4,881 | 0,529 | 10,397 | 30,621 | 5,193 | 7,901 | 10,220 | 16,806 | 10,672 | 4,209 |
| 5,699 | 20,952 | 12,542 | 7,316 | 0,272 | 4,380 | 9,699 | 9,466 | 7,928 | 13,086 |
| 8,871 | 13,000 | 16,132 | 9,950 | 8,449 | 8,301 | 16,127 | 22,698 | 4,335 | |

Tikrinsime hipotezę

$$H : F \in \{G : G(x;\theta) = 1 - \exp\{-x/\theta\},\ \ x, \theta > 0\},$$

kad lempučių darbo laikas pasiskirstęs pagal eksponentinį dėsnį.

Randame $\hat{\theta} = \bar{X} = 8,231$. Parenkame $k = 6$ intervalus. Tada $p_i = 1/6$, $P_i = i/6$, $i = 1, ..., 6$. Tarpiniai skaičiavimo rezultatai pateikti lentelėje.

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $P_i$ | 0,0000 | 0,1667 | 0,3333 | 0,5000 | 0,6667 | 0,8333 | 1,0000 |
| $a_i$ | 0,0000 | 1,5005 | 3,3377 | 5,7049 | 9,0426 | 14,7483 | $\infty$ |
| $z_i$ | 0,0000 | 0,1823 | 0,4055 | 0,6931 | 1,0986 | 1,7918 | $\infty$ |
| $b_i$ | – | 0,1519 | 0,1184 | 0,0762 | 0,0196 | -0,0676 | -0,2986 |
| $U_i$ | – | 5 | 8 | 18 | 13 | 18 | 8 |

Gauname $v = -1,6259, \lambda = 0,1778$, $X_n^2 = 13,1429$, $Q_n = 0,2124$ ir $Y_n^2 = 13,3553$. Asimptotinė $P$ reikšmės $pv_a = \mathbf{P}\{\chi_5^2 > 13,3553\} = 0,0203$. Hipotezė atmestina.

Šiame pavyzdyje pataisos $Q_n$ buvo galima neskaičiuoti, nes $\mathbf{P}\{\chi_5^2 > 13,1429\} = 0,0221$, t. y. hipotezė atmestina remiantis vien statistikos $X_n^2$ stebiniu.

### 2.3.3. Skirstiniai, priklausantys nuo poslinkio ir mastelio parametrų

Tikrinsime hipotezę

$$H_0 : F \in \{G : G(x) = \{G_0((x-\mu)/\sigma)),\ x \in \mathbf{R},\ -\infty < \mu < +\infty,\ 0 < \sigma < \infty\};$$

čia $G_0$ yra žinoma pasiskirstymo funkcija. Taigi hipotezėje tvirtinama, kad imties elemento $X_i$ skirstinys priklauso specialiai skirstinių, priklausančių tik nuo poslinkio ir mastelio parametrų, šeimai. Gautas kriterijus tiks ir hipotezėms

$$H_0^* : F \in \{G : G(x) = G_0((\frac{x}{\theta})^\nu),\ x > 0,\ 0 < \theta, \nu < \infty\}$$

tikrinti, nes, atlikus logaritminę transformaciją $Y_i = \ln X_i$, pastaroji skirstinių šeima suvedama į šeimą, priklausančią tik nuo poslinkio ir mastelio parametrų.

Tarkime, kad hipotezė $H_0$ yra teisinga. Pažymėkime $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})^T$ parametro DT įvertinį. Remdamiesi formulėmis (2.3.6) gauname:

$$a_i = \hat{\mu} + z_i \hat{\sigma}, \quad z_i = G_0^{-1}(P_i), \quad \frac{\partial a_i}{\partial \mu} = 1, \quad \frac{\partial a_i}{\partial \sigma} = z_i, \quad g(x) = G_0'(x), \quad (2.3.11)$$

$$c_{i1} = \frac{g(z_i) - g(z_{i-1})}{\hat{\sigma}} = \frac{b_{i1}}{\hat{\sigma}}, \quad c_{i2} = \frac{z_i g(z_i) - z_{i-1} g(z_{i-1})}{\hat{\sigma}} = \frac{b_{i2}}{\hat{\sigma}}.$$

Tegu

$$j_{rs} = \int_{-\infty}^{\infty} x^r \left[\frac{g'(x)}{g(x)}\right]^s g(x)dx, \quad r = 0, 1, 2; \quad s = 1, 2.$$

Tada Fišerio informacinės matricos elementai $\boldsymbol{i}(\hat{\boldsymbol{\theta}}_n)$ yra tokie:

$$i_{11} = \frac{j_{02}}{\hat{\sigma}^2}, \quad i_{12} = \frac{j_{12}}{\hat{\sigma}^2}, \quad i_{22} = \frac{j_{22} + 2j_{11} + 1}{\hat{\sigma}^2}. \tag{2.3.12}$$

Taigi statistika $Y_n^2$ turi tokį pavidalą

$$Y_n^2 = X_n^2 + Q_n, \quad Q_n = \frac{\lambda_1 \beta^2 - 2\lambda_3 \alpha\beta + \lambda_2 \alpha^2}{n(\lambda_1 \lambda_2 - \lambda_3^2)}; \tag{2.3.13}$$

čia

$$\alpha = \sum_{i=1}^k \frac{b_{i1} U_i}{p_i}, \quad \beta = \sum_{i=1}^k \frac{b_{i2} U_i}{p_i},$$

$$\lambda_1 = j_{02} - \sum_{i=1}^k \frac{b_{i1}^2}{p_i}, \ \lambda_2 = j_{22} + 2j_{11} + 1 - \sum_{i=1}^k \frac{b_{i2}^2}{p_i}, \ \lambda_3 = j_{12} - \sum_{i=1}^k \frac{b_{i1} b_{i2}}{p_i}.$$

Pateiksime išraiškas, reikalingas kriterijaus statistikai apskaičiuoti keleto dažnai naudojamų skirstinių atveju.

**Normalusis skirstinys:** $F(x; \mu, \sigma) = \Phi((x - \mu)/\sigma)$, $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-u^2/2} du.$

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\sigma} = s, \quad s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

$$G(x) = \Phi(x), \quad g(x) = \varphi(x) = \Phi'(x),$$

$$j_{02} = 1, \quad j_{12} = 0, \quad j_{22} + 2j_{11} + 1 = 2. \tag{2.3.14}$$

**Lognormalusis skirstinys:** $F(x; \theta, \nu) = \Phi(\ln(x/\theta)^\nu) = \Phi((\ln x - \mu)/\sigma)$, $x > 0$. Atlikę transformaciją $Y_i = \ln X_i$, gauname normaliųjų skirstinių šeimą.

**Logistinis skirstinys:** $F(x; \mu, \sigma) = (1 + e^{-(x-\mu)/\sigma})^{-1}$.

$$G(x) = (1 + e^{-x})^{-1}, \quad g(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{e^x}{(1 + e^x)^2},$$

$$j_{02} = \frac{1}{3}, \quad j_{12} = 0, \quad j_{22} + 2j_{11} + 1 = \frac{\pi^2 + 3}{9}.$$

DT įvertiniai $\hat{\mu}$ ir $\hat{\sigma}$ maksimizuoja logtikėtinumo funkciją

$$\ell(\mu,\sigma) = -n\ln\sigma - \sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} - 2\sum_{i=1}^{n}\ln(1+e^{\frac{X_i-\mu}{\sigma}}).$$

**Loglogistinis skirstinys:** $F(x;\theta,\nu) = 1-(1+(\frac{x}{\theta})^{\nu})^{-1}$, $x>0$.
Atlikę transformaciją $Y_i = \ln X_i$, gauname logistinių skirstinių šeimą.

**Ekstremalių reikšmių skirstinys:** $F(x;\mu,\sigma) = 1-e^{-e^{\frac{x-\mu}{\sigma}}}$.

$$G(x) = 1-e^{-e^{x}}, \quad g(x) = e^{x}e^{-e^{x}},$$

$$j_{02} = 1, \quad j_{12} = \Gamma'(1)+1, \quad j_{22}+2j_{11}+1 = \Gamma''(1)+2\Gamma'(1)+1.$$

DT įvertiniai $\hat{\mu}$ ir $\hat{\sigma}$ maksimizuoja logtikėtinumo funkciją

$$\ell(\mu,\sigma) = -n\ln\sigma - \sum_{i=1}^{n}e^{\frac{X_i-\mu}{\sigma}} + \sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}.$$

**Veibulo skirstinys:** $F(x;\theta,\nu) = 1-e^{-(\frac{x}{\theta})^{\nu}}$, $x>0$.
Atlikę transformaciją $Y_i = \ln X_i$, gauname ekstremaliųjų reikšmių skirstinių šeimą.

**Koši skirstinys:** $F(x;\mu,\sigma) = \frac{1}{\pi}\,(arctg\,\frac{x-\mu}{\sigma}+\frac{\pi}{2})$.

$$G(x) = \frac{1}{\pi}\,(arctg\,x+\frac{\pi}{2}), \;\; g(x) = \frac{1}{\pi(1+x^2)}, \;\; j_{02} = \frac{1}{2}, \;\; j_{12} = 0, \;\; j_{22}+2j_{11}+1 = \frac{1}{2}.$$

DT įvertiniai $\hat{\mu}$ ir $\hat{\sigma}$ maksimizuoja logtikėtinumo funkciją

$$\ell(\mu,\sigma) = -n\ln\sigma - \sum_{i=1}^{n}\ln(1+(\frac{X_i-\mu}{\sigma})^2).$$

Pateiksime modifikuoto chi kvadrato kriterijaus naudojimo rekomendacijas, kai tikimybinių skirstinių šeima priklauso tik nuo poslinkio ir mastelio parametrų, o alternatyvos nėra tiksliai suformuluotos.

1) Parenkame grupavimo intervalų skaičių $k$ ir tikimybes $p_i = 1/k$, $i = 1,...,k$. Tada $P_i = i/k$. Intervalų skaičių $k$ rekomenduojame parinkti taip, kad būtų tenkinama nelygybė $n/k > 5$.

2) Randame grupavimo intervalų galus: $a_i = \hat{\mu}+z_i\hat{\sigma}$; čia $z_i = G^{-1}(P_i)$, $\hat{\mu},\hat{\sigma}$ – parametrų DT įvertiniai, surasti pagal pradinius (negrupuotus) duomenis.

3) Randame stebėjimų, patekusių į intervalus $(a_{i-1},\ a_i]$, skaičius $U_i$.

4) Apskaičiuojame statistikos

$$X_n^2 = \sum_{i=1}^{k}\frac{U_i^2}{np_i} - n = \frac{k}{n}\sum_{i=1}^{k}U_i^2 - n.$$

reikšmę. Jeigu $X_n^2 > \chi^2_\alpha(k-1)$ (arba, ekvivalentiškai, $pv_a = \mathbf{P}\{\chi^2_{k-1} > x_n^2\} < \alpha$; čia $x_n^2$ yra statistikos $X_n^2$ stebinys), tai hipotezę atmetame.

5) Jeigu $X_n^2 < \chi^2_\alpha(k-1)$, tai apskaičiuojame statistikos $Q_n$ reikšmę pagal formulę (2.3.13). Pabrėšime, kad $\lambda_3 = 0$, kai $p_i = 1/k$, ir $g(-x) = g(x)$.

Hipotezė atmetama, kai

$$pv_a = \mathbf{P}\{\chi^2_{k-1} > y_n^2\} < \alpha;$$

čia $y_n^2$ yra statistikos $Y_n^2$ stebinys
(arba, ekvivalentiškai, kai $\tilde{Y}_n^2 = X^2 + Q_n > \chi^2_\alpha(k-1)$).

**2.3.2 pavyzdys.** Matuojama per vienodus laiko intervalus iš gręžinio gautos naftos kiekis $V$. Lentelėje pateikiami gauti rezultatai $V_1, \ldots, V_{49}$ (sąlyginiais vienetais).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 8,7 | 6,6 | 10,0 | 24,3 | 7,9 | 1,3 | 26,2 | 8,3 | 0,9 | 7,1 |
| 5,9 | 16,8 | 6,0 | 13,4 | 31,7 | 8,3 | 28,3 | 17,1 | 16,7 | 19,7 |
| 5,2 | 18,9 | 1,0 | 3,5 | 2,7 | 12,0 | 8,3 | 14,8 | 6,3 | 39,3 |
| 4,3 | 19,4 | 6,5 | 7,4 | 3,4 | 7,6 | 8,3 | 1,9 | 10,3 | 3,2 |
| 0,7 | 19,0 | 26,2 | 10,0 | 17,7 | 14,1 | 44,8 | 3,4 | 3,5 | |

Reikia patikrinti hipotezę, kad stebimo a. d. $V$ skirstinys yra a) normalusis; b) lognormalusis; c) a. d. $V^{1/4}$ skirstinys yra normalusis.

a)
1) Parenkame $k = 6$, $p_i = 1/6$.
2 – 3) Gauname $\bar{X} = 12,018$ ir $s = 9,930$. Grupavimo intervalų rėžiai $a_i$ ir $U_i$ pateikti lentelėje.

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $a_i$ | $-\infty$ | 2,4117 | 7,7411 | 12,0180 | 16,2949 | 21,6243 | $+\infty$ |
| $U_i$ | | 5 | 16 | 10 | 3 | 8 | 7 |

4) Randame $X_n^2 = \frac{6}{49}\sum_{i=1}^6 U_i^2 - 49 = 12,5918$. Kadangi $k = 6$ ir

$$\mathbf{P}\{\chi^2_5 > 12,5918\} = 0,0275,$$

tai normališkumo hipotezė atmestina (neskaičiuojant statistikos $Y_n^2$). Vis dėlto, jei pratęsime analizę, tai gausime $Q_n = 5,0515$, $Y_n^2 = 17,6433$ ir $pv_a = \mathbf{P}\{\chi^2_5 > 17,6433\} = 0,0034$. Taigi kriterijus, grindžiamas statistika $Y_n^2$, atmeta hipotezę dar labiau.
b) Atliekame transformaciją $\ln V_1, \ldots, \ln V_{49}$.
1) Parenkame $k = 6$, $p_i = 1/6$.
2 – 3) Parametrų $\mu$ ir $\sigma$ DT įvertiniai: $\bar{X} = 2,1029$ ir $s = 0,9675$. Rėžiai $a_i$ ir dažniai $U_i$ pateikti lentelėje.

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $a_i$ | $-\infty$ | 1,1675 | 1,6865 | 2,1030 | 2,5195 | 3,0385 | $+\infty$ |
| $U_i$ | | 7 | 6 | 9 | 9 | 11 | 7 |

4) Gauname $X_n^2 = \frac{6}{49}\sum_{i=1}^6 U_i^2 - 49 = 2,0612$ ir $\mathbf{P}\{\chi^2_5 > 2,0612\} = 0,8406$. Hipotezė neatmetama.

5) Pratęsiame analizę apskaičiuodami statistikos $Y_n^2$ reikšmę. Pagal (2.3.13) normaliojo skirstinio atveju gauname $Q_n = 3,5695$ ir $Y_n^2 = X_n^2 + Q_n = 5,6307$. Asimptotinė $P$ reikšmė $pv_a = \mathbf{P}\{\chi^2_5 > 5,6307\} = 0,3438$. Hipotezė neatmetama, jei kriterijaus reikšmingumo lygmuo neviršija 0,3438.
c) Atliekame transformaciją $V_1^{1/4}, \ldots, V_{49}^{1/4}$.
1) Parenkame $k = 6$, $p_i = 1/6$.

2 – 3) Parametrų $\mu$ ir $\sigma$ DT įvertiniai: $\bar{X} = 1,7394$ ir $s = 0,3944$. Rėžiai $a_i$ ir dažniai $U_i$ pateikti lentelėje.

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $a_i$ | $-\infty$ | 1,3579 | 1,5695 | 1,7394 | 1,9093 | 2,1209 | $+\infty$ |
| $U_i$ | | 7 | 8 | 12 | 4 | 11 | 7 |

4) Statistika $X_n^2 = \frac{6}{49}\sum_{i=1}^{6} U_i^2 - 49 = 5,2449$ ir $\mathbf{P}\{\chi_5^2 > 5,2449\} = 0,3867$. Hipotezė neatmetama.

5) Pratęsiame analizę apskaičiuodami statistikos $Y_n^2$ reikšmę. Gauname: $Q_n = 0,4681$, $Y_n^2 = X_n^2 + Q_n = 5,7130$. Asimptotinė $P$ reikšmė grindžiama statistika $Y_n^2$ yra

$$pv = \mathbf{P}\{\chi_5^2 > 5,7130\} = 0,3352$$

gana didelė. Todėl nėra pagrindo atmesti hipotezę.

**2.3.3 pavyzdys.** Reikia patikrinti hipotezę, kad pateiktieji $n = 100$ skaičių gauti stebint normalųjį a. d.

237,34 247,43 251,30 257,64 258,87 261,01 263,05 265,37 265,77 265,95 271,59 273,84 278,85
282,56 283,10 283,18 283,22 285,99 287,81 288,24 291,15 291,86 294,32 295,36 295,47 295,90
296,92 297,63 298,75 300,52 302,95 303,58 304,46 304,55 305,24 305,24 306,25 306,64 306,80
307,31 307,96 308,49 309,70 310,25 310,32 312,18 313,18 313,37 313,61 313,63 315,03 316,35
317,91 318,34 319,42 322,38 324,55 325,02 325,47 326,14 327,82 327,83 337,45 340,73 341,14
342,14 343,50 344,21 346,49 346,72 346,81 348,46 350,19 350,20 351,25 352,23 353,04 353,44
353,79 355,09 355,63 357,48 365,92 366,76 370,46 371,77 373,10 373,92 380,89 381,26 387,94
391,21 402,68 406,04 406,60 409,58 414,93 415,18 418,82 444,66

1) Tegu $k = 8$ ir $p_1 = ... = p_8 = 1/8 = 0,125$.

2) Parametrų $\mu$ ir $\sigma$ DT įvertiniai yra $\hat{\mu} = 324,3367$, $\hat{\sigma} = 42,9614$,

$$P_i = i/8, \quad z_i = \Phi^{-1}(P_i), \quad a_i = \hat{\mu} + z_i\hat{\sigma}.$$

Reikšmės $z_i, a_i$ ir $U_i$ pateikiamos lentelėje.

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $z_i$ | $-\infty$ | -1,150 | -0,674 | -0,319 | 0,000 | 0,319 | 0,674 | 1,150 | $+\infty$ |
| $a_i$ | $-\infty$ | 274,919 | 295,360 | 310,650 | 324,337 | 338,024 | 353,314 | 373,755 | $+\infty$ |
| $U_i$ | | 12 | 11 | 22 | 11 | 7 | 14 | 10 | 13 |

4) Gauname

$$X_n^2 = \frac{8}{100}\sum_{i=1}^{6} U_i^2 - 100 = 10,72$$

ir

$$\mathbf{P}\{\chi_7^2 > 10,72\} = 0,1513$$

nėra maža, todėl skaičiuojame statistikos $Y_n^2$ reikšmę.

5) Pagal (2.3.13) formulę, kai skirstinys normalusis, gauname: $Q_n = 2,6503$, $Y_n^2 = X_n^2 + Q_n = 13,3703$.

Asimptotinė $P$ reikšmė grindžiama statistika $Y_n^2$ yra $pv_a = \mathbf{P}\{\chi_7^2 > 13,3703\} = 0,0636$. Hipotezė neatmestina, jei kriterijaus reikšmingumo lygmuo neviršija 0,0636.

Tarkime, kad hipotezei tikrinti taikome skyrelio 2.3.2. Pirsono kriterijų imdami surastus DT įvertinius ir intervalus. Formaliai žiūrint tai nėra korektiška, nes įvertiniai surasti ne pagal grupuotus duomenis, o intervalų galai nėra konstantos. Jeigu nekreipdami dėmesio į šias aplinkybes gautąją statistiką remdamiesi teorema 2.2.1 aproksimuotume chi kvadrato skirstiniu su $k - s - 1 = 5$ laisvės laipsniais (įvertinti 2 parametrai), tai gautume

$$pv_a = \mathbf{P}\{\chi_5^2 > 10,72\} = 0,0572.$$

Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,0572. Matome, kad nekorektiškai taikydami chi kvadrato suderinamumo kriterijų galime padaryti klaidingą išvadą.

## 2.4. Chi kvadrato nepriklausomumo kriterijus

Tarkime

$$\mathcal{A} = \{A_1, \ldots, A_s : A_i \cap A_j = \emptyset, i \neq j = 1, \ldots s, \cup_{i=1}^{s} A_i = \Omega\},$$

$$\mathcal{B} = \{B_1, \ldots, B_r : B_i \cap B_j = \emptyset, i \neq j = 1, \ldots r, \cup_{i=1}^{r} B_i = \Omega\}$$

yra dvi nesutaikomų, sudarančių pilną grupę įvykių, kuriuos galime stebėti eksperimento metu, sistemos. Pavyzdžiui, jei dvimačio a. v. $(X, Y)^T$ komponenčių reikšmių aibės padalytos atitinkamai į poaibius $I_i$, $i = 1, \ldots, s$ ir $J_j$, $j = 1, \ldots, r$, tai galime apibrėžti $A_i = \{X \in I_i\}$, $B_j = \{Y \in J_j\}$.

Konkretesnis pavyzdys: įvykiai $A_1, \ldots, A_5$ gali reikšti atitinkamai, kad sutuoktinių pora turi 0, 1, 2, 3 arba ne mažiau kaip 4 vaikus ($X$ reiškia vaikų skaičių), o įvykiai $B_1, \ldots, B_4$ gali reikšti, kad sutuoktinių metinės pajamos (eurais) atitinkamai patenka į intervalus

$$[0,\, 400], \quad (400,\, 900], \quad (900,\, 1500] \quad \text{ir} \quad (1500, \infty),$$

($Y$ – pajamas).

Atliekama $n$ stebėjimų. Pažymėkime $U_{ij}$ įvykio $A_i \cap B_j$ pasirodymo skaičių. Pavyzdžiui, $U_{23}$ reiškia skaičių šeimų, kurios turi du vaikus, o pajamos yra intervale (900, 1500] tarp $n$ stebėtųjų.

Stebėjimo rezultatus galime surašyti į pavidalo **2.4.1** lentelę.

**2.4.1 lentelė.** Stebėjimo rezultatai

| $A_i \backslash B_j$ | $B_1$ | $B_2$ | ... | $B_r$ | $\Sigma$ |
|---|---|---|---|---|---|
| $A_1$ | $U_{11}$ | $U_{12}$ | ... | $U_{1r}$ | $U_{1.}$ |
| $A_2$ | $U_{21}$ | $U_{22}$ | ... | $U_{2r}$ | $U_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_s$ | $U_{s1}$ | $U_{s2}$ | ... | $U_{sr}$ | $U_{s.}$ |
| $\Sigma$ | $U_{.1}$ | $U_{.2}$ | ... | $U_{.r}$ | $n$ |

Čia

$$U_{i.} = \sum_{j=1}^{r} U_{ij}, \quad i = 1, ..., s; \quad U_{.j} = \sum_{i=1}^{s} U_{ij}, \quad j = 1, ..., r.$$

Pažymėkime

$$\pi_{ij} = \mathbf{P}\{A_i \cap B_j\},$$

$$\pi_{i.} = \sum_{j=1}^{r} \pi_{ij}, \quad \pi_{.j} = \sum_{i=1}^{s} \pi_{ij}, \quad \sum_{i=1}^{s} \pi_{i.} = 1, \quad \sum_{j=1}^{r} \pi_{.j} = 1.$$

Atsitiktinis vektorius

$$\boldsymbol{U} = (U_{11}, ..., U_{1r},\ U_{21}, ..., U_{2r}, ..., U_{s1}, ..., U_{sr})^T$$

turi polinominį skirstinį (žr. (2.1.2)):

$$\mathcal{P} = \{\mathcal{P}_{s\times r}(n, \boldsymbol{\pi}), \quad \boldsymbol{\pi} = (\pi_{11}, ..., \pi_{sr})^T, \quad 0 < \pi_{ij} < 1, \quad \sum_{i=1}^{s}\sum_{j=1}^{r}\pi_{ij} = 1\}.$$

**Nepriklausomumo hipotezė** (dviejų atsitiktinių įvykių sistemų)

$$H_0' : \pi_{ij} = \mathbf{P}\{A_i \cap B_j\} = \mathbf{P}\{A_i\}\mathbf{P}\{B_j\} = \pi_{i\cdot}\pi_{\cdot j}, \quad i = 1, \ldots, s;\ j = 1, \ldots, r. \tag{2.4.1}$$

Jeigu sistemos $\mathcal{A}$ ir $\mathcal{B}$ apibrėžiamos naudojant a. d. $X$ ir $Y$ patekimą į intervalų sistemas, tai esant neteisingai hipotezei $H_0'$ tuo labiau neteisinga

**Nepriklausomumo hipotezė** (dviejų atsitiktinių dydžių)

$$H_0 : F(x, y) = \mathbf{P}\{X \le x, Y \le y\} = F_1(x)F_2(y), \quad \forall x, y \in \mathbf{R}. \tag{2.4.2}$$

Tikrinsime hipotezę $H_0'$. Atmetus hipotezę $H_0'$, natūralu atmesti ir hipotezę $H_0$. Hipotezės $H_0'$ alternatyva yra

$$H_1' : \pi_{ij} \ne \pi_{i\cdot}\pi_{\cdot j} \quad \text{su kuriais nors } i, j.$$

Polinominio skirstinio tikimybių $\pi_{ij}$ DT įvertiniai, maksimizuojantys tikėtinumo funkciją

$$L(\boldsymbol{\pi}) = \frac{n!}{U_{11}! \cdots U_{sr}!}\prod_{i=1}^{s}\prod_{j=1}^{r}\pi_{ij}^{U_{ij}},$$

yra $\hat{\pi}_{ij} = U_{ij}/n$.

Jeigu hipotezė $H_0'$ teisinga, tai $\pi_{ij}$ yra $s + r - 2$ parametrų

$$\boldsymbol{\theta} = (\pi_{1\cdot}, ..., \pi_{s-1\cdot}, \pi_{\cdot 1}, ..., \pi_{\cdot r-1})^T,$$

funkcijos, t. y.

$$\pi_{ij} = \pi_{ij}(\boldsymbol{\theta}) = \pi_{i\cdot}\pi_{\cdot j}.$$

Taigi, kai teisinga hipotezė $H_0'$, tikėtinumo funkcija turi tokį pavidalą (plg. (2.2.4))

$$\tilde{L}(\boldsymbol{\theta}) = L(\boldsymbol{\pi}|\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}) = \frac{n!}{U_{11}! \cdots U_{sr}!}\prod_{i=1}^{s}\pi_{i\cdot}^{U_{i\cdot}}\prod_{j=1}^{r}\pi_{\cdot j}^{U_{\cdot j}},$$

o jos logaritmas

$$\tilde{\ell}(\boldsymbol{\theta}) = C + \sum_{i=1}^{s}U_{i\cdot}\ln\pi_{i\cdot} + \sum_{j=1}^{r}U_{\cdot j}\ln\pi_{\cdot j}.$$

Parametrų $\pi_{i\cdot}$ DT įvertiniai tenkina lygtis

$$\frac{\partial\tilde{\ell}}{\partial\pi_{i\cdot}} = \frac{U_{i\cdot}}{\pi_{i\cdot}} - \frac{U_{s\cdot}}{\pi_{s\cdot}} = 0, \quad i = 1, 2, \ldots, s-1,$$

o kartu ir lygtis

$$U_{i\cdot}\pi_{s\cdot} = U_{s\cdot}\pi_{i\cdot}, \quad i = 1, 2, \ldots, s.$$

Sumuodami pagal $i = 1, \ldots s$ ir atsižvelgę į sąlygas

$$\sum_{i=1}^{s} \pi_{i\cdot} = 1, \quad \sum_{i=1}^{s} U_{i\cdot} = n,$$

gauname tikimybių $\pi_{i\cdot}$ DT įvertinius:

$$\hat{\pi}_{i\cdot} = U_{i\cdot}/n, \quad i = 1, ..., s.$$

Analogiškai

$$\hat{\pi}_{\cdot j} = U_{\cdot j}/n, \quad j = 1, ..., r.$$

Taigi, kai teisinga hipotezė $H_0'$, polinominio skirstinio tikimybių $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ DT įvertiniai yra

$$\hat{\pi}_{ij} = \pi_{ij}(\boldsymbol{\theta}^*) = \hat{\pi}_{i\cdot} \cdot \hat{\pi}_{\cdot j} = \frac{U_{i\cdot}}{n}\frac{U_{\cdot j}}{n}.$$

Naudodami šiuos įvertinius gauname statistiką (žr. (2.2.6))

$$X_n^2 = X_n^2(\boldsymbol{\theta}^*) = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{(U_{ij} - n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j})^2}{n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{r} \frac{U_{ij}^2}{U_{i\cdot}U_{\cdot j}} - 1\right). \tag{2.4.3}$$

Pagal teoremą 2.2.1 gautoji statistika asimptotiškai (kai $n \to \infty$) pasiskirsčiusi pagal chi kvadrato skirstinį su

$$rs - 1 - (r + s - 2) = (r - 1)(s - 1)$$

laisvės laipsnių.

**Chi kvadrato nepriklausomumo kriterijus:** hipotezė $H_0'$ atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai

$$X_n^2 > \chi_\alpha^2((r-1)(s-1)). \tag{2.4.4}$$

Jeigu hipotezė $H_0'$ atmetama, tai natūralu atmesti ir hipotezę $H_0$.

Jeigu $s = r = 2$, tai statistikos $X_n^2$ išraiška paprastesnė:

$$X_n^2 = \frac{n(U_{11}U_{22} - U_{12}U_{21})^2}{U_{1\cdot}U_{2\cdot}U_{\cdot 1}U_{\cdot 2}}.$$

**2.4.1 pavyzdys.** Lentelėje pateikti skaičiai sutuoktinių, sugrupuotų pagal vaikų skaičių (požymis $A$) ir metines pajamas (požymis $B$). Reikia patikrinti hipotezę dėl požymių $A$ ir $B$ nepriklausomumo.

**2.4.2 lentelė.** Statistiniai duomenys

| | [0, 400] | (400, 900] | (900, 1500] | $(1500, \infty)$ | |
|---|---|---|---|---|---|
| 0 | 2161 | 3577 | 2184 | 1635 | 9558 |
| 1 | 2755 | 5081 | 2222 | 1052 | 11110 |
| 2 | 936 | 1753 | 640 | 306 | 3635 |
| 3 | 225 | 419 | 96 | 38 | 778 |
| $\geq 4$ | 39 | 98 | 31 | 14 | 182 |
| | 6116 | 10928 | 5173 | 3046 | 25263 |

Iš šios lentelės gauname

$$X_n^2 = 25263\left(\frac{2161^2}{9558 \cdot 6116} + \frac{3577^2}{9558 \cdot 10928} + ... + \frac{14^2}{182 \cdot 3046} - 1\right) = 568,5.$$

Kai hipotezė teisinga, šios statistikos skirstinys aproksimuojamas chi kvadrato skirstiniu su $(4-1)(5-1) = 12$ laisvės laipsnių. Kadangi

$$pv_a = 1 - \mathbf{P}\{\chi^2_{12} > 568,5\} < 10^{-16}$$

tai hipotezė atmetama.

## 2.5. Chi kvadrato homogeniškumo kriterijus

Sakykime, kad yra $s$ nepriklausomų objektų grupių; $i$-osios grupės objektų skaičių pažymėkime $n_i$. Tarkime, kad

$$\mathcal{B} = \{B_1, \ldots, B_r : B_i \cap B_j = \emptyset, i \neq j = 1, \ldots r, \cup_{i=1}^r B_i = \Omega\}$$

yra pilna nesutaikomų įvykių grupė. Atlikus bet kurio objekto stebėjimą, žinoma, kuris iš įvykių $B_1, \ldots, B_r$ įvyko.

Pažymėkime

$$\pi_{ij} = \mathbf{P}\{B_j \,|\, j\text{-asis objektas priklauso } i\text{-ajai grupei}\}. \tag{2.5.1}$$

Pavyzdžiui, $s$ skirtingų profesijų atstovų pildo psichologinį testą. Kiekvienas asmuo gali būti įvertintas balais $1, \ldots, 5$. Šiuo atveju

$$B_j = \{\text{asmuo surinko } j \quad \text{balų}\},$$

$j = 1, \ldots, r = 5$, ir $\pi_{ij}$ yra tikimybė, kad $i$-osios profesijos atstovas surinks $j$ balų. Šiame pavyzdyje dominantis faktorius $B$ yra surinkti balai.

**Homogeniškumo hipotezė** (faktoriaus $B$ atžvilgiu):

$$H_0 : \pi_{1j} = ... = \pi_{sj} := \pi_j, \; j = 1, ..., r, \tag{2.5.2}$$

kuri reiškia, kad esant fiksuotam $j$ įvykio $B_j$ tikimybė yra ta pati visų grupių objektams.

Minėto pavyzdžio atveju hipotezė reiškia, kad tikimybė gauti $j$ balų visų profesijų atstovams vienoda, kad ir kokį paimtume $j$.

**2.5.1 pastaba.** Kriterijai naudojami hipotezei $H_0$ tikrinti, dažnai taikomi ir atsitiktinių dydžių skirstinių lygybės hipotezei tikrinti.

Tarkime, kad

$$(X_{i1}, \; X_{i2}, ..., X_{in_i})^T, \; i = 1, ..., s,$$

yra $s$ nepriklausomų paprastųjų imčių. Sudalinkime abscisių ašį į intervalus taškais $-\infty = a_0 < a_1 < ... < a_r = +\infty$. Įvykis $B_j$ reiškia, kad stebimas a. d. $X$ patenka į $j$-ąjį intervalą $I_j = (a_{j-1}, a_j]$.

Tegu $F_i(x)$ yra a. d. $X_{ij}$, $i = 1, ..., s$ pasiskirstymo funkcija.

**Homogeniškumo hipotezė** (nepriklausomų imčių):

$$H_0' : F_1(x) \equiv F_2(x) \equiv \cdots \equiv F_s(x). \tag{2.5.3}$$

Jeigu hipotezė $H_0'$ yra teisinga, tai tuo labiau teisinga hipotezė $H_0$, nes

$$\pi_{ij} = \mathbf{P}\{X_i \in (a_{j-1},\ a_j]\} = F_i(a_j) - F_i(a_{j-1}), \quad i = 1, ..., s,\ j = 1, ..., r.$$

Taigi, atmetus hipotezę $H_0$, natūralu atmesti ir hipotezę $H_0'$. Jei hipotezė $H_0$ priimama, tegalima tvirtinti, kad sugrupuoti duomenys neprieštarauja hipotezei $H_0'$.

Pažymėkime $U_{ij}$ $i$-osios grupės objektų, kuriems įvyko įvykis $B_j$, $U_{i1} + ... + U_{ir} = n_i$ skaičių, t. y. $U_{ij}$ yra i-osios imties elementų, patekusių į intervalą $I_j = (a_{j-1}, a_j]$, skaičius.

Stebėjimo rezultatus galime surašyti į analogišką **2.4.1** lentelę.

**2.5.1 lentelė.** Stebėjimo duomenys

| | 1 | 2 | ... | $r$ | $\Sigma$ |
|---|---|---|---|---|---|
| 1 | $U_{11}$ | $U_{12}$ | ... | $U_{1r}$ | $n_1$ |
| 2 | $U_{21}$ | $U_{22}$ | ... | $U_{2r}$ | $n_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $s$ | $U_{s1}$ | $U_{s2}$ | ... | $U_{sr}$ | $n_s$ |
| $\Sigma$ | $U_{\cdot 1}$ | $U_{\cdot 2}$ | ... | $U_{\cdot r}$ | $n$ |

Atsitiktinis vektorius $\boldsymbol{U}_i = (U_{i1}, ..., U_{ir})^T$ turi polinominį skirstinį:

$$\boldsymbol{U}_i \sim \mathcal{P}_r(n_i,\ \boldsymbol{\pi}_i), \quad \boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{ir})^T.$$

Kai hipotezė $H_0$ teisinga, tikimybės $\pi_{ij}$ yra $r-1$ parametro $\boldsymbol{\theta} = (\pi_1, ..., \pi_{r-1})^T$ funkcijos. Tikėtinumo funkcija turi tokį pavidalą

$$\tilde{L}(\pi_1, \ldots, \pi_{r-1}) = \prod_{i=1}^{s} \frac{n_i!}{U_{i1}! \cdots U_{ir}!} \pi_1^{U_{i1}} \ldots \pi_r^{U_{ir}} = \prod_{i=1}^{s} \prod_{j=1}^{r} \frac{n_i!}{U_{ij}!} \pi_j^{U_{ij}}, \quad \pi_r = 1 - \sum_{i=1}^{r-1} \pi_i,$$

o jos logaritmas

$$\tilde{\ell}(\pi_1, \ldots, \pi_{r-1}) = C + \sum_{i=1}^{s} \sum_{j=1}^{r} U_{ij} \ln \pi_j = C + \sum_{j=1}^{r} U_{\cdot j} \ln \pi_j,$$

Parametrų $\pi_j$ DT įvertiniai tenkina lygtis

$$\frac{\partial \ell}{\partial \pi_j} = \frac{U_{\cdot j}}{\pi_j} - \frac{U_{\cdot r}}{\pi_r} = 0, \quad j = 1, 2, \ldots, r-1,$$

o kartu ir lygtis

$$U_{\cdot j} \pi_r = U_{\cdot r} \pi_j, \quad j = 1, 2, \ldots, r.$$

Sumuodami pagal $j = 1, \ldots, r$ ir atsižvelgę į lygybes

$$\sum_{j=1}^{r} \pi_j = 1, \quad \sum_{i=1}^{s} U_{\cdot j} = n,$$

gauname, kad parametrų $\pi_j$ DT įvertiniai yra

$$\hat{\pi}_j = U_{\cdot j}/n, \ j = 1, ..., r.$$

Taigi, kai teisinga hipotezė $H_0$, tikimybių $\pi_{ij}$ DT įvertiniai yra

$$\hat{\pi}_{ij} = \hat{\pi}_j = \frac{U_{\cdot j}}{n}, \quad j = 1, \ldots, r.$$

Naudodami šiuos įvertinius sudarome statistiką (plg. (2.2.6))

$$X_n^2 = \sum_{i=1}^{s} \sum_{j-1}^{r} \frac{(U_{ij} - n_i \hat{\pi}_j)^2}{n_i \hat{\pi}_j} = n \left( \sum_{i=1}^{s} \sum_{j-1}^{r} \frac{U_{ij}^2}{n_i U_{\cdot j}} - 1 \right). \tag{2.5.4}$$

Pagal šiek tiek bendresnę už 2.2.1 teoremą (kai vietoje chi kvadrato kvadratų sumos, atitinkančios vieną polinominį skirstinį, nagrinėjama suma chi kvadrato kvadratų sumų,atitinkančių kelis nepriklausomus polinominius skirstinius) gauname, kad gautoji statistika asimptotiškai ($n_i \to \infty, \ i = 1, ..., s$) turi chi kvadrato skirstinį su

$$s(r-1) - (r-1) = (r-1)(s-1)$$

laisvės laipsnių.

**Chi kvadrato homogeniškumo kriterijus:** homogeniškumo hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$X_n^2 > \chi_\alpha^2((r-1)(s-1)). \tag{2.5.5}$$

Jeigu hipotezė $H_0$ atmetama, tai natūralu atmesti ir hipotezę $H_0'$.

Kartais nagrinėjamos siauresnės už $H_0'$ hipotezės:

1. Hipotezė

$$H_1' : F_1 = ... = F_s = F_0,$$

kurioje tvirtinama, kad pasiskirstymo funkcijos $F_1, ..., F_s$ ne tik sutampa tarpusavyje, bet ir lygios žinomai pasiskirstymo funkcijai $F_0$.

Tikimybių $\pi_{ij}$ terminais formuluojama platesnė hipotezė:

$$H_1 : \pi_{1j} = ... = \pi_{sj} = \pi_{j0} = F_0(a_j) - F_0(a_{j-1}), \quad j = 1, ..., r.$$

Kadangi jokių nežinomų parametrų nėra, tai kriterijaus statistiką sudarome pagal skyrelį 2.1.:

$$X_n^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(U_{ij} - n_i \pi_{j0})^2}{n_i \pi_{j0}}. \tag{2.5.6}$$

Jeigu $H_1$ yra teisinga ir $n_i \to \infty$, $i = 1, ..., s$, tai vidinės sumos asimptotiškai pasiskirsčiusios pagal chi kvadrato skirstinius su $r-1$ laisvės laipsniu, o $X_n^2$ — su $s(r-1)$ laisvės laipsnių. Hipotezė $H_1$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$X_n^2 > \chi_\alpha^2(s(r-1)). \tag{2.5.7}$$

Jeigu hipotezė $H_1'$ atmetama, tai natūralu atmesti ir hipotezę $H_1$.

2. Hipotezė

$$H_2' : F_1(x) \equiv ... \equiv F_s(x) \equiv F(x, \boldsymbol{\theta}),$$

kurioje tvirtinama, kad pasiskirstymo funkcijos $F_1, ..., F_s$ ne tik sutampa tarpusavyje, bet ir sutampa su pasiskirstymo funkcija $F(x; \boldsymbol{\theta})$, kurios funkcinis pavidalas yra žinomas, tačiau ji priklauso nuo nežinomo baigtinės dimensijos parametro $\boldsymbol{\theta} = (\theta_1, ..., \theta_l)^T$, $0 < l < r$.

Tikimybių $\pi_{ij}$ terminais formuluojama platesnė hipotezė:

$$H_2 : \pi_{1j} = ... = \pi_{sj} = \pi_j(\boldsymbol{\theta}) = F(a_j;\ \boldsymbol{\theta}) - F(a_{j-1};\ \boldsymbol{\theta}), \quad j = 1, ..., r.$$

DT ar chi kvadrato minimumo metodu pagal grupuotus duomenis įvertinę parametrą $\boldsymbol{\theta}$, sukonstruojame statistiką

$$X_n^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(U_{ij} - n_i \pi_j(\hat{\boldsymbol{\theta}}))^2}{n_i \pi_j(\hat{\boldsymbol{\theta}})}, \tag{2.5.8}$$

kurios asimptotinis skirstinys, kai ($n_i \to \infty$), remiantis 2.2.1 teorema yra chi kvadrato su $s(r-1) - l$ laisvės laipsnių. Hipotezė $H_2$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$X_n^2 > \chi_\alpha^2(s(r-1) - l). \tag{2.5.9}$$

Jeigu hipotezė $H_2'$ atmetama, tai natūralu atmesti ir hipotezę $H_2$.

**2.5.1 pavyzdys.** Lentelėje pateikti vaikų (berniukų ir mergaičių), gimusių Švedijoje kiekvieną 1935 metų mėnesį, skaičiai (žr. [9]). Reikia patikrinti hipotezę, kad berniuko gimimo tikimybė per metus yra pastovi.

| Mėnuo | $X_{i1}$ | $X_{i2}$ | $X_{i.}$ | Mėnuo | $X_{i1}$ | $X_{i2}$ | $X_{i.}$ |
|---|---|---|---|---|---|---|---|
| 1 | 3743 | 3537 | 7280 | 7 | 3964 | 3621 | 7585 |
| 2 | 3550 | 3407 | 6957 | 8 | 3797 | 3596 | 7393 |
| 3 | 4017 | 3866 | 7883 | 9 | 3712 | 3491 | 7203 |
| 4 | 4173 | 3711 | 7884 | 10 | 3512 | 3391 | 6903 |
| 5 | 4117 | 3775 | 7892 | 11 | 3392 | 3160 | 6552 |
| 6 | 3944 | 3665 | 7609 | 12 | 3761 | 3371 | 7132 |

Šioje lentelėje $X_{i1}$ – berniukų skaičiai; $X_{i2}$ – mergaičių skaičiai; $s = 12$, $r = 2$. Kai hipotezė teisinga, yra tik vienas nežinomas parametras $p$ – berniuko gimimo tikimybė. Pagal (2.5.4) gauname

$$X_n^2 = \sum_{i=1}^{12} \left( \frac{(U_{i1} - n_i\hat{p})^2}{n_i\hat{p}} + \frac{(U_{i2} - n_i(1-\hat{p}))^2}{n_i(1-\hat{p})} \right) = \sum_{i=1}^{12} \frac{(U_{i1} - n_i\hat{p})^2}{n_i\hat{p}(1-\hat{p})} =$$

$$= \frac{1}{1-\hat{p}}\left(\frac{1}{\hat{p}}\sum_{i=1}^{12}\frac{U_{i1}^2}{n_i} - U_{\cdot 1}\right) = 14,6211; \quad \hat{p} = \frac{U_{i1}}{n} = \frac{45682}{88273} = 0,51745.$$

Laisvės laipsnių skaičius yra $(s-1)(r-1) = 11$. Kadangi

$$pv_a = \mathbf{P}\{\chi^2_{11} > 14,6211\} = 0,2005,$$

tai atmesti hipotezę nėra pagrindo.

## 2.6. Pratimai

**2.1.** Raskite Pirsono statistikos $X_n^2$ iš (2.1.5) pirmuosius du momentus, kai tikrinamoji hipotezė: a) teisinga; b) neteisinga.

**2.2.** (**2.1** pratimo tęsinys.) Įrodykite, kad jei tikimybės $\pi_{i0} = 1/k$, tai $X_n^2$ dispersija yra $\boldsymbol{V}(X_n^2) = 2k^2(n-1)\{2(n-2)\sum_i \pi_i^3 - (2n-3)(\sum_i \pi_i^2)^2 + \sum_i \pi_i^2\}/n$ , o jeigu ir $\pi_i = \pi_{i0} = 1/k$, tai $\boldsymbol{V}(X_n^2) = 2(k-1)(n-1)/n$.

**2.3.** Įrodykite, kad teoremos 2.1.1 sąlygomis statistika

$$\tilde{X}_n^2 = n\sum_{i=1}^{k}(1-\pi_{i0})[H(U_i/n) - H(\pi_{i0})]^2$$

asimptotiškai$(n \to \infty)$ pasiskirsčiusi pagal chi kvadrato skirstinį su $k-1$ laisvės laipsnių; čia $H(x) = \arcsin x$.

**2.4.** Įrodykite, kad matrica $\mathbf{\Sigma} = \mathbf{E}_k - \sqrt{\boldsymbol{\pi}}(\sqrt{\boldsymbol{\pi}})^T$ yra idempotentinė, o jos rangas $k-1$; čia $\mathbf{E}_k$ – vienetinė $k \times k$ matrica, $\sqrt{\boldsymbol{\pi}} = (\sqrt{\pi_1}, ..., \sqrt{\pi_k})^T$, $0 < \pi_i < 1$, $i = 1, ..., k$, $\pi_1 + ... + \pi_k = 1$.

**2.5.** Skaitmenys 0, 1, 2, ..., 9 tarp pirmųjų 800 skaičiaus $\pi$ ženklų kartojasi atitinkamai 74, 92, 83, 79, 80, 73, 77, 75 ,76, 91 kartą. Ar galima šiuos duomenis interpretuoti kaip a. v. $\boldsymbol{U} \sim \mathcal{P}_{10}(800, \boldsymbol{\pi})$, $\boldsymbol{\pi} = (1/10, ..., 1/10)^T$ realizaciją?

**2.6.** Tikrinama hipotezė apie atsitiktinių skaičių lentelės korektiškumą, t. y. hipotezė, kad lentelėje skaitmenys 0, 1, 2, ..., 9 pasitaiko su vienodomis tikimybėmis $p = 0,1$. Hipotezė tikrinama $\chi^2$ kriterijumi. Koks turi būti imties didumas, kad ta hipotezė būtų atmesta su tikimybe, ne mažesne už 0,95, jei žinoma, kad 5 skaitmenys lentelėje pasirodo su tikimybėmis 0,11, o kiti 5 – su tikimybėmis 0,09 (kriterijaus reikšmingumo lygmuo yra 0,05).

**2.7.** Mendelis stebėjo, kokios žirnių sėklos gaunamos kryžminant augalus, kurių sėklos geltonos ir apvalios, su augalais, kurių sėklos raukšlėtos ir žalios. Rezultatai pateikti lentelėje kartu su teorinėmis tikimybėmis, apskaičiuotomis remiantis Mendelio paveldimumo teorija.

| Sėklos | Dažnumai | Tikimybės |
|---|---|---|
| Geltonos ir apvalios | 315 | 9/16 |
| Geltonos ir raukšlėtos | 101 | 3/16 |
| Žalios ir apvalios | 108 | 3/16 |
| Žalios ir raukšlėtos | 32 | 1/16 |
| $\Sigma$ | 556 | 1 |

Ar stebėjimo duomenys neprieštarauja Mendelio paveldimumo teorijai?

**2.8.** Kryžminant du kukurūzų tipus gauti keturi skirtingi augalų tipai. Pagal paprastąją Mendelio paveldimumo teoriją šie tipai turėtų pasirodyti su tikimybėmis 9/16, 3/16, 3/16 ir 1/16. Stebint 1301 augalą gauti tokie dažniai 773, 231, 238 ir 59. Su kokiu reikšmingumo lygmeniu $\chi^2$ kriterijus neprieštarauja Mendelio modeliui?

**2.9.** Nuskaitant prietaiso skalės parodymus, kai paskutinis skaitmuo įvertinamas iš akies, stebėtojai kartais nesąmoningai suteikia pirmenybę kai kuriems skaičiams. Lentelėje pateikti paskutiniųjų skaitmenų dažniai tam tikram stebėtojui atlikus 200 stebėjimų.

| Skaitmuo | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Dažnis | 35 | 16 | 15 | 17 | 17 | 19 | 11 | 16 | 30 | 24 |

Iš lentelės matome, kad skaitmenys 0 ir 8 pasirodo šiek tiek dažniau, palyginti su kitais. Ar galima daryti išvadą, kad stebėtojas daro sisteminę paklaidą?

**2.10.** Per 8000 bandymų nesutaikomi, sudarantys pilną įvykių grupę, įvykiai $A$, $B$ ir $C$ pasirodė 2014, 5012 ir 974 kartus. $\chi^2$ kriterijumi patikrinkite hipotezę, kad įvykių pasirodymo tikimybės yra $p_A = 0,5 - 2\alpha$, $p_B = 0,5 + \alpha$, $p_c = \alpha$, $0 < \alpha < 0,25$.

**2.11.** Tarp 2020 šeimų buvo užregistruota 527 šeimos, kuriose abu vaikai berniukai, 476 šeimos, kuriose abu vaikai mergaitės, ir 1017 šeimų, kuriose vaikai skirtingų lyčių. Ar galima tvirtinti, kad berniukų skaičius šeimose su dviem vaikais yra a) binominis a. d.; b) binominis, kai berniuko ir mergaitės gimimo tikimybės vienodos.

**2.12.** Diskretaus a. d. penkios nepriklausomos realizacijos yra 47, 46, 49, 53 ir 50. Ar galima tvirtinti, kad buvo stebimas Puasono a. d.?

**2.13.** Rezerfordo ir Geigerio bandymuose buvo registruojamas radioaktyvios medžiagos per 2608 ilgio 7.5 sek. periodus išspinduliuotų $\alpha$ dalelių skaičius. Rezultatai pateikti lentelėje ($i$ – išspinduliuotų dalelių skaičius, $V_i$ – periodų, per kuriuos buvo stebima $i$ dalelių, skaičius).

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $V_i$ | 57 | 203 | 383 | 525 | 532 | 408 | 273 | 139 | 45 | 27 | 10 | 4 | 2 |

Ar neprieštarauja gauti duomenys, kad buvo stebimas Puasono a. d.?

**2.14.** Kontroliniu prietaisu buvo išmatuotas atstumas $r$ (mikronais) nuo detalės svorio centro iki jos išorinio cilindro ašies. Matavimo rezultatai pateikti lentelėje ($r_i$ – reikšmės, $n_i$ – dažniai).

| $r_i$ | $n_i$ | $r_i$ | $n_i$ |
|---|---|---|---|
| 0 – 16 | 40 | 80 – 96 | 45 |
| 16 – 32 | 129 | 96 – 112 | 19 |
| 32 – 48 | 140 | 112 – 128 | 8 |
| 48 – 64 | 126 | 128 – 144 | 3 |
| 64 – 80 | 91 | 144 – 160 | 1 |

Remdamiesi $\chi^2$ kriterijumi, patikrinkite, ar stebėjimo rezultatai neprieštarauja prielaidai, kad stebimi atstumai pasiskirstę pagal Relėjaus dėsnį.

**2.15.** Nustatant 200 elektros lempučių degimo laiką $T$, gauti rezultatai pateikti lentelėje ($(a_{i-1}, a_i]$ – degimo laiko intervalai, $n_i$ – dažniai).

| $(a_{i-1} - -a_i$ | $n_i$ | $T_i$ | $n_i$ |
|---|---|---|---|
| 0 – 300 | 53 | 1800 – 2100 | 9 |
| 300 – 600 | 41 | 2100 – 2400 | 7 |
| 600 – 900 | 30 | 2400 – 2700 | 5 |
| 900 – 1200 | 22 | 2700 – 3000 | 3 |
| 1200 – 1500 | 16 | 3000 – 3300 | 2 |
| 1500 – 1800 | 12 | 3300 – 3600 | 0 |

Remdamiesi $\chi^2$ kriterijumi, patikrinkite, ar stebėjimo rezultatai neprieštarauja prielaidai, kad lemputės degimo laikas pasiskirstęs pagal eksponentinį dėsnį.

**2.16.** Lentelėje pateikti prapuolimo kampai 209 pašto balandžių, kai atliekant bandymą buvo bandoma paveikti jų „vidinį laikrodį“ (žr. [21]).

| Kryptis | Dažnis | Kryptis | Dažnis |
|---|---|---|---|
| 0° – | 26 | 180° – | 14 |
| 30° – | 22 | 210° – | 11 |
| 60° – | 26 | 240° – | 12 |
| 90° – | 30 | 270° – | 5 |
| 120° – | 29 | 300° – | 5 |
| 150° – | 18 | 330° – | 11 |

Duomenys sugrupuoti į 30° ilgio intervalus. Lentelėje nurodyti kampai $\varphi_i$, atitinkantys i-ojo intervalo pradžią ir patekusių į i-ąjį intervalą dažniai $n_i$, $i = 1, \ldots 12$. Patikrinkite hipotezę, kad turimi duomenys neprieštarauja prielaidai, jog prapuolimo kampas turi Mizeso skirstinį $M(\mu, \theta)$.

**2.17.** Lentelėje pateikta smėlio grūdelių orientacija plokštumoje (žr. [21]).

| Kampas | Kiekis | Kampas | Kiekis | Kampas | Kiekis |
|---|---|---|---|---|---|
| 0° – | 244 | 60° – | 326 | 120° – | 322 |
| 10° – | 262 | 70° – | 340 | 130° – | 295 |
| 20° – | 246 | 80° – | 371 | 140° – | 230 |
| 30° – | 290 | 90° – | 401 | 150° – | 256 |
| 40° – | 284 | 100° – | 382 | 160° – | 263 |
| 50° – | 314 | 110° – | 332 | 170° – | 281 |

Kampai sugrupuoti į 10° ilgio intervalus (nurodoma grupavimo intervalo pradžia). Gretimuose stulpeliuose nurodomi smėlio grūdelių, kurių orientacija patenka į atitinkamus intervalus skaičiai.

Padvigubinę kampus perveskite duomenis į intervalą $[0° - 360°]$. Patikrinkite hipotezę, kad stebėtas atsitiktinis kampas turi Mizeso skirstinį $M(\mu, \theta)$.

**2.18.** Didumo $n = 100$ imties realizacija pateikta lentelėje.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 338 | 336 | 312 | 322 | 381 | 302 | 296 | 360 | 342 | 334 |
| 348 | 304 | 323 | 310 | 368 | 341 | 298 | 312 | 322 | 350 |
| 304 | 302 | 336 | 334 | 304 | 292 | 324 | 331 | 324 | 334 |
| 314 | 338 | 324 | 292 | 298 | 342 | 338 | 331 | 325 | 324 |
| 326 | 314 | 312 | 362 | 368 | 321 | 352 | 304 | 302 | 332 |
| 314 | 304 | 312 | 381 | 290 | 322 | 326 | 316 | 328 | 340 |
| 324 | 320 | 364 | 304 | 340 | 290 | 318 | 332 | 354 | 324 |
| 304 | 321 | 356 | 366 | 328 | 332 | 304 | 282 | 330 | 314 |
| 342 | 322 | 362 | 298 | 316 | 298 | 332 | 342 | 316 | 326 |
| 308 | 321 | 302 | 304 | 322 | 296 | 322 | 338 | 324 | 323 |

Modifikuotuoju $\chi^2$ kriterijumi (grupavimo intervalų skaičius $k = 8$) patikrinkite hipotezę, kad buvo stebimas normalusis atsitiktinis dydis.

**2.19.** Lentelėje pateikti duomenys, apibūdinantys tam tikro elemento koncentraciją nesureagavusiame likutyje pasibaigus cheminiam procesui.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 51 | 8 | 47 | 8 | 5 | 56 | 12 | 4 | 5 | 4 | 4 | 7 | 6 | 9 |
| 30 | 25 | 12 | 3 | 22 | 5 | 15 | 4 | 4 | 29 | 15 | 4 | 2 | 18 | 41 |
| 3 | 5 | 54 | 110 | 24 | 16 | 2 | 37 | 20 | 2 | 6 | 7 | 16 | 2 | 14 |
| 68 | 10 | 16 | 11 | 78 | 6 | 17 | 7 | 11 | 21 | 15 | 24 | 6 | 32 | 8 |
| 11 | 4 | 14 | 45 | 17 | 10 | 15 | 20 | 4 | 65 | 10 | 3 | 5 | 11 | 13 |
| 35 | 11 | 34 | 3 | 4 | 12 | 7 | 6 | 62 | 13 | 36 | 26 | 6 | 11 | 6 |
| 13 | 1 | 4 | 36 | 18 | 10 | 37 | 28 | 4 | 12 | 31 | 14 | 3 | 11 | 6 |
| 4 | 10 | 38 | 6 | 11 | 24 | 9 | 4 | 5 | 8 | 135 | 22 | 6 | 18 | 49 |
| 17 | 9 | 32 | 27 | 2 | 12 | 8 | 93 | 3 | 9 | 10 | 3 | 14 | 33 | 72 |
| 14 | 4 | 9 | 10 | 19 | 2 | 5 | 21 | 8 | 25 | 30 | 20 | 12 | 19 | 16 |

Modifikuotuoju $\chi^2$ kriterijumi (grupavimo intervalų skaičius $k = 10$) patikrinkite hipotezę, kad buvo stebimas lognormalusis atsitiktinis dydis.

**2.20.** Ląsteles veikiant rentgeno spinduliais, jose keičiasi kai kurios chromosomos. Lentelėje pateikti kelių nepriklausomų bandymų serijų duomenys ($i$ – chromosomų pasikeitimų skaičius, $n_{ik}$ – ląstelių su $i$ pasikeitimų $k$-ajame eksperimente skaičius).

| $i$ | 0 | 1 | 2 | $\geq 3$ | $\sum n_{ik}$ |
|---|---|---|---|---|---|
| $n_{i1}$ | 280 | 75 | 12 | 1 | 368 |
| $n_{i2}$ | 593 | 143 | 20 | 3 | 759 |
| $n_{i3}$ | 639 | 141 | 13 | 0 | 793 |
| $n_{14}$ | 359 | 109 | 13 | 1 | 482 |

Patikrinkite hipotezę, kad visos 4 imtys gautos stebint atsitiktinius dydžius, kurių skirstiniai yra a) Puasono; b) tie patys Puasono.

**2.21.** Modifikuotuoju chi kvadrato kriterijumi patikrinkite hipotezę, kad pateikti $n = 100$ skaičių yra normaliojo a. d. realizacija.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 41 | 30 | 37 | 25 | 32 | 28 | 35 | 28 | 51 | 36 | 26 | 43 | 25 | 27 |
| 39 | 21 | 45 | 39 | 25 | 29 | 43 | 66 | 25 | 24 | 56 | 29 | 31 | 41 | 41 |
| 36 | 57 | 36 | 48 | 25 | 36 | 48 | 24 | 48 | 22 | 40 | 7 | 31 | 24 | 32 |
| 53 | 33 | 46 | 22 | 33 | 25 | 37 | 34 | 32 | 41 | 36 | 19 | 32 | 25 | 19 |
| 19 | 37 | 20 | 21 | 48 | 44 | 35 | 19 | 44 | 34 | 29 | 48 | 38 | 43 | 48 |
| 35 | 42 | 37 | 35 | 36 | 58 | 45 | 34 | 40 | 37 | 21 | 41 | 11 | 41 | 27 |
| 50 | 24 | 37 | 39 | 33 | 45 | 39 | 43 | 21 | 34 | | | | | |

Pakoreguokite kriterijų atsižvelgdami į tai, kad duomenys suapvalinti.

**2.22.** Dviejose nepriklausomose didumo 500 imtyse buvo užregistruota laikrodžių, išstatytų įvairių taisyklų vitrinose, rodmenys. Duomenys sugrupuoti į 12 intervalų (0 reiškia intervalą nuo 0 h iki 1 h; 1 – intervalą nuo 1 h iki 2 h ir t. t.) ir surašyti į lentelę.

| Imtis | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | $\Sigma$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 41 | 34 | 54 | 39 | 49 | 45 | 41 | 33 | 37 | 41 | 47 | 39 | 500 |
| 2 | 36 | 47 | 41 | 47 | 49 | 45 | 32 | 37 | 40 | 41 | 37 | 48 | 500 |

Remdamiesi $\chi^2$ kriterijumi patikrinkite hipotezę, kad abiejose imtyse laikrodžių rodmenų patekimo į visus intervalus tikimybės yra vienodos.

**2.23.** Viename sraute iš 300 stojančiųjų pažymius „nepatenkinamai“, „patenkinamai“, „gerai“ ir „labai gerai“ gavo atitinkamai 33, 43, 80 ir 144; kito srauto stojantieji atitinkamai 39, 35, 72 ir 154. Ar galima laikyti, kad abiejų srautų stojantieji pasirengę vienodai?

**2.24.** Paleidus raketą 87 kartus, buvo gauti tokie duomenys apie atstumą $X$(m) ir nukrypimą $Y$(kampo minutės).

| $x_i \backslash y_j$ | (-250,-50) | (-50,50) | (50,250) | $\Sigma$ |
|---|---|---|---|---|
| 0 – 1200 | 5 | 9 | 7 | 21 |
| 1200 – 1800 | 7 | 5 | 5 | 21 |
| 1800 – 2700 | 8 | 21 | 16 | 45 |
| $\Sigma$ | 20 | 35 | 32 | 87 |

Ar šie požymiai nepriklausomi?

**2.25.** Tiriant granulometrinę kvarco sudėtį Anykščių ir Afrikos smėlio pavyzdžiuose, buvo gauta duomenų apie jo grūdelių didžiosios ašies ilgį. Remiantis pavyzdžių granulometrine sudėtimi daromos tam tikros išvados apie geologines smėlio susidarymo sąlygas. Pateikiami duomenys sugrupuoti vienodo ilgio intervalais ($X_i$ – $i$-ojo intervalo vidurys).

| $X_i$ | 9 | 13 | 17 | 21 | 25 | 29 | 33 | 37 | 41 | 45 | 49 | $\Sigma$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anykščių smėlis | 4 | 12 | 35 | 61 | 52 | 23 | 7 | 4 | 2 | 1 | 0 | 201 |
| Afrikos smėlis | 0 | 6 | 10 | 12 | 13 | 12 | 15 | 12 | 11 | 7 | 4 | 102 |

Remdamiesi $\chi^2$ kriterijumi patikrinkite hipotezę, kad imtys yra to paties atsitiktinio dydžio.

## 2.7. Atsakymai ir nurodymai.

**2.1.** a) $\mathbf{E}(X_n^2) = k-1$, $\boldsymbol{V}(X_n^2) = 2(k-1) + [-2(k-1) - k^2 + \sum_i (1/\pi_{i0})]/n = 2(k-1) + O(1/n)$; b) $\mathbf{E}(X_n^2) = k - 1 + n\sum_i((\pi_i - \pi_{i0})^2/\pi_{i0}) + \sum_i((\pi_i - \pi_{i0})(1-\pi_i)/\pi_{i0})$, $\boldsymbol{V}(X_n^2) = (2(n-1)/n)\{2(n-2) - (2n-3)(\sum_i(\pi_i^2/\pi_{i0}))^2 - 2\sum_i(\pi_i^2/\pi_{i0})\sum_i(\pi_i/\pi_{i0}) + 3\sum_i(\pi_i^2/\pi_{i0}^2)\} - [(\sum_i(\pi_i/\pi_{i0}))^2 - \sum_i(\pi_i/\pi_{i0}^2)]/n$. **Nurodymas.**

$$\boldsymbol{V}(X_n^2) = \mathbf{E}(\sum_i (U_i^2/(n\pi_{i0})))^2 - (\sum_i(\mathbf{E}U_i^2/(n\pi_{i0})))^2.$$

Raskite momentus $\mathbf{E}(U_i^2)$, $\mathbf{E}(U_i^4)$, $\mathbf{E}(U_i^2U_j^2)$ (pavyzdžiui, naudodami faktorialinių momentų išraiškas) ir sutraukite panašiuosius narius. **2.3. Nurodymas.** Įrodykite, kad a. v.

$$\sqrt{n}(\sqrt{1-\pi_{10}}(H(U_1/n) - H(\pi_{10})), \ldots, \sqrt{1-\pi_{k0}}(H(U_k/n) - H(\pi_{k0})))^T$$

yra asimptotiškai normalusis $N_k(\mathbf{0}, \boldsymbol{\Sigma})$ su ta pačia kovariacine matrica $\boldsymbol{\Sigma}$ kaip ir 2.1.1 teoremoje. **2.5.** Statistika $X_n^2$ įgijo reikšmę 5,125 ir $pv_a = \mathbf{P}\{\chi_9^2 > 5,125\} = 0,8233$; duomenys neprieštarauja iškeltai hipotezei. **2.6.** $n \geq 881$. **2.7.** Statistika $X_n^2$ įgijo reikšmę 0,47 ir $pv_a = \mathbf{P}\{\chi_3^2 > 0,47\} = 0,9254$; duomenys neprieštarauja iškeltai hipotezei. **2.8.** Statistika $X_n^2$ įgijo reikšmę 9,2714 ir $pv_a = \mathbf{P}\{\chi_3^2 > 9,2714\} = 0,0259$; hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,0259. **2.9.** Tikrinant hipotezę $H : \pi_i = 1/10$, $i = 1, \ldots, 10$ statistika $X_n^2$ įgijo reikšmę 24,9 ir $pv_a = \mathbf{P}\{\chi_9^2 > 24,9\} = 0,0031$; hipotezė atmestina. Atlikdami tolesnę analizę patikrinkime hipotezę, kad skaitmenų 0 arba 8 pasirodymo tikimybė 0,2. Esant teisingai hipotezei $S = U_1 + U_8 \sim B(n, 0,2)$ ir įgijo reikšmę 65. Taigi $pv = \mathbf{P}\{S \geq 65\} = 0,00002$. Išvada: stebėtojas daro sisteminę paklaidą. **2.10.** Parametro $\alpha$ įvertis yra $\hat{\alpha} = 0,1235$. Statistika $X_n^2(\hat{\alpha})$ įgijo reikšmę 0,3634 ir $pv_a = \mathbf{P}\{\chi_1^2 > 0,3634\} = 0,5466$; duomenys neprieštarauja iškeltai hipotezei. **2.11.** Berniuko gimimo tikimybės įvertis yra $\hat{p} = 0,5126$. Statistika $X_n^2(\hat{p})$ įgijo reikšmę 0,1159 ir $pv_a = \mathbf{P}\{\chi_1^2 > 0,1159\} = 0,7335$; duomenys neprieštarauja iškeltai hipotezei. Jeigu tartume, kad berniuko ir mergaitės gimimo tikimybė vienoda ir lygi 1/2, tai jokių parametrų vertinti nereikia. Statistika $X_n^2$ įgijo reikšmę 2,6723 ir $pv_a = \mathbf{P}\{\chi_2^2 > 2,6723\} = 0,2629$; duomenys neprieštarauja ir šiai hipotezei. **2.12.** Statistika $X_n^2$, turinti asimptotinį chi kvadrato skirstinį su 4 laisvės laipsniais, įgijo reikšmę 0,6122 ir $pv_a = \mathbf{P}\{\chi_4^2 > 0,6122\} = 0,9617$; duomenys neprieštarauja iškeltai hipotezei. **Nurodymas.** Pasiremkite tokiu faktu: esant teisingai hipotezei imties $(X_1, \ldots, X_n)^T$ sąlyginis skirstinys, kai suma $S = X_1 + \ldots + X_n$ fiksuota, yra polinominis $\mathcal{P}_n(S, \boldsymbol{\pi}_0)$, $\boldsymbol{\pi}_0 = (1/n, \ldots, 1/n)^T$. **2.13.** Parametro $\lambda$ įvertis yra $\hat{\lambda} = 3,8666$. Statistika $X_n^2(\hat{\lambda})$ įgijo reikšmę 13,0146 ir $pv_a = \mathbf{P}\{\chi_{10}^2 > 13,0146\} = 0,2229$; duomenys neprieštarauja iškeltai hipotezei. Skaičiuojant statistikos reikšmę du paskutinieji intervalai buvo sujungti. **2.14.** Parametro $\sigma^2$ DT įvertis yra $\hat{\sigma}^2 = 1581,65$. Statistika $X_n^2(\hat{\sigma}^2)$ įgijo reikšmę 2,6931 ir $pv_a = \mathbf{P}\{\chi_7^2 > 2,6931\} = 0,9119$; duomenys neprieštarauja iškeltai hipotezei. Skaičiuojant statistikos reikšmę du paskutinieji intervalai buvo sujungti. **2.15.** Parametro $\theta$ DT įvertis yra $\hat{\theta} = 878,4$. Statistika $X_n^2(\hat{\theta})$ įgijo reikšmę 4,0477 ir $pv_a = \mathbf{P}\{\chi_8^2 > 4,0477\} = 0,8528$; duomenys neprieštarauja iškeltai hipotezei. Skaičiuojant statistikos reikšmę trys paskutinieji intervalai buvo sujungti. **2.16.**

Parametrų įverčiai $\hat{\mu} = 96,16°, \hat{\theta} = 0,6854$; statistikos $R_n(\hat{\mu},\hat{\theta})$ ir $X_n^2(\hat{\mu},\hat{\theta})$ įgijo reikšmes 9,960 ir 10,175; atitinkamos asimptotinės $P$ reikšmės $pv_a = 0,354$ ir $pv_a = 0,337$. Hipotezė neatmetama. **2.17.** Parmetrų įverčiai $\hat{\mu} = 180,8°, \hat{\theta} = 0,2047$; statistikos $R_n(\hat{\mu},\hat{\theta})$ ir $X_n^2(\hat{\mu},\hat{\theta})$ įgijo reikšmes 24,858 ir 24,641; atitinkamos asimptotinės $P$ reikšmės $pv_a = 0,0519$ ir $pv_a = 0,0550$. Reikšmingumo lygmens $\alpha = 0,05$ kriterijumi hipotezė neatmetama. Turint omenyje tokį didelį stebėjimų skaičių, matyt, galima daryti išvadą, kad tokio tipo duomenims aprašyti Mizeso modelis yra tinkamas. **2.18.** Parametrų $\mu$ ir $\sigma$ DT įverčiai yra $\hat{\mu} = \bar{X} = 324,57$ ir $\hat{\sigma} = 20,8342$. Parenkame $k = 8$ intervalus. Tada $X_n^2 = 8,0$, $Q_n = 2,8302$, $Y_n^2 = 10,8302$ ir $pv_a = \mathbf{P}\{\chi_7^2 > 10,8302\} = 0,1462$. Hipotezė neatmetama, jei kriterijaus reikšmingumo lygmuo neviršija 0,1462. **2.19.** Perėję prie logaritmų $Y_i = \ln(X_i)$ gauname DT įverčius $\hat{\mu} = \bar{Y} = 2,4589$ ir $\hat{\sigma} = 0,9529$. Parenkame $k = 10$ intervalų. Tada $X_n^2 = 4,1333$, $Q_n = 1,0668$, $Y_n^2 = 5,2001$ ir $pv_a = \mathbf{P}\{\chi_9^2 > 5,2001\} = 0,8165$. Duomenys neprieštarauja iškeltai hipotezei. **2.20.** a) kiekvienoje imtyje įvertiname parametrą $\lambda$, apskaičiuojame statistikų $X_{n_i}^2(\hat{\lambda}_i)$ reikšmes ir jas sudedame. Gauname statistikos, kuri esant taisingai hipotezei asimptotiškai turi chi kvadrato skirstinį su 4 laisvės laipsniais, realizaciją. Gautoji reikšmė yra 2,5659 ir $pv_a = \mathbf{P}\{\chi_4^2 > 2,5659\} = 0,6329$. Hipotezę atmesti nėra pagrindo; b) įvertiname parametrą $\lambda$ pagal jungtinę imtį ir gauname $\hat{\lambda} = 0,2494$. Apskaičiuojame statistikų $X_{n_i}(\hat{\lambda})$ reikšmes ir jas sudedame; gauname 10,2317. Kadangi $pv_a = \mathbf{P}\{\chi_7^2 > 10,2317\} = 0,1758$, tai ir ši hipotezė neatmetama. Skaičiuojant du paskutinieji intervalai buvo sujungti. **2.21.** Neatsižvelgiant į duomenų apvalinimą gaunama $X_n^2 = 4,160$, $Q_n = 0,172$, $Y_n^2 = 4,332$ ir $pv_a = \mathbf{P}\{\chi_7^2 > 4,332\} = 0,741$. Duomenys neprieštarauja iškeltai hipotezei. Atlikę korekciją atsižvelgdami į duomenų apvalinimą, gauname $X_n'^2 = 3,731$, $Q_n' = 0,952$, $Y_n'^2 = 4,683$ ir $pv_a' = \mathbf{P}\{\chi_7^2 > 4,683\} = 0,699$. Duomenys neprieštarauja iškeltai hipotezei. Reikia pažymėti, kad $P$ reikšmės $pv$ ir $pv'$ gerokai skiriasi. **Nurodymas.** Kadangi duomenys suapvalinti iki sveikųjų skaičių, tai gautuosius intervalų galus $a_i$ reikia pastumti iki artimiausių $m \pm 0,5$ pavidalo rėžių ($m$ – sveikasis skaičius) ir apskaičiuoti statistikos reikšmę naudojant naujai gautus rėžius $a_i'$. **2.22.** Statistika (2.5.4) įgijo reikšmę 8,51 ir $pv_a = \mathbf{P}\{\chi_{11}^2 > 8,51\} = 0,6670$; duomenys neprieštarauja iškeltai hipotezei. **2.23.** Statistika (2.5.4) įgijo reikšmę 2,0771 ir $pv_a = \mathbf{P}\{\chi_3^2 > 2,0771\} = 0,5566$; duomenys neprieštarauja iškeltai hipotezei. **2.24.** Statistika (2.4.3) įgijo reikšmę 3,719 ir $pv_a = \mathbf{P}\{\chi_4^2 > 3,719\} = 0,4454$; duomenys neprieštarauja iškeltai hipotezei. **2.25.** Statistika (2.5.4) įgijo reikšmę 75,035 (trys paskutinieji intervalai sujungti) ir $pv_a = \mathbf{P}\{\chi_7^2 > 75,035\} < 10^{-12}$; hipotezė atmetama.

# 3 skyrius

# Glodūs Neimano ir Bartono kriterijai

Ankstesniame skyriuje nagrinėtas $\chi^2$ suderinamumo kriterijus yra gana bendras ir turi geras asimptotines savybes. Tikrinant paprastąją suderinamumo hipotezę $H_0 : X \sim F_0(x)$, kriterijaus statistikos $X_n^2$ asimptotinis skirstinys yra $\chi^2$ skirstinys su $k-1$ laisvės laipsniu, nepriklausomai nuo to, kokia yra pasiskirstymo funkcija $F_0(x)$. Analogiškai, tikrinant sudėtinę suderinamumo hipotezę $H_0 : X \sim F_0(x|\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_s)^T$, statistikos $X^2(\hat{\boldsymbol{\theta}})$ skirstinys yra $\chi^2$ skirstinys su $k-1-s$ laisvės laipsniu, nepriklausomai nuo funkcijos $F_0$ ir parametro $\boldsymbol{\theta}$, jeigu parametras vertinamas DT (ar jam ekvivalenčiu) metodu naudojant *grupuotąją imtį*. Jeigu parametrui $\boldsymbol{\theta}$ vertinti naudojamas DT metodas pradiniams (negrupuotiems) duomenims, tai statistikos $X^2(\hat{\boldsymbol{\theta}})$ skirstinys netgi ir asimptotiškai priklauso ir nuo funkcijos $F_0$, ir nuo parametro $\boldsymbol{\theta}$.

Nagrinėjant skirstinius, priklausančius tik nuo poslinkio ir mastelio parametrų, 2.3 skyrelyje pateiktas modifikuotas $\chi^2$ kriterijus, kurio statistika $Y_n^2$ asimptotiškai turi $\chi^2$ skirstinį su $k-1$ laisvės laipsniu nepriklausomai nuo hipotetinės pasiskirstymo funkcijos ir nuo nežinomo parametro. Be to, grupavimo intervalų galai gali priklausyti nuo imties.

Pagrindinis $\chi^2$ tipo suderinamumo kriterijų trūkumas yra tai, kad jie sudaromi remiantis mažiau informatyvia grupuotąja imtimi. Be to, kriterijus yra asimptotinis ir, norint pasiekti reikiamą aproksimacijos tikslumą, į kiekvieną intervalą turi patekti vidutiniškai pakankamai stebinių. Todėl intervalai negali būti trumpi ir hipotezė apie polinominio skirstinio parametrų reikšmes (2.2.2) gali gerokai skirtis nuo tikrinamos hipotezės (2.2.1) (žr. 2.1.1 pastabą).

## 3.1. Suderinamumo kriterijai, remiantis negrupuotais duomenimis

Tarkime, kad $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis, gauta stebint a. d. $X$. Tikriname paprastąją suderinamumo hipotezę

$$H_0 : X \sim F_0(x), \tag{3.1.1}$$

čia $F_0(x)$ žinoma absoliučiai tolydi pasiskirstymo funkcija su tankio funkcija $f_0(x) = F_0'(x)$, kai sudėtinė alternatyva yra

$$H : X \sim F(x) \in \mathcal{F}, \tag{3.1.2}$$

čia $\mathcal{F}$ yra absoliučiai tolydžių skirstinių aibė su tankio funkcijomis $f(x) = F'(x)$. Aibės $\mathcal{F}$ sudėtį aptarsime vėliau.

Suderinamumo kriterijai, naudojantys pradinę negrupuotą imtį, tiesiogiai ar netiesiogiai susiję su integraline transformacija

$$Y_i = F_0(X_i) = \int_{-\infty}^{X_i} f_0(x)dx, \quad i = 1, ..., n. \tag{3.1.3}$$

Jeigu hipotezė $H_0$ teisinga ir $X_i \sim F_0(x)$, tai a. d. $Y_1, ..., Y_n$ yra nepriklausomi vienodai tolygiai pasiskirstę intervale $[0,\ 1]$, t. y. $Y_i \sim U(0,\ 1), i = 1, ..., n$.

Jeigu hipotezė $H_0$ neteisinga ir $X_i \sim F(x) \in \mathcal{F}$, tai a. d. $Y_1, ..., Y_n$ taip pat nepriklausomi ir vienodai pasiskirstę intervale $[0,\ 1]$. Tačiau jų skirstinys nėra tolygusis. Remiantis transformacija (3.1.3), a. d. $Y_i$ tankio funkcija

$$g(y) = \frac{f(F_0^{-1}(y))}{f_0(F_0^{-1}(y))}, \quad 0 \le y \le 1. \tag{3.1.4}$$

Pradinis uždavinys suvedamas į tokį: remiantis imtimi $Y_1, ..., Y_n$ reikia patikrinti hipotezę $H_0 : Y_i \sim U(0,\ 1)$, kai sudėtinė alternatyva yra

$$Y_i \sim g(y) \in \mathcal{G}, \quad 0 \le y \le 1, \tag{3.1.5}$$

čia $\mathcal{G}$ yra aibė tankių $g(y)$, kurie gaunami įrašant į (3.1.4) šeimos $\mathcal{F}$ tankius $f(x)$.

Pažymėsime, kad kriterijai, kurių statistikos yra imties $Y_1, ..., Y_n$ funkcijos, kai hipotezė teisinga, nepriklauso nuo $F_0$ ne tik asimptotiškai, bet ir esant baigtiniam imties didumui $n$.

## 3.2. Neimano ir Bartono suderinamumo kriterijus

### Kriterijaus sudarymo idėja

Tikrinant suderinamumo hipotezes sunku apibrėžti alternatyviųjų hipotezių aibę $\mathcal{F}$. Neimanas [23] pasiūlė apibrėžti tolygiojo skirstinio alternatyvas imant pakankamai plačią intervalo [0, 1] skirstinių, priklausančių nuo keleto parametrų, aibę, kuriems kintant skirstiniai gali būti glodžiai priartinti prie tolygiojo skirstinio:

$$\mathcal{G} = \{g(y|\boldsymbol{\theta}), \quad 0 \leq y \leq 1, \quad \boldsymbol{\theta} \in \mathbf{R}^k\}, \tag{3.2.1}$$

čia tankio funkcijos

$$g(y|\boldsymbol{\theta}) = \frac{1}{c(\boldsymbol{\theta})} \exp\{\sum_{r=1}^{k} \theta_r \pi_r(y - 1/2)\}, \quad 0 \leq y \leq 1, \quad k = 1, 2, 3, ... \tag{3.2.2}$$

Normuojanti konstanta $c(\boldsymbol{\theta}) = c(\theta_1, ..., \theta_k)$ parenkama taip, kad integralas nuo tankio būtų lygus 1:

$$c(\boldsymbol{\theta}) = \int_0^1 \exp\{\sum_{r=1}^{k} \theta_r \pi_r(y - 1/2)\} dy,$$

o $\pi_r(z)$ – ortonormuoti intervale $[-1/2,\ 1/2]$ Ležandro polinomai.

**3.2.1 pastaba.** Ležandro polinomas $L_r(z)$ apibrėžiamas formule

$$L_r(z) = \frac{1}{r!2^r} \frac{d^r}{dz^r} (z^2 - 1)^r$$

ir tenkina ortogonalumo sąlygas

$$\int_{-1}^{1} L_r(z) L_s(z) dz = \begin{cases} 0, & r \neq s, \\ \frac{2}{2r+1}, & r = s. \end{cases}$$

Atlikę keitimą $\pi_r(z) = \sqrt{2r+1} L_r(2z)$, gauname ortonormuotų intervalo $[-1/2,\ 1/2]$ polinomų sistemą $\pi_1(z), \pi_2(z), ...$ Pirmieji keturi polinomai yra

$$\pi_1(z) = 2\sqrt{3}z, \quad \pi_2(z) = \sqrt{5}(6z^2 - 1/2),$$

$$\pi_3(z) = \sqrt{7}(20z^3 - 3z), \quad \pi_4(z) = 3(70z^4 - 15z^2 + 3/8). \tag{3.2.3}$$

Prilyginę (3.1.4) tankiui (3.2.2) ir atlikę atvirkštinį keitimą $x = F_0^{-1}(y)$ grįžtame prie pradinio uždavinio. Gauname, kad Neimano pasiūlymas reiškia, kad tikrinama paprastoji hipotezė $H_0 : X \sim F_0(x)$, kai alternatyvų aibės $\mathcal{F}$ skirstinių tankiai $f(x)$ turi tokį pavidalą:

$$f(x) = f_0(x) \frac{1}{c(\boldsymbol{\theta})} \exp\{\sum_{r=1}^{k} \theta_r \pi_r(F_0(x) - 1/2)\}, \quad \boldsymbol{\theta} \neq \mathbf{0}, \quad \boldsymbol{\theta} \in \mathbf{R}^k. \tag{3.2.4}$$

Matome, kad alternatyvos gaunamos deformuojant hipotetinį tankį $f_0(x)$, t. y. padauginant jį iš $F_0(x)$ funkcijos, priklausančios nuo parametro $\boldsymbol{\theta}$. Alternatyvų aibė vienareikšmiškai (neskaitant parametro $\boldsymbol{\theta}$) apibūdinama naudojant hipotetinę funkciją $F_0(x)$.

**3.1.2 pastaba.** Skirstiniai (3.2.4) nėra iš tų, kurie naudojami stebimų a. d. skirstiniams apibūdinti. Tačiau kadangi šeima (3.2.4) gana plati, tai joje atsiras skirstinių, kurie bus artimi praktiškai įdomioms alternatyvoms. Todėl tikėtina, jei surastas kriterijus bus galingas su visomis aibės (3.2.4) alternatyvomis, tai jis bus galingas ir su kitomis praktiškai įdomiomis alternatyvomis, nors jos ir nepriklausys aibei (3.2.4).

**Kriterijaus statistika**

Hipotezė $H_0 : Y_i \sim U(0,\ 1)$, kai sudėtinė alternatyva yra (3.2.1), tampa parametrine hipoteze

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0, \tag{3.2.5}$$

kai alternatyvoje tvirtinama, kad nors vienas iš $\theta_i \neq 0$.

Tankis (3.2.1) priklauso $k$-parametrei kanoninio pavidalo eksponentinių skirstinių šeimai. Tikėtinumo funkcija

$$L = L(\boldsymbol{\theta}) = \exp\{\sum_{r=1}^{k} \theta_r T_r - n \ln c(\boldsymbol{\theta})\},$$

o logtikėtinumo funkcija

$$\ell = \ell(\boldsymbol{\theta}) = \{\sum_{r=1}^{k} \theta_r T_r - n \ln c(\boldsymbol{\theta})\}, \tag{3.2.6}$$

čia $\boldsymbol{T} = (T_1, ..., T_k)^T$ yra parametro $\boldsymbol{\theta}$ pakankamoji statistika

$$T_r = \sum_{i=1}^{n} \pi_r(Y_i - 1/2), \quad r = 1, ..., k.$$

Gauname lygčių sistemą parametro $\boldsymbol{\theta}$ DT įvertiniui $\hat{\boldsymbol{\theta}}$ rasti

$$\dot{\ell}_{\theta_r} = T_r - n[\ln c(\boldsymbol{\theta})]'_{\theta_r} = 0, \quad r = 1, ..., k. \tag{3.2.7}$$

Fišerio informacinė matrica

$$I(\boldsymbol{\theta}) = -\mathbf{E}\ddot{\ell}(\boldsymbol{\theta}) = n[(\ln c(\boldsymbol{\theta}))''_{\theta_r \theta_s}]_{k \times k}. \tag{3.2.8}$$

Tikrinant hipotezę (3.2.5) natūralu naudoti tikėtinumų santykio kriterijų. Tikėtinumų santykio statistika

$$-2 \ln \Lambda = -2 \ln \frac{\max_{\theta_1 = ... = \theta_k = 0} L(\theta_1, ..., \theta_k)}{\max_{\theta_1, ..., \theta_k} L(\theta_1, ..., \theta_k)} = 2 \ln L(\hat{\theta}_1, ..., \hat{\theta}_k) \xrightarrow{d} \chi^2_k,$$

kai hipotezė $H_0$ teisinga (žr. A priedą, (7.1.10)). Vietoje tikėtinumų santykio statistikos patogiau naudoti jam ekvivalentaus informantinio kriterijaus statistiką (žr. A priedą, (7.1.9)), kuriai rasti nereikia įvertinių $\hat{\theta}_1, ..., \hat{\theta}_k$. Kriterijaus statistika

$$R_I = \dot{\ell}^T(\boldsymbol{\theta}_0)(-\ddot{\ell}(\boldsymbol{\theta}_0))^{-1}\dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_k, \tag{3.2.9}$$

čia $\boldsymbol{\theta}_0$ yra hipotetinė parametro $\boldsymbol{\theta}$ reikšmė.

Kai $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = \mathbf{0}$, tai, remdamiesi polinomų $\pi_r(z)$ ortonormuotumu, gauname

$$c(\mathbf{0}) = 1, \quad \dot{c}(\mathbf{0}) = \mathbf{0}, \quad \ddot{c}(\mathbf{0}) = \boldsymbol{I}.$$

Taigi statistika $R_I$ turi tokį pavidalą

$$R_I = \frac{1}{n}(T_1^2 + ... + T_k^2). \tag{3.2.10}$$

**Neimano ir Bartono kriterijus**

Hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$R_I > \chi^2_\alpha(k), \tag{3.2.11}$$

arba P reikšmių terminais, kai

$$pv_a = \mathbf{P}\{\chi^2_k > r_I\} < \alpha,$$

čia $r_I$ yra statistikos $R_I$ stebinys.

**3.2.1 pavyzdys** (**2.1.1** pavyzdžio tęsinys). Patikrinsime prielaidą, kad 2.1.1 pavyzdžio duomenys gauti stebint a.d. $Y \sim U(0,\ 1)$.

Apskaičiuojame statistikų realizacijas

$$\frac{T_1^2}{n} = 0,00011, \quad \frac{T_1^2 + T_2^2}{n} = 0,0401, \quad \frac{T_1^2 + T_2^2 + T_3^2}{n} = 0,0858, \quad \frac{T_1^2 + T_2^2 + T_3^2 + T_4^2}{n} = 0,2954.$$

Atitinkamos asimptotinės P reikšmės

$$pv_a = \mathbf{P}\{\chi^2_1 > 0,00011\} = 0,9915, \quad pv_a = \mathbf{P}\{\chi^2_2 > 0,0401\} = 0,9801,$$

$$pv_a = \mathbf{P}\{\chi^2_3 > 0,0858\} = 0,9935, \quad pv_a = \mathbf{P}\{\chi^2_4 > 0,2954\} = 0,9901.$$

Hipotezė neatmetama kriterijais, kuriuos sudarant naudoti 1, 2, 3, 4 parametrai.

## 3.3. Suderinamumo kriterijai, grindžiami beta skirstiniu

Atlikus integralinę transformaciją (3.1.1) ir perėjus prie hipotezės $H_0 : Y_i \sim U(0,\ 1)$ tikrinimo, alternatyvomis galima imti ir kitokias negu Neimano pasiūlyta (3.1.4) intervalo $[0,\ 1]$ skirstinių šeimas. Viena iš tokių alternatyvų galėtų būti intervale $[0,\ 1]$ apibrėžtas beta skirstinys $Be(\gamma, \eta)$, priklausantis nuo dviejų parametrų $\gamma, \eta > 0$. Abu parametrai yra skirstinio formos parametrai ir jiems kintant tankis įgyja įvairius pavidalus. Tolygusis skirstinys gaunamas imant $\gamma = \eta = 1$. Tikrinama paprastoji hipotezė $H_0 : Y_i \sim U(0,\ 1)$, kai sudėtinė alternatyva yra $H : Y_i \sim g(y)$, čia tankio funkcija

$$g(y) \in \mathcal{G} = \{\frac{1}{B(\gamma,\eta)} y^{\gamma-1}(1-y)^{\eta-1}, \quad 0 < y < 1, \quad \gamma, \eta > 0\}. \tag{3.3.1}$$

Normuojanti konstanta yra beta funkcija

$$B(\gamma,\eta) = \int_0^1 y^{\gamma-1}(1-y)^{\eta-1} dy.$$

Prilyginę (3.1.2) beta skirstinio tankiui ir atlikę atvirkštinį keitimą $x = F_0^{-1}(y)$, gauname pradinio uždavinio alternatyvų aibę $\mathcal{F}$. Jai priklauso tokio pavidalo tankiai

$$f(x) = f_0(x) \frac{1}{B(\gamma,\eta)} (F_0(x))^{\gamma-1}(1-F_0(x))^{\eta-1}, \quad \gamma, \eta > 0.$$

Iš šios išraiškos gerai matyti, kaip deformuojamas hipotetinis tankis $f_0(x)$ formuluojant alternatyvas. Pavyzdžiui, jeigu tankis $f_0(x)$ simetrinis ir domina simetrinės alternatyvos, tai reikėtų imti $\gamma = \eta = \beta$. Kai $\beta > 1$, tai tankis $f(x)$ greičiau, o kai $\beta < 1$ – lėčiau artėja į nulį, kai $|x| \to \infty$, negu hipotetinis tankis $f_0(x)$.

**Kriterijaus statistika**

Hipotezė $H_0$, kai sudėtinė alternatyva yra (3.3.1), tampa parametrine hipoteze

$$H_0 : \gamma = \eta = 1, \tag{3.3.2}$$

kai alternatyvioje tvirtinama, kad nors vienas iš šių parametrų nelygus vienetui.

Tikėtinumo funkcija

$$L = L(\gamma,\eta) = \frac{1}{B^n(\gamma,\eta)} \left( \prod_{i=1}^n Y_i \right)^{\gamma-1} \left( \prod_{i=1}^n (1-Y_i) \right)^{\eta-1},$$

o logtikėtinumo funkcija

$$\ell = \ell(\gamma,\eta) = (\gamma-1)\sum_{i=1}^n \ln Y_i + (\eta-1)\sum_{i=1}^n \ln(1-Y_i) - n\ln B(\gamma,\eta).$$

Parametrų $\gamma$ ir $\eta$ DT įvertiniams rasti turime lygčių sistemą

$$\dot{\ell}_\gamma = \sum_{i=1}^n \ln Y_i - n(\ln B(\gamma,\eta))'_\gamma = 0,$$

$$\dot{\ell}_\eta = \sum_{i=1}^n \ln(1-Y_i) - n(\ln B(\gamma,\eta))'_\eta = 0. \tag{3.3.3}$$

Fišerio informacinė matrica

$$\boldsymbol{I} = \boldsymbol{I}(\gamma, \eta) = [I_{rs}(\gamma, \eta)]_{2\times 2}, \quad I_{11}(\gamma, \eta) = n[\ln B(\gamma, \eta)]''_{\gamma\gamma},$$

$$I_{22}(\gamma, \eta) = n[\ln B(\gamma, \eta)]''_{\eta\eta}, \quad I_{12} = I_{21} = n[\ln B(\gamma, \eta)]''_{\gamma\eta}. \tag{3.3.4}$$

Kriterijaus statistika imkime informantinę statistiką

$$R_I = (\dot\ell_\gamma(1,\ 1), \dot\ell_\eta(1,\ 1))(-\ddot\ell(1,\ 1))^{-1}(\dot\ell_\gamma(1,\ 1), \dot\ell_\eta(1,\ 1))^T.$$

Kai hipotezė $H_0 : \gamma = \eta = 1$ yra teisinga, tai $B(1,\ 1) = 1$,

$$\dot B_\gamma(1,\ 1) = \int_0^1 \ln x dx = -1, \quad \dot B_\eta(1,\ 1) = \int_0^1 \ln(1-x)dx = -1,$$

$$\ddot B_{\gamma,\gamma}(1,\ 1) = \int_0^1 \ln^2 x dx = 2, \quad \ddot B_{\eta,\eta}(1,\ 1) = \int_0^1 \ln^2(1-x)dx = 2,$$

$$\ddot B_{\gamma,\eta}(1,\ 1) = \int_0^1 \ln x \ln(1-x)dx = 2 - \pi^2/6.$$

Gauname informacinės matricos elementus $I_{11} = I_{22} = n$, $I_{12} = I_{21} = n(1 - \pi^2/6)$ ir informantinę statistiką

$$R_I = \frac{36}{\pi^2(12-\pi^2)}(T_1^2 + \frac{\pi^2-6}{3}T_1T_2 + T_2^2), \tag{3.3.5}$$

čia

$$T_1 = \frac{1}{\sqrt n}\sum_{i=1}^n(\ln Y_i + 1), \quad T_2 = \frac{1}{\sqrt n}\sum_{i=1}^n(\ln(1-Y_i) + 1). \tag{3.3.6}$$

Jeigu hipotezė $H_0$ teisinga, tai (žr. A priedą, (7.1.9))

$$R_I \xrightarrow{d} \chi^2_2.$$

**Suderinamumo kriterijus**

Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$R_I > \chi^2_\alpha(2), \tag{3.3.7}$$

arba $P$ reikšmių terminais, kai

$$pv_a = \mathbf{P}\{\chi^2_2 > r_I\} < \alpha,$$

čia $r_I$ yra statistikos $R_I$ stebinys.

**3.3.1 pastaba.** Jeigu apie tikrinamą hipotezę ir alternatyvas turima papildomos informacijos, tai kriterijų galima patikslinti. Pavyzdžiui, tegu žinoma, kad skirstinys $F_0$ ir aibės $\mathcal{F}$ skirstiniai yra simetriški. Tada vietoje alternatyvų aibės (3.3.1) natūralu imti simetriškus beta skirstinius

$$\mathcal{G} = \{\frac{1}{B(\gamma,\gamma)} y^{\gamma-1}(1-y)^{\gamma-1}, \quad \gamma \neq 1, \quad \gamma > 0\}.$$

Tikėtinumo funkcija

$$L = L(\gamma) = \exp\{(\gamma-1)\sum_{i=1}^{n} \ln(Y_i(1-Y_i)) - n\ln B(\gamma,\gamma)\}$$

turi monotoninį tikėtinumo santykį pakankamosios statistikos $\sum_{i=1}^n \ln(Y_i(1-Y_i))$ atžvilgiu. Tikrinant hipotezę $H_0 : \gamma = 1$ su vienpusėmis alternatyvomis $H_1 : \gamma > 1$ arba $H_2 : \gamma < 1$ egzistuoja TG kriterijai (žr. I dalies 4.3.1 skyrelį). Hipotezė atmetama, kai

$$T > t_\alpha \quad \text{arba} \quad T < t_{1-\alpha}, \tag{3.3.8}$$

čia

$$T = \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}[\ln(Y_i(1-Y_i)) + 2], \quad \sigma^2 = \boldsymbol{V}(\ln(Y_i(1-Y_i))) = 4 - \frac{\pi^2}{3},$$

o $t_\alpha$ statistikos $T$ lygmens $\alpha$ kritinė reikšmė.

Jeigu alternatyva dvipusė $H_3 : \gamma \neq 1$, tai egzistuoja TGN kriterijus (žr. I dalies 4.3.2 skyrelį), kurio pavidalas (keičiant simetrišku kriterijumi) yra toks: hipotezė atmetama, kai

$$T < t_{1-\alpha/2} \quad \text{arba} \quad T > t_{\alpha/2}. \tag{3.3.9}$$

Asimptotiškai ($n \to \infty$) statistikos $T$ skirstinys konverguoja į standartinį normalųjį skirstinį. Taigi asimptotinis kriterijus gaunamas (3.3.8) ir (3.3.9) keičiant $t_\alpha$ ir $t_{1-\alpha}$ į $z_\alpha$ ir $-z_\alpha$.

Analogiški rezultatai teisingi, kai vienas iš beta skirstinio parametrų nežinomas, o kitas lygus 1.

**3.3.1 pavyzdys** (**2.1.1** pavyzdžio tęsinys). Patikrinsime prielaidą, kad 2.1.1 pratimo duomenys gauti stebint a. d $Y \sim U(0,\ 1)$. Statistikos (3.3.4) realizacija yra $R_I = 0,3056$; asimptotinė $P$ reikšmė $pv_a = \mathbf{P}\{\chi_2^2 > 0,3056\} = 0,8583$. Hipotezė neatmetama.

## 3.4. Modifikuotieji kriterijai

Retai tenka tikrinti paprastąsias suderinamumo hipotezes $H_0 : X \sim F_0(x)$. Dažniau reikia tikrinti sudėtines suderinamumo hipotezes $H_0 : X \sim F(x) \in \mathcal{F}_0$, kai $\mathcal{F}_0 = \{F(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbf{R}^m\}$, o $F(x|\boldsymbol{\theta})$ žinomos analizinės išraiškos pasiskirstymo funkcija, priklausanti nuo baigtinės dimensijos parametro $\boldsymbol{\theta}$. Pavyzdžiui, $\mathcal{F}_0$ gali būti normaliųjų, gama, Veibulo ir kt. skirstinių šeima. Paprastųjų hipotezių atvejis gali būti svarbus teoriniu požiūriu. Jis gali nurodyti kriterijų statistikų paieškos kryptis tikrinant sudėtines hipotezes.

Apsiribosime skirstinių šeimomis, priklausančiomis tik nuo poslinkio ir mastelio parametrų.

Tarkime, kad paprastoji imtis $\mathbf{X} = (X_1, ..., X_n)^T$ gauta stebint a. d. $X$. Tikriname sudėtinę suderinamumo hipotezę

$$H_0 : X \sim F(x) \in \mathcal{F}_0, \quad \mathcal{F}_0 = \{F_0(\frac{x-\mu}{\sigma}), \mu \in \mathbf{R}, \sigma > 0\}, \tag{3.4.1}$$

o $F_0(x)$ yra absoliučiai tolydi žinomos analizinės formos pasiskirstymo funkcija, kurios tankis $f_0(x)$. Šeimų $\mathcal{F}_0$ pavyzdžiai gali būti normaliųjų, Koši, logistinių, ekstremaliųjų reikšmių ir kt. skirstinių šeimos.

Jeigu tikrosios parametrų reikšmės yra $\mu$ ir $\sigma$, tai $\varepsilon_i = (X_i - \mu)/\sigma \sim F_0(x)$ ir hipotezė $H_0$ tampa paprastąja. A. d. $Y_i = F_0(\varepsilon_i)$ yra nepriklausomi vienodai tolygiai pasiskirstę intervale $[0,\ 1]$, t. y. $Y_i \sim U(0,\ 1)$, $i = 1, ..., n$.

**Kriterijaus sudarymo idėja**

Skyreliuose 3.2, 3.3 paprastosios hipotezės tikrinimo kriterijai buvo sudaromi tokiu būdu. Parenkama tam tikra a. d. $Y_i = F_0(\varepsilon_i)$ transformacija $G(\varepsilon_i) = L(F_0(\varepsilon_i))$ ir asimptotinis kriterijus sudaromas remiantis statistikos

$$T = \frac{1}{\sqrt{n\boldsymbol{V}(G(\varepsilon_i))}} \sum_{i=1}^{n} (G(\varepsilon_i) - \mathbf{E}(G(\varepsilon_i))) \xrightarrow{d} Z \sim N(0,\ 1) \tag{3.4.2}$$

asimptotiniu normalumu. Bendriau, parenkama keletas transformacijų $G_j(\varepsilon_i) = L_j(F_0(\varepsilon_i)), j = 1, ..., k$, ir kriterijus grindžiamas tuo, kad kvadratinė forma

$$(T_1, ..., T_k)\mathbf{\Sigma}_0^{-1}(T_1, ..., T_k)^T \xrightarrow{d} \chi_k^2 \tag{3.4.3}$$

konverguoja į a. d., turintį $\chi^2$ skirstinį su $k$ laisvės laipsnių; čia $T_j$ yra (3.4.2) statistika, atitinkanti transformaciją $G_j, j = 1, ..., k$, o $\mathbf{\Sigma}_0$ yra a. v. $(T_1, ..., T_k)^T$ kovariacinė matrica.

Kai parametrai $\mu$ ir $\sigma$ nežinomi, kriterijaus statistiką sudarysime analogiškai (3.4.2), pakeisdami nežinomus parametrus jų DT įvertiniais $\hat{\mu}$ ir $\hat{\sigma}$. Gauname statistiką

$$\hat{T} = \frac{1}{\sqrt{n\boldsymbol{V}(G(\hat{\varepsilon}_i))}} \sum_{i=1}^{n} (G(\hat{\varepsilon}_i) - \mathbf{E}(G(\varepsilon_i))), \tag{3.4.4}$$

čia $\hat{\varepsilon}_i = (X_i - \hat{\mu})/\hat{\sigma}$. Atsitiktiniai dydžiai $Y_i = F_0(\hat{\varepsilon}_i), i = 1, ..., n$ yra vienodai pasiskirstę intervale $[0,\ 1]$. Tačiau jie yra priklausomi ir jų skirstiniai nėra tolygieji.

Kad kriterijus grindžiamas statistika $\hat{T}$ būtų pritaikomas, reikia įsitikinti, kad jo skirstinys (bent jau asimptotiškai) nepriklauso nuo nežinomų parametrų $\mu$ ir $\sigma$ ir nuo pasiskirstymo funkcijos $F_0$.

**3.4.1 teorema.** *Tarkime, kad tankis $f_0(x) = F_0'(x)$ tenkina įprastines reguliarumo sąlygas (žr. A priedą, sąlygas A) ir parametrų $\mu, \sigma$ DT įvertiniai yra $\hat\mu, \hat\sigma$. Tada statistikų, kurios yra a. d. $Y_i = F_0(\hat\varepsilon_i), i = 1, ..., n$ funkcijos, skirstiniai nepriklauso nuo nežinomų parametrų $\mu$ ir $\sigma$.*

**Įrodymas.** Tikėtinumo ir logtikėtinumo funkcijos yra

$$L = L(\mu, \sigma) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0\left(\frac{X_i - \mu}{\sigma}\right),$$

$$\ell = \ell(\mu, \sigma) = -n \ln \sigma + \sum_{i=1}^{n} \ln f_0\left(\frac{X_i - \mu}{\sigma}\right).$$

Informantės

$$\dot\ell_\mu = -\frac{1}{\sigma} \sum_{i=1}^{n} (\ln f_0)'(\varepsilon_i), \quad \dot\ell_\sigma = -\frac{n}{\sigma} - \frac{1}{\sigma} \sum_{i=1}^{n} \varepsilon_i (\ln f_0)'(\varepsilon_i).$$

Taigi parametrų $\mu$ ir $\sigma$ įvertiniai $\hat\mu$ ir $\hat\sigma$ tenkina lygčių sistemą

$$\sum_{i=1}^{n} (\ln f_0)'(\hat\varepsilon_i) = 0, \quad n + \sum_{i=1}^{n} \hat\varepsilon_i (\ln f_0)'(\hat\varepsilon_i) = 0. \tag{3.4.5}$$

Įvertinių vektorius $(\hat\mu, \hat\sigma)^T$ asimptotiškai turi normalųjį skirstinį

$$\sqrt{n}((\hat\mu, \hat\sigma)^T - (\mu, \sigma)^T) \xrightarrow{d} \boldsymbol{Z} \sim N_2(\boldsymbol{0}, \boldsymbol{i}^{-1}(\mu, \sigma)), \tag{3.4.6}$$

čia $\mu, \sigma$ tikrosios parametrų reikšmės, o $\boldsymbol{i}(\mu, \sigma)$ vieno imties elemento Fišerio informacijos matrica

$$\boldsymbol{i}(\mu, \sigma) = \boldsymbol{I}(\mu, \sigma)/n, \quad \boldsymbol{I}(\mu, \sigma) = [\boldsymbol{I}_{rs}]_{2\times 2}, \tag{3.4.7}$$

$$\boldsymbol{I}_{11}(\mu, \sigma) = -\mathbf{E}\ddot\ell_{\mu\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{E}(\ln f_0)''(\varepsilon_i),$$

$$\boldsymbol{I}_{12}(\mu, \sigma) = -\mathbf{E}\ddot\ell_{\mu\sigma} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{E}((\varepsilon_i \ln f_0)''(\varepsilon_i) + (\ln f_0)'(\varepsilon_i)),$$

$$\boldsymbol{I}_{22}(\mu, \sigma) = -\mathbf{E}\ddot\ell_{\sigma\sigma} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathbf{E}((\varepsilon_i^2 \ln f_0)''(\varepsilon_i) + 2\varepsilon_i (\ln f_0)'(\varepsilon_i) + 1).$$

Atsitiktinių dydžių $\varepsilon_i = (X_i - \mu)/\sigma$ pasiskirstymo funkcija yra $F_0(x)$ ir nepriklauso nuo nežinomų parametrų. Atlikę keitimą $X_i = \sigma\varepsilon_i + \mu$, lygčių sistemą (3.4.5) užrašome tokiu pavidalu:

$$\sum_{i=1}^{n} (\ln f_0)'(\frac{\sigma}{\hat\sigma}\varepsilon_i + \frac{\mu - \hat\mu}{\hat\sigma}) = 0, \quad n + \sum_{i=1}^{n} (\frac{\sigma}{\hat\sigma}\varepsilon_i + \frac{\mu - \hat\mu}{\hat\sigma})(\ln f_0)'(\frac{\sigma}{\hat\sigma}\varepsilon_i + \frac{\mu - \hat\mu}{\hat\sigma}) = 0.$$

Atsitiktinių dydžių $\sigma/\hat{\sigma}$ ir $(\hat{\mu}-\mu)/\hat{\sigma}$ skirstiniai nuo nežinomų parametrų nepriklauso. Remdamiesi sąryšiu

$$Y_i = F_0\left(\frac{X_i-\hat{\mu}}{\hat{\sigma}}\right) = F_0\left(\frac{\sigma}{\hat{\sigma}}\varepsilon_i + \frac{\mu-\hat{\mu}}{\sigma}\right)$$

darome išvadą, kad $Y_i$ skirstinys nuo nežinomų parametrų nepriklauso. Tada ir a. d. $Y_1, ..., Y_n$ funkcija $\hat{T}$ nuo nežinomų parametrų nepriklauso. ▲

Gavome, kad statistikos $\hat{T}$ skirstinys nepriklauso nuo nežinomų parametrų ne tik asimptotiškai, bet ir su bet kokiu baigtiniu imties didumu $n$. Lieka ištirti statistikos, kuri yra a. d. $Y_1, ..., Y_n$ funkcija, asimptotines savybes.

**Statistikos asimptotinis skirstinys**

Nagrinėsime statistiką

$$\hat{T} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(G(\hat{\varepsilon}_i) - \mathbf{E}(G(\varepsilon_i))), \tag{3.4.8}$$

čia

$$G(\hat{\varepsilon}_i) = L(Y_i) = L(F_0(\hat{\varepsilon}_i)), \quad i = 1, ..., n.$$

**3.4.2 teorema.** *Tarkime, kad hipotezė (3.4.1) teisinga ir įvykdytos sąlygos.*

*1) Funkcija $L$ du kartus diferencijuojama ir egzistuoja dispersija $\boldsymbol{V}(G(\varepsilon_i)) = \sigma_0^2 < \infty$.*

*2) Matrica $\boldsymbol{i} = \boldsymbol{i}(\mu,\sigma)$ neišsigimusi; atvirkštinė matrica $\boldsymbol{i}^{-1} = [i^{rs}]_{2\times 2}$.*

*3) Įvertinys $(\hat{\mu},\hat{\sigma})^T \xrightarrow{P} (\mu,\sigma)^T$ ir (žr. A priedas, (7.1.2))*

$$(\sqrt{n}(\hat{\mu}-\mu), \sqrt{n}(\hat{\sigma}-\sigma))^T = \boldsymbol{i}^{-1}(\mu,\sigma)\left(\frac{1}{\sqrt{n}}\dot{\ell}_\mu, \frac{1}{\sqrt{n}}\dot{\ell}_\sigma\right)^T + o_P(1).$$

*4) $|\Delta_i| < \infty$, $i$ = 1, 2, 3, kai*

$$\Delta_i = \int_{-\infty}^{\infty} x^{i-1}g'(x)dF_0(x), \quad g(x) = G'(x) = L'(F_0(x))f_0(x).$$

*5) $|A_i| < \infty$, $i$ = 1, 2, kai*

$$A_1 = \int_{-\infty}^{\infty} g(x)dF_0(x), \quad A_2 = \int_{-\infty}^{\infty} xg(x)dF_0(x).$$

*6) Egzistuoja konstanta $\delta > 0$, kad $|A_i| < \infty$, $i$ = 3, 4, kai*

$$A_3 = \int_{-\infty}^{\infty} |g(x)|^{2+\delta}dF_0(x), \quad A_4 = \int_{-\infty}^{\infty} |1+(\ln f_0)'(x)+x(\ln f_0)'(x)|^{2+\delta}dF_0(x).$$

*Tada*

$$\hat{T} \xrightarrow{d} Z \sim N(0,\sigma_B^2), \tag{3.4.9}$$

$$\sigma_B^2 = \sigma_0^2 - [A_1^2 j^{11} + 2A_1A_2 j^{12} + A_2^2 j^{22}], \quad j^{rs} = i^{rs}/\sigma^2, \quad r,s = 1,\ 2.$$

**Įrodymas.** Užrašykime statistiką $\hat{T}$ dviejų dėmenų suma $\hat{T} = B_1 + B_2$,

$$B_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [G(\hat{\varepsilon}_i) - G(\varepsilon_i)], \quad B_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [G(\varepsilon_i) - \mathbf{E}(G(\varepsilon_i))]. \tag{3.4.10}$$

Dėmuo $B_2$ yra suma centruotų vienodai pasiskirsčiusių n. a. d. $G(\varepsilon_i)$, t. y. statistika, kuri gaunama tikrinant paprastąją suderinamumo hipotezę. Pirmasis dėmuo $B_1$ apibūdina paklaidas, kurių atsiranda keičiant parametrus $\mu$ ir $\sigma$ jų DT įvertiniais.

Skleisdami $B_1$ Teiloro eilute taško $(\mu, \sigma)$ aplinkoje, gauname

$$B_1 = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} [g(\varepsilon_i)(\hat{\mu} - \mu) + \varepsilon_i g(\varepsilon_i)(\hat{\sigma} - \sigma)] + R. \tag{3.4.11}$$

Įvertinsime liekaną $R$ imdami tolesnį skleidinio narį

$$R = \frac{\sqrt{n}}{\sigma^2} [\Delta_{1n}(\hat{\mu} - \mu)^2 + 2\Delta_{2n}(\hat{\mu} - \mu)(\hat{\sigma} - \sigma) + \Delta_{3n}(\hat{\sigma} - \sigma)^2] + o_P(1),$$

čia

$$\Delta_{1n} = \frac{1}{n} \sum_{i=1}^{n} g'(\varepsilon_i) \xrightarrow{P} \Delta_2,$$

$$\Delta_{2n} = \frac{1}{n} \sum_{i=1}^{n} (g(\varepsilon_i) + \varepsilon_i g'(\varepsilon_i)) \xrightarrow{P} A_1 + \Delta_2,$$

$$\Delta_{3n} = \frac{1}{n} \sum_{i=1}^{n} (2\varepsilon_i g(\varepsilon_i) + \varepsilon_i^2 g'(\varepsilon_i)) \xrightarrow{P} 2A_2 + \Delta_3,$$

pagal tikimybę artėja į aprėžtas konstantas. Likusieji daugikliai, pavyzdžiui,

$$\sqrt{n}(\hat{\mu} - \mu)^2 = (\sqrt{n}(\hat{\mu} - \mu))^2 \frac{1}{\sqrt{n}} = O_P(\frac{1}{\sqrt{n}}) = o_P(1).$$

Taigi liekana $R = o_P(1)$, ir statistikos $B_1$ asimptotinis skirstinys sutampa su (3.4.11) pirmojo dėmens skirstiniu. Pertvarkome jį taip:

$$B_1 = -\frac{\sqrt{n}}{\sigma}(\hat{\mu} - \mu)\frac{1}{n} \sum_{i=1}^{n} g(\varepsilon_i) - \frac{\sqrt{n}}{\sigma}(\hat{\sigma} - \sigma)\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(\varepsilon_i) + o_P(1).$$

Remdamiesi 3) sąlyga ir didžiųjų skaičių dėsniu, gauname

$$\frac{1}{n} \sum_{i=1}^{n} g(\varepsilon_i) \xrightarrow{P} A_1, \quad \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i g(\varepsilon_i) \xrightarrow{P} A_2$$

ir

$$B_1 = -\frac{1}{\sigma} A_1 \sqrt{n}(\hat{\mu} - \mu) - \frac{1}{\sigma} A_2 \sqrt{n}(\hat{\sigma} - \sigma) + o_P(1). \tag{3.4.12}$$

Pasinaudoję 4) sąlyga, gauname

$$B_1 = -\left(\frac{A_1}{\sigma}, \frac{A_2}{\sigma}\right)(\sqrt{n}(\hat{\mu}-\mu), \sqrt{n}(\hat{\sigma}-\sigma))^T + o_P(1)$$

$$= -\left(\frac{A_1}{\sigma}, \frac{A_2}{\sigma}\right)\boldsymbol{i}^{-1}(\mu,\sigma)(\frac{1}{\sqrt{n}}\dot{\ell}_\mu, \frac{1}{\sqrt{n}}\dot{\ell}_\sigma)^T + o_P(1)$$

$$= C_1\frac{1}{\sqrt{n}}\dot{\ell}_\mu + C_2\frac{1}{\sqrt{n}}\dot{\ell}_\sigma + o_P(1),$$

čia

$$C_1 = -\frac{A_1}{\sigma}i^{11} - \frac{A_2}{\sigma}i^{12}, \quad C_2 = -\frac{A_1}{\sigma}i^{12} - \frac{A_2}{\sigma}i^{22}.$$

Statistikos $B_1$ skirstinys asimptotiškai sutampa su skirstiniu a. d.

$$B_1^0 = C_1\frac{1}{\sqrt{n}}\dot{\ell}_\mu + C_2\frac{1}{\sqrt{n}}\dot{\ell}_\sigma = \frac{1}{\sqrt{n}}\sum_{j=1}^n \xi_j, \tag{3.4.13}$$

$$\xi_j = C_1\dot{\ell}_{j\mu} + C_2\dot{\ell}_{j\sigma}$$

čia $\dot{\ell}_{j\mu}$ ir $\dot{\ell}_{j\sigma}$ yra $\dot{\ell}_\mu$ ir $\dot{\ell}_\sigma$ $j$-osios komponentės.

Statistika

$$B_2 = \frac{1}{\sqrt{n}}\sum_{j=1}^n [G(\varepsilon_j) - \mathbf{E}(G(\varepsilon_j))] = \frac{1}{\sqrt{n}}\sum_{j=1}^n \eta_j. \tag{3.4.14}$$

Atsitiktinio vektoriaus $(B_1, B_2)^T$ asimptotinis skirstinys sutampa su asimptotiniu vektoriaus $(B_1^0, B_2)$ skirstiniu. Rasime atsitiktinio vektoriaus $(B_1^0, B_2)^T$ kovariacinę matricą $\boldsymbol{\Sigma} = [\sigma_{kl}]_{2\times 2}$. Turime $\boldsymbol{V}B_2 = \sigma_0^2$,

$$\boldsymbol{V}B_1^0 = \boldsymbol{V}(C_1\frac{1}{\sqrt{n}}\dot{\ell}_\mu + C_2\frac{1}{\sqrt{n}}\dot{\ell}_\sigma) = C_1^2 i_{11} + 2C_1C_2 i_{12} + C_2^2 i_{22}$$

$$= \frac{A_1^2}{\sigma^2}i^{11} + 2\frac{A_1A_2}{\sigma^2}i^{12} + \frac{A_2^2}{\sigma^2}i^{22} = \sigma_{11}. \tag{3.4.15}$$

Remdamiesi (3.4.5) ir (3.4.14), gauname

$$\mathbf{Cov}\,(\frac{1}{\sqrt{n}}\dot{\ell}_\mu, B_2) = \mathbf{E}(-\frac{1}{\sigma}(\ln f_0)'(\varepsilon_i)G(\varepsilon_i)) = -\frac{1}{\sigma}\int_{-\infty}^{\infty} G(x)f_0'(x)dx = \frac{A_1}{\sigma},$$

$$\mathbf{Cov}\,(\frac{1}{\sqrt{n}}\dot{\ell}_\sigma, B_2) = \mathbf{E}(-\frac{1}{\sigma}[1 + \varepsilon_i(\ln f_0)'(\varepsilon_i)(G(\varepsilon_i) - \mathbf{E}(G(\varepsilon_i)))]$$

$$= -\frac{1}{\sigma}\int_{-\infty}^{\infty} x(G(x) - \mathbf{E}(G(\varepsilon_i)))f_0(x)dx = \frac{A_2}{\sigma}.$$

Tada

$$\sigma_{12} = \mathbf{Cov}\,(C_1\frac{1}{\sqrt{n}}\dot{\ell}_\mu + C_2\frac{1}{\sqrt{n}}\dot{\ell}_\sigma, B_2) = \frac{1}{\sigma}(C_1A_1 + C_2A_2)$$

$$= -(\frac{A_1^2}{\sigma^2} i^{11} + 2\frac{A_1 A_2}{\sigma^2} i^{12} + \frac{A_2^2}{\sigma^2} i^{22}) = -\sigma_{11}. \quad (3.4.16)$$

Taigi vektoriaus $(B_1^0, B_2)^T$ kovariacinė matrica turi tokį pavidalą

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & -\sigma_{11} \\ -\sigma_{11} & \sigma_0^2 \end{pmatrix}.$$

Tam kad vektorius $(B_1^0, B_2)^T$ asimptotiškai turėtų dvimatį normalųjį skirstinį, pakanka, jog būtų įvykdyta Liapunovo sąlyga: egzistuoja toks $\delta > 0$, kad $n \to \infty$

$$\frac{\sum_{j=1}^n \mathbf{E}|\eta_j|^{2+\delta}}{(\sum_{j=1}^n \boldsymbol{V}\eta_j)^{1+\delta/2}} \to 0, \quad \frac{\sum_{j=1}^n \mathbf{E}|\xi_j|^{2+\delta}}{(\sum_{j=1}^n \boldsymbol{V}\xi_j)^{1+\delta/2}} \to 0.$$

Remiantis 6) teoremos sąlyga, a. d. $\eta_1, ..., \eta_n$ tenkina šią sąlygą, nes tai vienodai pasiskirstę nepriklausomi a. d. su baigtine dispersija, o $\mathbf{E}|\eta_j|^{2+\delta}$ aprėžtas.

Kadangi

$$\frac{1}{n}\sum_{j=1}^n \boldsymbol{V}\xi_j = \boldsymbol{V}B_1^0 = \sigma_{11} > 0,$$

pakanka įrodyti, kad $\mathbf{E}|\xi_j|^{2+\delta} \le C < \infty$ su visais $j = 1, ..., n$. Tai ekvivalentu nelygybei

$$\mathbf{E}|(A_1 i^{11} + A_2 i^{12})(\ln f_0)'(\varepsilon_j) + (A_1 i^{12} + A_2 i^{22})(1 + \varepsilon_j(\ln f_0)'(\varepsilon_j))|^{2+\delta} \le C.$$

Tarkime $(A_1 i^{11} + A_2 i^{12})$ ir $(A_1 i^{12} + A_2 i^{22})$ neviršija konstantos $K$. Tada, remiantis 6) teoremos sąlyga, šis reiškinys neviršija $C = KA_4$.

Taigi a. v. $(B_1, B_2)^T$ asimptotiškai dvimatis normalusis $N_2(\mathbf{0}, \mathbf{\Sigma})$. Tada a. d.

$$\hat{T} = B_1 + B_2 \xrightarrow{d} Z \sim N(0, \sigma_B^2), \quad (3.4.17)$$

$$\sigma_B^2 = \boldsymbol{V}B_1 + 2\mathbf{Cov}\,(B_1, B_2) + \boldsymbol{V}B_2$$
$$= \sigma_{11} - 2\sigma_{11} + \sigma_0^2 = \sigma_0^2 - \sigma_{11}.$$

▲

**Modifikuotasis suderinamumo kriterijus**

Sudėtinė suderinamumo hipotezė (3.4.2) atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$\frac{|\hat{T}|}{\sigma_B} > z_{\alpha/2}, \quad \text{arba} \quad \frac{\hat{T}^2}{\sigma_B^2} > \chi_\alpha^2(1), \quad (3.4.18)$$

arba P reikšmių terminais, kai

$$pv_a = \mathbf{P}\{\chi_1^2 > t\} < \alpha,$$

čia $t$ yra statistikos $\hat{T}^2/\sigma_B^2$ realizacija.

**3.4.1 pastaba.** Dispersijos sumažėjimą imant modifikuotojo kriterijaus statistiką galima paaiškinti taip. Vertinant parametrus $\mu$ ir $\sigma$ modelis prisitaiko prie turimos imties, todėl $\hat{T}$ sklaida apie nulį yra mažesnė už $\sigma_0^2 = \boldsymbol{V}(G(\varepsilon_i))$, kai tikrinama paprastoji hipotezė.

**3.4.2 pastaba.** Jeigu imtis nėra didelė ir kyla abejonių dėl aproksimacijos (3.4.17) tikslumo, kriterijų galima patikslinti atliekant kompiuterinį modeliavimą. Tarkime, kad spręsdami konkretų suderinamumo uždavinį gavome statistikos $\hat{T}^2/\sigma_B^2$ realizaciją $t$. Modeliuojama $N$ a. d. $X \sim F_0(\varepsilon)$ paprastųjų didumo $n$ imčių (kadangi statistikos skirstinys nepriklauso nuo nežinomų parametrų, tai modeliuojant galima imti, pavyzdžiui, $\mu = 0, \sigma = 1$). Randame kiekvienos sumodeliuotos imties statistikos $\hat{T}^2/\sigma_B^2$ realizaciją $t^*$ ir kiekvieną kartą patikriname nelygybę $t^* > t$. Tarkime, ši nelygybė teisinga $M$ kartų. Tada P reikšmės įverčiu imamas dažnis $\hat{pv} = M/N$. Hipotezė atmetama, kai $\hat{pv} < \alpha$. Kad kiekvieną kartą nereikėtų modeliuoti, galima tam tikram imties didumų $n$ rinkiniui modeliuojant įvertinti kritines reikšmes ir jų lenteles įdėti į kompiuterio atmintį.

### Modifikuotojo kriterijaus, grindžiamo keletu transformacijų, statistika

Tikrinant paprastąją suderinamumo hipotezę $H_0 : X \sim F_0((x-\mu)/\sigma)$, kai $\mu$ ir $\sigma$ žinomi, buvo naudojami kriterijai, kurių statistikos gaunamos imant keletą transformacijų $G_j(\varepsilon_i) = L_j(F_0(\varepsilon_i))$, $\varepsilon_i = (X_i - \mu)/\sigma$, $j = 1, ..., k$. Kriterijaus statistika yra kvadratinė forma (3.3.3).

Kai nežinomi parametrai $\mu$ ir $\sigma$ keičiami jų DT įvertiniais $\hat{\mu}$ ir $\hat{\sigma}$, pagal analogiją nagrinėsime transformacijų rinkinį

$$G_j(\hat{\varepsilon}_i), \quad \hat{\varepsilon}_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}} \quad j = 1, ..., k. \tag{3.4.19}$$

Pažymėkime $\boldsymbol{\Sigma}_0$ atsitiktinio vektoriaus $(G_1(\varepsilon_i), ..., G_k(\varepsilon_i))^T$ kovariacinę matricą. Tegu $A_{j1}$ ir $A_{j2}$ yra 3.4.2 teoremos $A_1$ ir $A_2$ analogai, surasti imant transformaciją $G_j(\hat{\varepsilon}_i), j = 1, ..., k$.

**3.4.3 teorema.** *Tarkime, kad kiekviena transformacija $G_j$ tenkina 3.4.2 teoremos sąlygas. Tada kvadratinė forma*

$$T = (\hat{T}_1, ..., \hat{T}_k)\boldsymbol{\Sigma}^{-1}(\hat{T}_1, ..., \hat{T}_k)^T \xrightarrow{d} \chi_k^2, \tag{3.4.20}$$

*čia $\hat{T}_j$ yra statistikos (3.4.8) analogas imant transformaciją $G_j$. Kovariacinė matrica*

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 - \tilde{\boldsymbol{\Sigma}}, \tag{3.4.21}$$

*čia $\tilde{\boldsymbol{\Sigma}} = [\tilde{\sigma}_{rs}]_{k\times k}$ yra pataisų kovariacinė matrica,*

$$\tilde{\sigma}_{ss} = A_{s1}^2 j^{11} + 2A_{s1}A_{s2}j^{12} + A_{s2}^2 j^{22}, \quad s = 1, ..., k;$$

$$\tilde{\sigma}_{rs} = A_{r1}A_{s1}j^{11} + (A_{r1}A_{s2} + A_{r2}A_{s1})j^{12} + A_{r2}A_{s2}j^{22}, \quad r \neq s = 1, ..., k. \tag{3.4.22}$$

**Įrodymas.** Pakartoję 3.4.2 teoremos įrodymą kiekvienai iš transformacijų gauname, kad vietoje dvimačio a. v. $(B_1^0, B_2)^T$ tenka nagrinėti asimptotinį skirstinį $(2k)$-mačio vektoriaus

$$(B_{11}^0, B_{12}, B_{21}^0, B_{22}, ..., B_{k1}^0, B_{k2})^T, \tag{3.4.23}$$

čia $B_{j1}^0$ ir $B_{j2}$ yra $B_1^0$ ir $B_2$ analogai imant transformaciją $G_j$:

$$B_{j1}^0 = C_{j1}\frac{1}{\sqrt{n}}\dot{\ell}_\mu + C_{j2}\frac{1}{\sqrt{n}}\dot{\ell}_\sigma, \quad C_{j1} = -\frac{A_{j1}}{\sigma}i^{11} - \frac{A_{j2}}{\sigma}i^{12},$$

$$C_{j2} = -\frac{A_{j1}}{\sigma}i^{12} - \frac{A_{j2}}{\sigma}i^{22}, \quad B_{j2} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[G_j(\varepsilon_i) - \mathbf{E}(G_j(\varepsilon_i))]. \tag{3.4.24}$$

Minėto $2k$-mačio vektoriaus asimptotinis skirstinys gaunamas analogiškai teoremai 3.4.2, tereikia rasti jo kovariacinę matricą. Vektoriaus (3.4.23) elementų $B_{j1}^0$ dispersijos $\boldsymbol{V}(B_{j1}^0)$ ir kovariacijos $\mathbf{Cov}\,(B_{j1}^0, B_{j2})$, $j = 1, ..., k$, surastos 3.4.2 teoremoje. Vektoriaus $(B_{12}, ..., B_{k2})^T$ kovariacinę matricą pažymėjome $\mathbf{\Sigma}_0$. Įsitikiname, kad likusios kovariacijos:

$$\mathbf{Cov}\,(B_{r1}^0, B_{s1}^0) = \tilde{\sigma}_{rs}, \quad \mathbf{Cov}\,(B_{r1}^0, B_{s2}) = -\tilde{\sigma}_{rs}, \ \ r \neq s = 1, ..., k.$$

Tada a. v. $(B_{11}^0 + B_{12}, ..., B_{k1}^0 + B_{f2})^T$ asimptotinis skirstinys yra $N_k(\mathbf{0}, \mathbf{\Sigma})$, kai $\mathbf{\Sigma}$ apibrėžta (3.4.21) lygybe. Tokį pat asimptotinį skirstinį turi a. v. $(\hat{T}_1, ..., \hat{T}_k)^T$. ▲

**Modifikuotasis suderinamumo kriterijus, grindžiamas keletu transformacijų**

Sudėtinė suderinamumo hipotezė (3.4.2) atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$T = (\hat{T}_1, ..., \hat{T}_k)\mathbf{\Sigma}^{-1}(\hat{T}_1, ..., \hat{T}_k)^T > \chi^2_\alpha(k), \tag{3.4.25}$$

arba $P$ reikšmių terminais, kai

$$pv_a = \mathbf{P}\{\chi^2_k > t\} < \alpha,$$

čia $t$ yra statistikos $T$ realizacija.

## 3.5. Modifikuotųjų kriterijų pavyzdžiai

Remiantis 3.4.2 ir 3.4.3 teoremomis, norint pritaikyti sudėtinių hipotezių suderinamumo kriterijus, grindžiamus viena ar keliomis transformacijomis $G_j(\hat{\varepsilon}_i) = L_j(F_0(\hat{\varepsilon}_i)), j = 1, ..., k$, reikia atlikti tokius veiksmus:

1) remiantis (3.4.7) rasti Fišerio informacinės matricos atvirkštinę $\boldsymbol{i}^{-1} = [i^{rs}]_{2\times 2}$ ir $j^{rs} = i^{rs}/\sigma^2$, $r, s = 1,\ 2$;

2) apskaičiuoti kiekvienos transformacijos $G_j$ konstantas

$$A_{j1} = \int_{-\infty}^{\infty} g_j(x)dF_0(x), \quad A_{j2} = \int_{-\infty}^{\infty} xg_j(x)dF_0(x), \tag{3.5.1}$$

$$g_j(x) = G_j'(x) = L'(F_0(x))f_0(x), \quad j = 1, ..., k;$$

3) rasti kovariacinę matricą $\mathbf{\Sigma}_0$ (žr. 3.4.3 teoremą);

4) apskaičiuoti pataisų kovariacinę matricą $\tilde{\mathbf{\Sigma}}$ ir kovariacinę matricą $\mathbf{\Sigma}$ (žr. (3.4.21));

5) rasti statistikos $T$ iš (3.4.20) realizaciją $t$;

6) remiantis (3.4.25) kriterijumi, arba modeliavimo būdu (žr. 3.4.2 pastabą) įvertinus P reikšmę $pv$, priimti sprendimą apie sudėtinės hipotezės (3.4.1) teisingumą ar klaidingumą.

Pateiksime keleto dažnai naudojamų skirstinių suderinamumo kriterijus imdami pirmąsias dvi Neimano transformacijas (3.2.3) ir transformacijas (3.3.5), grindžiamas beta skirstiniu.

## 3.5.1. Normalusis skirstinys

Pagal didumo $n$ paprastąją imtį $\mathbf{X} = (X_1, ..., X_n)^T$ tikrinama sudėtinė suderinamumo hipotezė

$$H_0 : X \sim \Phi(\frac{x-\mu}{\sigma}), \quad \mu \in \mathbf{R}, \ \sigma > 0, \quad \varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{3.5.2}$$

1. *Neimano ir Bartono tipo kriterijus.* Imant $k = 2$ parametrus Neimano pasiūlytos transformacijos yra

$$G_1(z_i) = 2\sqrt{3}z_i, \quad G_2(z_i) = \sqrt{5}(6z_i^2 - 1/2), \tag{3.5.3}$$

čia $z_i = \Phi(\hat{\varepsilon}_i) - 1/2, \quad \hat{\varepsilon}_i = (X_i - \hat{\mu})/\hat{\sigma}, \quad i = 1, ..., n.$

Randame 1) $j^{11} = 1, \ \ j^{12} = 0, \ \ j^{22} = 1/2;$

$$2) \ A_{11} = 2\sqrt{3}\int_{-\infty}^{\infty} \varphi^2(x)dx = \sqrt{3/\pi}, \quad A_{12} = 2\sqrt{3}\int_{-\infty}^{\infty} x\varphi^2(x)dx = 0;$$

$$A_{21} = 0, \quad A_{22} = 12\sqrt{5}\int_{-\infty}^{\infty} x(\Phi(x) - 1/2)\varphi^2(x)dx = \sqrt{15}/\pi;$$

3) kadangi transformacijos ortogonalios ir normuotos, tai $\mathbf{\Sigma}_0 = \boldsymbol{I}$;

4) kovariacinė matrica

$$\mathbf{\Sigma} = \begin{pmatrix} 1 - A_{11}^2 & 0 \\ 0 & 1 - A_{22}^2/2 \end{pmatrix} = \begin{pmatrix} 1 - 3/\pi & 0 \\ 0 & 1 - 15/(2\pi^2) \end{pmatrix};$$

5) – 6) randame statistikos

$$T = T_1^2 + T_2^2 = \hat{T}_1^2/(1 - 3/\pi) + \hat{T}_2^2/(1 - 15/(2\pi^2)) \qquad (3.5.4)$$

realizaciją $t$. Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $t$ viršija kritinę reikšmę $\chi^2_\alpha(2)$ arba kai $pv_a = \mathbf{P}\{\chi^2_2 > t\} < \alpha$.

**3.5.1 pastaba.** Kriterijus galima sudaryti imant atskirai statistikas $T_1^2$ ir $T_2^2$. Priimant sprendimą šių statistikų realizacijas reikėtų palyginti su kritine reikšme $\chi^2_\alpha(1)$. Kriterijų rinkinį galima papildyti imant didesnį skaičių (3.2.3) tranformacijų. Kriterijų statistikas galima sudaryti imant po vieną atskiras transformacijas, komponuojant jų dvejetus, trejetus ir t. t.

**3.5.2 pastaba.** Kyla klausimas, kuriuos kriterijus reikėtų naudoti. Jeigu norima patikrinti hipotezę $H_0$, kai alternatyvos yra simetrinės, tai reikėtų imti lygines transformacijas $G_2, G_4$,..., nes tada aibės (3.2.1) alternatyvos yra simetrinės. Atvirkščiai, jei norima tikrinti su nesimetriškomis alternatyvomis, tai reikėtų imti nelygines transformacijas $G_1, G_3, \ldots$ Apskritai transformacijų skaičiaus didinimas turėtų sumažinti kriterijaus galią, nes tokiu atveju išplečiama alternatyvų (3.2.1) aibė. Mažinant transformacijų skaičių, atrodo, kad reikėtų imti statistiką $T_2^2$, kai alternatyvos simetrinės, statistiką $T_1^2$, kai alternatyvos nesimetrinės, ir statistiką $T$, jeigu apie alternatyvas nieko nežinoma. Tokios rekomendacijos, gautos modeliuojant ir lyginant įvairių kriterijų galias, siūlomos [16].

2. *Kriterijai, grindžiami beta skirstiniu.*

Remdamiesi 3.3 skyreliu, kriterijus sudarysime naudodami transformacijas

$$G_1(\hat{\varepsilon}_i) = \ln \Phi(\hat{\varepsilon}_i) + 1, \quad G_2(\hat{\varepsilon}_i) = \ln(1 - \Phi(\hat{\varepsilon}_i)) + 1.$$

2) randame

$$A_{11} = \int_{-\infty}^{\infty} \frac{\varphi^2(x)}{\Phi(x)} dx \approx 0,903197, \quad A_{12} = \int_{-\infty}^{\infty} \frac{x\varphi^2(x)}{\Phi(x)} dx \approx -0,595636,$$

$$A_{21} = \int_{-\infty}^{\infty} \frac{-\varphi^2(x)}{1 - \Phi(x)} dx = -A_{11}, \quad A_{22} = \int_{-\infty}^{\infty} \frac{-x\varphi^2(x)}{1 - \Phi(x)} dx = A_{12};$$

3) – 4)

$$\mathbf{\Sigma}_0 = \begin{pmatrix} 1 & 1 - \pi^2/6 \\ 1 - \pi^2/6 & 1 \end{pmatrix}, \ \mathbf{\Sigma} = \begin{pmatrix} 1 - A_{11}^2 - A_{12}^2/2 & 1 - \pi^2/6 + A_{11}^2 - A_{12}^2/2 \\ 1 - \pi^2/6 + A_{11}^2 - A_{12}^2/2 & 1 - A_{11}^2 - A_{12}^2/2 \end{pmatrix}$$

$$\approx \begin{pmatrix} 0,006844 & -0,006560 \\ -0,006560 & 0,006844 \end{pmatrix}, \quad \rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} \approx -0,958514.$$

5) – 6) Gauname kvadratinę formą

$$\tilde{T} = (\tilde{T}_1^2 - 2\rho\tilde{T}_1\tilde{T}_2 + \tilde{T}_2^2)/(1 - \rho^2), \qquad (3.5.5)$$

$$\tilde{T}_1 = \frac{1}{\sqrt{n\sigma_{11}}}\sum_{i=1}^{n}[\ln\Phi(\hat{\varepsilon}_i)+1], \quad \tilde{T}_2 = \frac{1}{\sqrt{n\sigma_{22}}}\sum_{i=1}^{n}[\ln(1-\Phi(\hat{\varepsilon}_i))+1].$$

Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $\tilde{T} > \chi^2_\alpha(2)$.

**3.5.3 pastaba.** Kriterijus galima sudaryti imant atskirai statistikas $\tilde{T}_1^2$ ir $\tilde{T}_2^2$. Priimant sprendimą šių statistikų realizacijas reikėtų palyginti su kritine reikšme $\chi^2_\alpha(1)$.

**Lognormalusis skirstinys.** Pasiskirstymo funkcija

$$F_0(x|\mu,\sigma) = \Phi(\frac{\ln x - \mu}{\sigma}), \quad x > 0, \quad \mu \in \mathbf{R}, \ \ \sigma > 0.$$

Atlikus transformaciją $Z = \ln X$, gaunamas normalusis skirstinys. Pritaikomi 3.5.1 skyrelio rezultatai, jeigu prieš tai atliekama kiekvieno imties elemento transformacija $Z_i = \ln X_i, i = 1, ..., n$.

**3.5.1 pavyzdys** (**2.3.2** pavyzdžio tęsinys). Pagal 2.3.2 pavyzdžio duomenis patikrinsime hipotezę, kad stebimo a. d. $V$ skirstinys yra a) normalusis; b) lognormalusis; c) a. d. $V^{1/4}$ skirstinys yra normalusis.

a) Gauname $\bar{X} = 12,0184, s = 9,9296$; statistika (3.5.4) įgijo reikšmę 17,2647; asimptotinė P reikšmė $pv_a = \mathbf{P}\{\chi^2_2 > 17,2647\} = 0,00018$; hipotezė atmestina. Naudodami kriterijų, kurio statistika yra $\tilde{T}_1^2$, gauname jos realizaciją 15,6712 ir asimptotinę P reikšmę $pv_a = \mathbf{P}\{\chi^2_1 > 15,6712\} = 0,000075$; kriterijus pasirodė galingesnis. O naudojant statistiką $T_2$, jos realizacija yra 1,5935 ir hipotezė neatmetama. Tai galima paaiškinti tuo, kad stebimo a. d. skirstinys asimetriškas (tai akivaizdu iš histogramos). Naudodami statistiką (3.5.5) randame $\tilde{T}$ realizaciją 19,1807 ir asimptotinę P reikšmę $pv_a = \mathbf{P}\{\chi^2_2 > 19,1807\} = 0,000068$. Hipotezė atmetama. Šiame pavyzdyje kriterijai, grindžiami statistikomis $T_1^2$ ir $\tilde{T}$ pasirodė gerokai galingesni už modifikuotąjį $\chi^2$ kriterijų.

b) Atliekame transformaciją $X_i = \ln V_i$ ir randame $\bar{X} = 2,1029, s = 0,9675$; statistikų (3.5.4) ir (3.5.5) realizacijos yra $T = 4,0410$, $\tilde{T} = 3,3480$, atitinkamos P reikšmės yra 0,1326 ir 0,1875. Hipotezė neatmetama.

c) Atliekame transformaciją $X_i = V_i^{1/4}$ ir randame $\bar{X} = 1,7394, s = 0,3944$; statistikų (3.5.4) ir (3.5.5) realizacijos yra $T = 0,2088$; $\tilde{T} = 0,4484$, atitinkamos P reikšmės yra 0,9009; 0,7992. Hipotezė neatmetama.

## 3.5.2. Logistinis skirstinys

Pagal didumo $n$ paprastąją imtį $\mathbf{X} = (X_1, ..., X_n)^T$ tikrinama sudėtinė suderinamumo hipotezė

$$H_0 : X \sim F_0(\frac{x-\mu}{\sigma}), \quad \mu \in \mathbf{R}, \quad \sigma > 0,$$

$$F_0(x) = \frac{e^x}{1+e^x}, \quad f_0(x) = F_0'(x) = \frac{e^x}{(1+e^x)^2}, \quad -\infty < x < \infty \qquad (3.5.6)$$

1. *Neimano ir Bartono tipo kriterijus.* Imant $k = 2$ parametrus Neimano pasiūlytos transformacijos yra

$$G_1(z_i) = 2\sqrt{3}z_i, \quad G_2(z_i) = \sqrt{5}(6z_i^2 - 1/2), \qquad (3.5.7)$$

čia $z_i = F_0(\hat{\varepsilon}_i) - 1/2, \quad \hat{\varepsilon}_i = (X_i - \hat{\mu})/\hat{\sigma}, \quad i = 1, ..., n.$

Randame
1) $j^{11} = 3, \;\; j^{12} = 0, \;\; j^{22} = 9/(\pi^2 + 3)$;

$$2) A_{11} = 2\sqrt{3}\int_{-\infty}^{\infty} \frac{e^{2x}}{(1+e^x)^4}dx = \sqrt{3}/3, \quad A_{12} = 2\sqrt{3}\int_{-\infty}^{\infty} \frac{xe^{2x}}{(1+e^x)^4}dx = 0;$$

$$A_{21} = 0, \quad A_{22} = 6\sqrt{5}\int_{-\infty}^{\infty} \frac{x(e^x-1)e^{2x}}{(1+e^x)^5}dx = \sqrt{5}/2.$$

Matome, kad $\sigma_{11} = 1 - A_{11}^2 j^{11} = 0$, t. y. asimptotinis skirstinys išsigimęs. Todėl vietoje $G_1$ imkime transformaciją $G_3(z_i) = \sqrt{7}(20z_i^3 - 3z_i)$. Tada

$$A_{31} = 60\sqrt{7}\int_{-\infty}^{\infty} (F_0(x) - 1/2)^2 f_0^2(x)dx - 3\sqrt{7}\int_{-\infty}^{\infty} f_0^2(x)dx = 0, \quad A_{32} = 0,$$

t. y. pataisų matricos elementai lygūs 0.
3) kadangi transformacijos ortogonalios ir normuotos, tai $\mathbf{\Sigma}_0 = \boldsymbol{I}$;
4) kovariacinė matrica

$$\mathbf{\Sigma} = \begin{pmatrix} 1 - A_{22}^2 j^{22} & 0 \\ 0 & 1 \end{pmatrix} \approx \begin{pmatrix} 0,1258 & 0 \\ 0 & 1 \end{pmatrix}.$$

5) – 6) randame statistikos

$$T = T_2^2 + T_3^2 = \hat{T}_2^2/0,1258 + \hat{T}_3^2 \tag{3.5.8}$$

realizaciją $t$. Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $t$ viršija kritinę reikšmę $\chi^2_\alpha(2)$ arba kai $pv_a = \mathbf{P}\{\chi^2_2 > t\} < \alpha$.

2. *Kriterijai, grindžiami beta skirstiniu.*
2) randame

$$A_{11} = \int_{-\infty}^{\infty} \frac{f_0^2(x)}{F_0(x)}dx = 1/2, \quad A_{12} = \int_{-\infty}^{\infty} \frac{xf_0^2(x)}{F_0(x)}dx = -1/2,$$

$$A_{21} = \int_{-\infty}^{\infty} \frac{-f_0^2(x)}{1-F_0(x)}dx = -A_{11}, \quad A_{22} = \int_{-\infty}^{\infty} \frac{-xf_0^2(x)}{1-F_0(x)}dx = A_{12}.$$

3) – 4) pataisų kovariacinės matricos $\tilde{\mathbf{\Sigma}}$ elementai

$$\tilde{\sigma}_{11} = \tilde{\sigma}_{22} = A_{11}^2 j^{11} + A_{12}^2 j^{22} = \frac{1}{4}(j^{11} + j^{22}) \approx 0,92483,$$

$$\tilde{\sigma}_{12} = \tilde{\sigma}_{21} = -A_{11}^2 j^{11} + A_{12}^2 j^{22} = -\frac{1}{4}(j^{11} - j^{22}) \approx -0,57517$$

ir

$$\mathbf{\Sigma} = \mathbf{\Sigma}_0 - \tilde{\mathbf{\Sigma}} = \begin{pmatrix} 1 - \tilde{\sigma_{11}} & 1 - \pi^2/6 - \tilde{\sigma}_{12} \\ 1 - \pi^2/6 - \tilde{\sigma}_{21} & 1 - \tilde{\sigma}_{22} \end{pmatrix} \approx \begin{pmatrix} 0,07517 & -0,06976 \\ -0,06976 & 0,07517 \end{pmatrix}.$$

5) – 6) Gauname kvadratinę formą

$$\tilde{T} = (\tilde{T}_1^2 - 2\rho\tilde{T}_1\tilde{T}_2 + \tilde{T}_2^2)/(1-\rho^2), \quad \rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} \approx -0,9280, \tag{3.5.9}$$

$$\tilde{T}_1 = \frac{1}{\sqrt{n\sigma_{11}}}\sum_{i=1}^{n}[\ln F_0(\hat{\varepsilon}_i)+1], \quad \tilde{T}_2 = \frac{1}{\sqrt{n\sigma_{22}}}\sum_{i=1}^{n}[\ln(1-F_0(\hat{\varepsilon}_i))+1].$$

Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $\tilde{T} > \chi^2_\alpha(2)$.

**Loglogistinis skirstinys.** Pasiskirstymo funkcija $F(x|\theta,\nu) = 1-(1+(x/\theta)^\nu)^{-1}$. Atlikę transformaciją $Z_i = \ln X_i$, gauname logistinių skirstinių šeimą.

### 3.5.3. Ekstremalių reikšmių skirstinys

Pagal didumo $n$ paprastąją imtį $\mathbf{X} = (X_1, ..., X_n)^T$ tikrinama sudėtinė suderinamumo hipotezė

$$H_0 : X \sim F_0(\frac{x-\mu}{\sigma}), \quad \mu \in \mathbf{R}, \quad \sigma > 0,$$

$$F_0(x) = 1 - e^{-e^x}, \quad f_0(x) = F_0'(x) = e^x e^{-e^x}, \quad -\infty < x < \infty. \tag{3.5.10}$$

1. *Neimano ir Bartono tipo kriterijus.* Imant $k = 2$ parametrus Neimano pasiūlytos transformacijos yra

$$G_1(z_i) = 2\sqrt{3}z_i, \quad G_2(z_i) = \sqrt{5}(6z_i^2 - 1/2), \tag{3.5.11}$$

čia $z_i = F_0(\hat{\varepsilon}_i) - 1/2, \quad \hat{\varepsilon}_i = (X_i - \hat{\mu})/\hat{\sigma}, \quad i = 1, ..., n.$

Randame 1)

$$j^{11} = \frac{1+2\Gamma'(1)+\Gamma''(1)}{\Gamma''(1)-(\Gamma'(1))^2} \approx 1,10866, \quad j^{12} = -\frac{1+\Gamma'(1)}{\Gamma''(1)-(\Gamma'(1))^2} \approx -0,25702,$$

$$j^{22} = \frac{1}{\Gamma''(1)-(\Gamma'(1))^2} \approx 0,60793.$$

2)

$$A_{11} = 2\sqrt{3}\int_{\infty}^{\infty} e^{2x}e^{-2e^x}dx = \sqrt{3}/2 \approx 0,86603,$$

$$A_{12} = 2\sqrt{3}\int_{\infty}^{\infty} xe^{2x}e^{-2e^x}dx = \sqrt{3}(\Gamma'(1) - \ln 2 + 1)/2 \approx -0,23415,$$

$$A_{21} = 12\sqrt{5}\int_{-\infty}^{\infty}(1/2 - e^{-e^x})e^{2x}e^{-2e^x}dx = \frac{\sqrt{5}}{6} \approx 0,37268,$$

$$A_{22} = 12\sqrt{5}\int_{-\infty}^{\infty}x(1/2 - e^{-e^x})e^{2x}e^{-2e^x}dx \approx 1,10799.$$

3) Kovariacinė matrica $\mathbf{\Sigma}_0 = \boldsymbol{I}$.

4) Pataisų kovariacinės matricos $\tilde{\mathbf{\Sigma}}$ elementai

$$\tilde{\sigma}_{11} = A_{11}^2 j^{11} + 2A_{11}A_{12}j^{12} + A_{21}^2 j^{22} \approx 0,96907,$$

$$\tilde{\sigma}_{12} = A_{11}A_{21}j^{11} + (A_{11}A_{22} + A_{12}A_{21})j^{12} + A_{12}A_{22}j^{22} \approx -0,02409,$$

$$\tilde{\sigma}_{22} = A_{21}^2 j^{11} + 2A_{21}A_{22}j^{12} + A_{22}^2 j^{22} \approx 0,68804,$$

ir kovariacinė matrica

$$\mathbf{\Sigma} = \mathbf{\Sigma}_0 - \tilde{\mathbf{\Sigma}} \approx \begin{pmatrix} 0,03093 & 0,02409 \\ 0,02409 & 0,32296 \end{pmatrix}, \quad \rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} \approx 0,24103.$$

5) – 6) Gauname kvadratinę formą

$$T = (T_1^2 - 2\rho T_1 T_2 + T_2^2)/(1-\rho^2). \tag{3.5.12}$$

Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $T > \chi^2_\alpha(2)$.

2. *Kriterijai, grindžiami beta skirstiniu.*

2) Randame

$$A_{11} = \int_{-\infty}^{\infty} \frac{e^{2x}e^{-2e^x}}{1-e^{-e^x}} dx = \frac{\pi^2}{6} - 1 \approx 0,64493,$$

$$A_{12} = \int_{-\infty}^{\infty} x \frac{e^{2x}e^{-2e^x}}{1-e^{-e^x}} dx \approx -0,24209,$$

$$A_{21} = \int_{-\infty}^{\infty} -e^{2x}e^{-e^x} dx = -1,$$

$$A_{22} = \int_{-\infty}^{\infty} -xe^{2x}e^{-e^x} dx = -(1+\Gamma'(1)) \approx -0,42278.$$

4) Pataisų kovariacinės matricos $\tilde{\mathbf{\Sigma}}$ elementas

$$\tilde{\sigma}_{22} = A_{21}^2 j^{11} + 2A_{21}A_{22}j^{12} + A_{22}^2 j^{22} = 1.$$

Tada kovariacinės matricos $\mathbf{\Sigma}$ elementas $\sigma_{22} = 1 - \tilde{\sigma}_{22} = 0$. Asimptotinis skirstinys išsigimęs, todėl sudarydami kriterijų naudosime tik pirmąją transformaciją.

$$\tilde{\sigma}_{11} = A_{11}^2 j^{11} + 2A_{11}A_{12}j^{12} + A_{12}^2 j^{22} \approx 0,57702, \quad \sigma_{11} \approx 0,42298.$$

5) – 6) Gauname statistiką

$$\tilde{T} = \tilde{T}_1^2, \quad \tilde{T}_1 = \frac{1}{\sqrt{n\sigma_{11}}} \sum_{i=1}^{n} [\ln(F_0(\hat{\varepsilon}_i)) + 1]. \tag{3.5.13}$$

Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $\tilde{T} > \chi^2_\alpha(1)$.

**Maksimaliųjų reikšmių skirstinys.** Nagrinėtas ekstremaliųjų reikšmių skirstinys dar vadinamas *minimaliųjų reikšmių* skirstiniu. Jis turi kairiąją asimetriją, asimetrijos koeficientas $\gamma = -1$. Kitas ekstremaliųjų reikšmių skirstinys vadinamas *maksimaliųjų reikšmių* skirstiniu. Pasiskirstymo ir tankio funkcijos yra

$$F_0^*(x) = e^{-e^{-x}}, \quad f_0^*(x) = e^{-x}e^{-e^{-x}}, \quad -\infty < x < \infty.$$

Minimaliųjų ir maksimaliųjų skirstinių pasiskirstymo funkcijos susietos lygybe

$$F_0(x) = 1 - F_0^*(-x).$$

Todėl tikrinant sudėtinę suderinamumo hipotezę

$$H_0 : X \sim F_0^*((x-\mu)/\sigma), \quad \mu \in \mathbf{R}, \sigma > 0, \tag{3.5.14}$$

pritaikoma šio skyrelio metodika. Tik įvertinus parametrus ir atlikus transformaciją $Y_i = F_0^*(-(X_i - \hat{\mu})/\hat{\sigma})$ reikia atlikti keitimą $Z_i = 1 - Y_i$, $i = 1, ..., n$ ir visose formulėse vietoje $Y_i$ įrašyti $Z_i$.

**Veibulo skirstinys.** Tikrinant sudėtinę suderinamumo hipotezę

$$H_0 : X \sim F_0(x|\nu/\sigma) = 1 - e^{-(x\sigma)^\nu}, \quad \nu, \sigma > 0, \tag{3.5.15}$$

taip pat pritaikoma šio skyrelio metodika, nes, atlikus transformaciją $Z = \ln X$, Veibulo skirstinių šeima tampa minimalių reikšmių skirstinių šeima.

### 3.5.4. Koši skirstinys

Pagal didumo $n$ paprastąją imtį $\mathbf{X} = (X_1, ..., X_n)^T$ tikrinama sudėtinė suderinamumo hipotezė

$$H_0 : X \sim F_0(\frac{x-\mu}{\sigma}), \quad \mu \in \mathbf{R}, \quad \sigma > 0,$$

$$F_0(x) = \frac{1}{\pi}\operatorname{arctg}(x) + \frac{1}{2}, \quad f_0(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad -\infty < x < \infty. \tag{3.5.16}$$

1. *Neimano ir Bartono tipo kriterijus.* Imant $k = 2$ parametrus Neimano pasiūlytos transformacijos yra

$$G_1(z_i) = 2\sqrt{3}z_i, \quad G_2(z_i) = \sqrt{5}(6z_i^2 - 1/2), \tag{3.5.17}$$

čia $z_i = F_0(\hat{\varepsilon}_i) - 1/2, \quad \hat{\varepsilon}_i = (X_i - \hat{\mu})/\hat{\sigma}, \quad i = 1, ..., n.$

Randame 1) $j^{11} = 2, \quad j^{12} = 0, \quad j^{22} = 2.$
2)

$$A_{11} = \frac{2\sqrt{3}}{\pi^2}\int_\infty^\infty \frac{dx}{(1+x^2)^2} = \frac{\sqrt{3}}{\pi}, \quad A_{12} = \frac{2\sqrt{3}}{\pi^2}\int_\infty^\infty \frac{xdx}{(1+x^2)^2} = 0,$$

$$A_{21} = 0, \quad A_{22} = \frac{12\sqrt{5}}{\pi^3} \int_{\infty}^{\infty} \frac{x \operatorname{arctg} x dx}{(1+x^2)^2} = \frac{3\sqrt{5}}{\pi^2},$$

3) kadangi transformacijos ortogonalios ir normuotos, tai $\mathbf{\Sigma}_0 = \boldsymbol{I}$;
4) kovariacinė matrica

$$\mathbf{\Sigma} = \begin{pmatrix} 1 - A_{11}^2 & 0 \\ 0 & 1 - A_{22}^2/2 \end{pmatrix} = \begin{pmatrix} 1 - 3/\pi^2 & 0 \\ 0 & 1 - 45/\pi^4 \end{pmatrix}.$$

5) – 6) randame statistikos

$$T = T_1^2 + T_2^2 = \hat{T}_1^2/(1 - 3/\pi 2) + \hat{T}_2^2/(1 - 45/\pi^4) \tag{3.5.18}$$

realizaciją $t$. Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $t$ viršija kritinę reikšmę $\chi_\alpha^2(2)$ arba kai $pv_a = \mathbf{P}\{\chi_2^2 > t\} < \alpha$.

2. *Kriterijai, grindžiami beta skirstiniu.*
2) Randame

$$A_{11} = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{((1/\pi) \operatorname{arctg} x + 1/2)^{-1}}{(1+x^2)^2} dx = 0,38796,$$

$$A_{12} = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{x((1/\pi) \operatorname{arctg} x + 1/2)^{-1}}{(1+x^2)^2} dx = -0,22571$$

$$A_{21} = -A_{11}, \quad A_{22} = A_{12}.$$

4) Pataisų kovariacinės matricos $\tilde{\mathbf{\Sigma}}$ elementai

$$\tilde{\sigma}_{11} = \tilde{\sigma}_{22} = A_{11}^2 j^{11} + A_{21}^2 j^{22} \approx 0,40292,$$

$$\tilde{\sigma}_{12} = \tilde{\sigma}_{21} = -A_{11}^2 j^{11} + A_{12}^2 j^{22} \approx -0,19914,$$

ir kovariacinė matrica

$$\mathbf{\Sigma} = \mathbf{\Sigma}_0 - \tilde{\mathbf{\Sigma}} \approx \begin{pmatrix} 0,59708 & -0,44580 \\ -0,44580 & 0,59708 \end{pmatrix}.$$

5) – 6) Gauname kvadratinę formą

$$\tilde{T} = (\tilde{T}_1^2 - 2\rho \tilde{T}_1 \tilde{T}_2 + \tilde{T}_2^2)/(1 - \rho^2), \quad \rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} \approx -0,74663, \tag{3.5.19}$$

$$\tilde{T}_1 = \frac{1}{\sqrt{n\sigma_{11}}} \sum_{i=1}^{n} [\ln(F_0(\hat{\varepsilon}_i)) + 1], \quad \tilde{T}_2 = \frac{1}{\sqrt{n\sigma_{22}}} \sum_{i=1}^{n} [\ln(1 - F_0(\hat{\varepsilon}_i)) + 1].$$

Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $\tilde{T} > \chi_\alpha^2(2)$.

# 3.6. Pratimai

**3.1 – 3.3 skyreliai**

**3.1.** Pagal paprastąją didumo $n$ imtį $\mathbf{X} = (X_1, ..., X_n)^T$ tikrinama paprastoji suderinamumo hipotezė $H_0 : X_i \sim \mathcal{E}(1)$, kai alternatyvų aibė yra $\{\mathcal{E}(\lambda),\ \lambda \neq 0,\ \lambda > 0\}$. Raskite alternatyvų aibę $\mathcal{G}$, kai atlikta a. d. $X_1, ..., X_n$ transformacija (3.1.3).

**3.2.** Tarkime, kad atlikę transformaciją (3.1.3) gavome alternatyvų aibę $\{Be(\lambda, 1),\ \lambda \neq 1,\ \lambda > 0\}$. Kokia yra pradinio uždavinio alternatyvų aibė $\mathcal{F}$?

**3.3.** Pagal paprastąją didumo $n$ imtį $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ tikrinama paprastoji suderinamumo hipotezė $H_0 : X_i \sim U(0,\ 1)$, kai Neimano tipo alternatyvių tankių aibė yra $\mathcal{G} = \{g(y|\theta) = \frac{1}{c(\theta)} e^{\theta(y-1/2)}, 0 < y < 1, \theta > 0\}$. Raskite TG kriterijų hipotezei $H_0$, kai alternatyva yra $H : Y_i \sim g \in \mathcal{G}$, tikrinti. Naudodami normaliąją aproksimaciją suformuluokite asimptotinį kriterijų.

**3.4.** (**3.3** pratimo tęsinys). Raskite 3.3 pratime surasto kriterijaus statistikos asimptotinį skirstinį, kai teisinga alternatyva. Naudodami normaliąją aproksimaciją raskite asimptotinio reikšmingumo lygmens $\alpha$ kriterijaus galios funkciją.

**3.5.** (**3.4** pratimo tęsinys). Apskaičiuokite asimptotinę reikšmingumo lygmens $\alpha = 0,05$ kriterijaus galią, kai a)$n = 50;\ \theta = 1,2;\ 1,5;\ 2;\ 3;$ b) $\theta = 0,5; n = 50;\ 100;\ 200.$

**3.6.** Pagal paprastąją didumo $n$ imtį $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ tikrinama paprastoji suderinamumo hipotezė $H_0 : X_i \sim U(0,\ 1)$, kai Neimano tipo alternatyvių tankių aibė yra $\mathcal{G} = \{g(y|\theta) = \frac{1}{c(\theta)} e^{\theta(y-1/2)^2}, 0 < y < 1, \theta > 0\}$. Raskite TG kriterijų hipotezei $H_0$, kai alternatyva yra $H : Y_i \sim g \in \mathcal{G}$, tikrinti. Naudodami normaliąją aproksimaciją suformuluokite asimptotinį kriterijų.

**3.7.** (**3.6** pratimo tęsinys). Raskite 3.6 pratime surasto kriterijaus statistikos asimptotinį skirstinį, kai teisinga alternatyva. Naudodami normaliąją aproksimaciją raskite asimptotinio reikšmingumo lygmens $\alpha$ kriterijaus galios funkciją.

**3.8.** (**3.7** pratimo tęsinys). Apskaičiuokite asimptotinę reikšmingumo lygmens $\alpha = 0,05$ kriterijaus galią, kai a)$n = 50;\ \theta = 1,2;\ 1,5;\ 2;\ 3;$ b) $\theta = 1,0; n = 50;\ 100;\ 200.$

**3.9.** Pagal paprastąją didumo $n$ imtį $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ tikrinama paprastoji suderinamumo hipotezė $H_0 : X_i \sim U(0,\ 1)$, kai alternatyvių tankių aibė yra $\mathcal{G} = \{g(y|\gamma) = \gamma(1-y)^{\gamma-1}, 0 < y < 1, \gamma > 1\}$. Raskite TG kriterijų hipotezei $H_0$, kai alternatyva yra $H : Y_i \sim g \in \mathcal{G}$, tikrinti ir jo galios funkciją.

**3.10.** Pagal paprastąją didumo $n$ imtį $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ tikrinama paprastoji suderinamumo hipotezė $H_0 : X_i \sim U(0,\ 1)$, kai alternatyvių tankių aibė yra $\mathcal{G} = \{g(y|\gamma) = \gamma(1-y)^{\gamma-1}, 0 < y < 1,\ 0 < \gamma < 1\}$. Raskite TG kriterijų hipotezei $H_0$, kai alternatyva yra $H : Y_i \sim g \in \mathcal{G}$, tikrinti ir jo galios funkciją.

**3.11.** Pagal paprastąją didumo $n$ imtį $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ tikrinama paprastoji suderinamumo hipotezė $H_0 : X_i \sim U(0,\ 1)$, kai alternatyvių tankių aibė yra 1) $\mathcal{G} = \{g(y|\gamma) = \gamma y^{\gamma-1}, 0 < y < 1,\ 0 < \gamma < 1\}$; 2) $\mathcal{G} = \{g(y|\gamma) = \gamma y^{\gamma-1}, 0 < y < 1,\ 1 < \gamma\}$. Raskite TG kriterijus ir jų galios funkcijas.

**3.4 – 3.5 skyreliai**

**3.12.** Raskite modifikuotojo Neimano ir Bartono tipo (2 parametrai) asimptotinį kriterijų eksponentiškumo hipotezei $H_0 : X_i \sim \mathcal{E}(1/\lambda), \lambda > 0$ tikrinti.

**3.13.** Raskite modifikuotąjį asimptotinį kriterijų, grindžiamą beta skirstiniu, eksponentiškumo hipotezei $H_0 : X_i \sim \mathcal{E}(1/\lambda), \lambda > 0$ tikrinti.

**3.14.** (**3.12** ir **3.13** pratimų tęsinys). Remdamiesi 3.12 ir 3.13 pratimuose rastais kriterijais patikrinkite hipotezę, kad 2.3.1 pavyzdžio duomenys gauti stebint eksponentinį a. d.

**3.15.** (**2.21** pratimo tęsinys). Modifikuotuoju Neimano ir Bartono tipo ir beta skirstiniu grindžiamais kriterijais patikrinkite hipotezę, kad 2.21 pratimo duomenys gauti stebint normalųjį a. d.

**3.16.** (**2.18** pratimo tęsinys). Modifikuotuoju Neimano ir Bartono tipo ir beta skirstiniu grindžiamais kriterijais patikrinkite hipotezę, kad 2.18 pratimo duomenys gauti stebint a) normalųjį a. d.; b) logistinį a. d.

**3.17.** (**2.19** pratimo tęsinys). Modifikuotuoju Neimano ir Bartono tipo ir beta skirstiniu grindžiamais kriterijais patikrinkite hipotezę, kad 2.19 pratimo duomenys gauti stebint a) lognormalųjį a. d.; b) loglogistinį a. d.

**3.18.** Sumodeliuokite didumo $n = 50$ paprastąją imtį, gautą stebint normalųjį a. d. ir, taikydami 3.5 skyrelio kriterijus, patikrinkite hipotezę, kad buvo sumodeliuotas a) normalusis a. d.; b) logistinis a. d.; c) Koši a. d.

**3.19.** Sumodeliuokite didumo $n = 50$ paprastąją imtį, gautą stebint normalųjį a. d. ir, taikydami 3.5 skyrelio kriterijus, patikrinkite hipotezę, kad buvo sumodeliuotas a) normalusis a. d.; b) logistinis a. d.; c) Koši a. d.

**3.20.** Sumodeliuokite didumo $n = 50$ paprastąją imtį, gautą stebint Koši a. d. ir, taikydami 3.5 skyrelio kriterijus, patikrinkite hipotezę, kad buvo sumodeliuotas a) normalusis a. d.; b) logistinis a. d.; c) Koši a. d.

**3.21.** Sumodeliuokite didumo $n = 50$ paprastąją imtį, gautą stebint Veibulo a. d. ir, taikydami 3.5 skyrelio kriterijus, patikrinkite hipotezę, kad buvo sumodeliuotas a) Veibulo a. d.; b) lognormalusis a. d.; c) maksimalių reikšmių a. d.

## 3.7. Atsakymai ir nurodymai

**3.1.** $\{Be(1,\lambda), \lambda \neq 1,\ \lambda > 0\}$. **3.2.** Alternatyvų aibę $\mathcal{F}$ sudaro tankiai $f(x) = \lambda e^{-x}(1-e^{-x})^{\lambda-1}, x > 0, \lambda \neq 1$. Jeigu a. d. $Y_i$ pakeistume a. d $Z_i = 1 - Y_i$, tai alternatyvų aibė būtų kaip **3.1** pratime $\{\mathcal{E}(\lambda), \lambda \neq 1, \lambda > 0\}$. **3.3.** Hipotezė atmetama, kai $T = 2sqrt3/n\sum_{i=1}^n (Y_i - 1/2)) > t_\alpha$, čia $t_\alpha$ yra statistikos $T$ lygmens $\alpha$ kritinė reikšmė. Asimptotinis kriterijus gaunamas pakeičiant $t_\alpha$ į $z_\alpha$. **3.4.** Pažymėkime $\mu(\theta) = \mathbf{E}_\theta(Y_i - 1/2)$ ir $\sigma^2(\theta) = \boldsymbol{V}_\theta(Y_i - 1/2)$. Jeigu $\theta$ tikroji parametro reikšmė, tai $T(\theta) = (1/(\sigma(\theta)\sqrt{n}))\sum_{i=1}^n (Y_i - 1/2 - \mu(\theta)) \xrightarrow{d} Z \sim N(0,\ 1)$. Asimptotinė kriterijaus galia $\beta(\theta) = \Phi(\sqrt{n}\mu(\theta)/\sigma(\theta) - z_\alpha/(\sqrt{12}\sigma(\theta)))$. **Nurodymas.** Randame normuojančią konstantą $c(\theta) = (e^{\theta/2} - e^{-\theta/2})/\theta$. Tada $\mu(\theta) = [\ln c(\theta)]'_\theta = [e^\theta(\theta - 2) + \theta + 2]/(2\theta(e^\theta - 1))$; $\sigma^2(\theta) = [\ln c(\theta)]''_{\theta^2} = e^\theta(e^\theta + e^{-\theta} - \theta^2 - 2)/(\theta(e^\theta - 1))^2$. **3.5.** a) 0,7807; 0,9164; 0,9919; 1,000; b)0,6451; 0,8897; 0,9925. **3.6.** Hipotezė atmetama, kai $T = \sum_{i=1}^n [(Y_i - 1/2)^2 - 1/12]6\sqrt{5}/\sqrt{n} > t_\alpha$, čia $t_\alpha$ yra statistikos $T$ lygmens $\alpha$ kritinė reikšmė. Asimptotinis kriterijus gaunamas pakeičiant $t_\alpha$ į $z_\alpha$. **3.7.** Pažymėkime $\mu(\theta) = \mathbf{E}_\theta(Y_i - 1/2)^2$ ir $\sigma^2(\theta) = \boldsymbol{V}_\theta(Y_i - 1/2)^2$. Jeigu $\theta$ tikroji parametro reikšmė, tai $T(\theta) = \sum_{i=1}^n [(Y_i - 1/2)^2 - \mu(\theta)]/(\sigma(\theta)\sqrt{n}) \xrightarrow{d} Z \sim N(0,\ 1)$. Asimptotinė kriterijaus galia $\beta(\theta) = \Phi(\sqrt{n}(\mu(\theta) - 1/12)/\sigma(\theta) - z_\alpha/(6\sqrt{5}\sigma(\theta)))$. **Nurodymas.** Normuojanti konstanta $c(\theta) = \int_0^1 \exp\{\theta(x - 1/2)^2\}dx$, $c'(\theta) = \int_0^1 (x - 1/2)^2 \exp\{\theta(x - 1/2)^2\}dx$, $c''(\theta) = \int_0^1 (x - 1/2)^4 \exp\{\theta(x - 1/2)^2\}dx$. Tada $\mu(\theta) = c'(\theta)/c(\theta)$, $\sigma^2(\theta) = c''(\theta)/c(\theta) - [c'(\theta)/c(\theta)]^2$. Kai $\theta$ žinomas, integralus galima apskaičiuoti skaitiniais metodais. **3.8.** a) 0,1668; 0,2122; 0,3016; 0,5148; b) 0,1403; 0,1950; 0,2912. **3.9.** Reikšmingumo lygmens $\alpha$ TG kriterijus atmeta hipotezę $H_0$, kai $T = -2\sum_{i=1}^n \ln(1-Y_i) < \chi^2_{1-\alpha}(2n)$. Galios funkcija $\beta(\gamma) = \mathbf{P}\{\chi^2_{2n} < \gamma\chi^2_{1-\alpha}(2n)\} \to 1$, kai $\gamma \to 0$. **Nurodymas.** Grįžkite prie eksponentinio skirstinio (žr. 3.1 pratimą). **3.10.** Reikšmingumo lygmens $\alpha$ TG kriterijus atmeta hipotezę $H_0$, kai $T = -2\sum_{i=1}^n \ln(1-Y_i) > \chi^2_\alpha(2n)$. Galios funkcija $\beta(\gamma) = \mathbf{P}\{\chi^2_{2n} > \gamma\chi^2_\alpha(2n)\} \to 1$, kai $\gamma \to \infty$. **3.11.** Atlikę keitimą $Z_i = 1 - Y_i$ gauname 3.9, 3.10 pratimų alternatyvų šeimas. **3.12.** Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $T = (T_1^2 - 2\rho T_1 T_2 + T_2^2)/(1-\rho^2) > \chi^2_\alpha(2)$, čia $T_1 = 4\sqrt{3}\sum_{i=1}^n (Y_i - 1/2)/\sqrt{n}$,, $T_2 = 6\sqrt{5}\sum_{i=1}^n [6(Y_1 - 1/2)^2 - 1/2]/\sqrt{31n}$, $Y_i = 1 - \exp(-X_i/\bar{X})$, $\rho = -0,6956$. **Nurodymas.** Pakartokite 3.4.2 ir 3.4.3 teoremų įrodymus, kai yra eksponentinis skirstinys (vienas mastelio parametras). **3.13.** Statistikos $\hat{T}_2 = \sum_i (\ln(1-Y_i)+1)/\sqrt{n}$ skirstinys išsigimęs. Todėl kriterijų sudarome naudodmi tik statistiką $\hat{T}_1 = \sum_i \ln Y_i + 1/\sqrt{n\sigma_{11}}$, $\sigma_{11} = \pi^2(12-\pi^2)/36$. Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $\hat{T}_1^2 > \chi^2_\alpha(1)$. **3.14.** Statistikos $T$ ir $\hat{T}_1^2$ įgijo reikšmes 11,80085 ir 7,1079; atitinkamos P reikšmės 0,0027 ir 0,0077. Hipotezė atmestina. **3.15.** Statistikos $T$ ir $\tilde{T}$ įgijo reikšmes 0,2714 ir 0,4506; atitinkamos P reikšmės 0,8731 ir 0,7983. Hipotezė neatmetama.

# 4 skyrius

# Kriterijai, grindžiami empiriniais procesais

Sakykime, $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis a. d. $X$, kurio pasiskirstymo funkcija $F$ priklauso absoliučiai tolydžių skirstinių šeimai $\mathcal{F}$. Tikrinsime paprastąją hipotezę

$$H_0 : \; F(x) \equiv F_0(x); \tag{4.0.1}$$

čia $F_0$ žinoma šeimos $\mathcal{F}$ pasiskirstymo funkcija.

## 4.1. Kriterijų, grindžiamų empiriniu procesu, statistikos

**Kriterijų, grindžiamų empiriniais procesais, kūrimo idėja.** Pažymėkime

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i)$$

empirinę pasiskirstymo funkciją.

Jei teisinga hipotezė $H_0$, tai pagal Glivenkos ir Kantelio teoremą

$$\sup_{x\in\mathbf{R}} |\hat{F}_n(x) - F_0(x)| \overset{b.t.}{\to} 0, \quad \text{kai} \quad n \to \infty.$$

Taigi, tikrinant hipotezę $H_0$, natūralu sudaryti kriterijus imant tam tikrus empirinio proceso $\mathcal{E}_n = \sqrt{n}(\hat{F}_n - F_0)$ funkcionalus.

**Kriterijaus statistikos.** Dažniausiai naudojami šie funkcionalai:

$$D_n = \sup_{x\in\mathbf{R}} |\hat{F}_n(x) - F_0(x)| \quad (\textit{Kolmogorovo ir Smirnovo statistika}), \tag{4.1.1}$$

$$C_n = \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 dF_0(x) \quad (\textit{Kramero ir Mizeso statistika}), \tag{4.1.2}$$

$$A_n = \int_{-\infty}^{\infty} \frac{(\hat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x) \quad \textit{(Anderseno ir Darlingo statistika)}, \tag{4.1.3}$$

arba, apibendrinant pastarąsias dvi,

$$\omega_n^2 = \omega_n^2(\psi) = \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 \psi(F_0(x)) dF_0(x) \quad (\omega^2 \ \textit{statistika}); \tag{4.1.4}$$

čia $\psi$ – neneigiama funkcija, apibrėžta intervale $(0,\ 1)$.

**4.1.1 teorema.** *Tegu paprastoji imtis $\mathbf{X} = (X_1, ..., X_n)^T$ gauta stebint absoliučiai tolydųjį a. d. $X$ su pasiskirstymo funkcija $F_0$. Tada statistikų (4.1.1) – (4.1.4) skirstiniai nepriklauso nuo pasiskirstymo funkcijos $F_0$, o priklauso tik nuo imties didumo $n$.*

**Įrodymas.** Žinoma, jei $X$ yra absoliučiai tolydus, tai a. d. $Y = F_0(X)$ yra tolygiai pasiskirstęs intervale $[0,\ 1]$, t. y. $Y \sim U(0,\ 1)$. Todėl empirinės pasiskirstymo funkcijos

$$\hat{G}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, y]}(Y_i), \quad Y_i = F_0(X_i), \quad i = 1, ..., n,$$

skirstinys nepriklauso nuo $F_0$.

Funkcija $F_0$ nemažėjanti. Kiekvienam $y \in [0,\ 1]$ apibrėžkime $x_y = \sup\{x : F_0(x) \le y\}$. Tada

$$X_i \le x_y \Longleftrightarrow F_0(X_i) \le F_0(x_y) \Longleftrightarrow Y_i \le F_0(x_y),$$

taigi $\hat{F}_n(x_y) = \hat{G}_n(F_0(x_y))$.

Funkcija $F_0$ yra tolydi. Todėl, kai $x_y$ prabėga galimas reikšmes iš intervalo $[-\infty, \infty]$, tai $y = F(x_y)$ užpildo intervalą $[0,\ 1]$. Gauname

$$D_n = \sup_{x_y \in \mathbf{R}} |\hat{F}_n(x_y) - F_0(x_y)| = \sup_{x_y \in \mathbf{R}} |\hat{G}_n(F_0(x_y)) - F_0(x_y)| = \sup_{y \in [0,\ 1]} |\hat{G}_n(y) - y|,$$

$$\omega_n^2 = \int_{-\infty}^{\infty} (\hat{G}_n(F_0(x)) - F_0(x))^2 \psi(F_0(x)) dF_0(x) = \int_0^1 (\hat{G}_n(y) - y)^2 \psi(y) dy.$$

Taigi statistikų $D_n$ ir $\omega^2$ skirstiniai nepriklauso nuo $F_0$, o priklauso tik nuo imties didumo $n$. ▲

Remiantis teorema hipotezė $H_0$ atmetama, kai minėtos statistikos įgyja dideles reikšmes, t. y. viršija atitinkamų statistikų lygmens $\alpha$ kritines reikšmes. Nedideliems $n$ kritines reikšmes galime surasti, pavyzdžiui, modeliuodami $\hat{G}_n(y)$ ir apskaičiuodami minėtų funkcionalų realizacijas.

Kai $n$ didelis, kritinių reikšmių skaičiavimas sudėtingas, todėl naudojamos statistikų asimptotinių $(n \to \infty)$ skirstinių kritinės reikšmės.

Kadangi visos statistikos yra empirinio proceso

$$\mathcal{E}_n(x) = \sqrt{n}(\hat{F}_n(x) - F_0(x)),\ x \in \mathbf{R}, \tag{4.1.5}$$

funkcionalai, tai, remiantis empirinio proceso invariantiškumo principu (žr. B priedą, 8.4.1 teoremą), gaunami statistikų asimptotiniai skirstiniai. Pagal 8.4.1 teoremą Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo statistikos turi tokias ribas:

$$\sqrt{n}D_n \xrightarrow{d} \sup_{0\le t\le 1}|B(t)|, \quad nC_n \xrightarrow{d} \int_0^1 B^2(t)dt, \quad nA_n \xrightarrow{d} \int_0^1 \frac{B^2(t)}{t(1-t)}dt, \quad (4.1.6)$$

čia $B$ yra intervalo $[0,\,1]$ Brauno tiltas (žr. B priedą, 8.2.3, 8.4, 8.5 skyrelius). Taigi statistikų asimptotiniai skirstiniai sutampa su atitinkamų Brauno tilto funkcionalų skirstiniais.

**Diskrečiojo skirstinio atvejis.** Tarkime, $X$ yra diskretusis a. d., įgyjantis reikšmes $a_1, ..., a_k$ su tikimybėmis $p_1, ..., p_k$, $\sum_i p_i = 1$. Jo pasiskirstymo funkcija $F(x)$ kinta didumo $p_i$ šuoliukais taškuose $a_i$, $i = 1, ..., k$.

Tegu $(X_1, ..., X_n)^T$ yra paprastoji didumo $n$ imtis, gauta stebint a. d. $X$. Pažymėkime $U_i$ reikšmės $a_i$ pasikartojimų skaičių imtyje, $\sum_i U_i = n$. Tada empirinė pasiskirstymo funkcija $\hat{F}_n(x)$ kinta didumo $U_i/n$ šuoliukais taškuose $a_i$, $i = 1, ..., k$.

Tikrinsime paprastąją hipotezę, kad a. d. $X$ pasiskirstymo funkcija $F$ sutampa su žinoma pasiskirstymo funkcija $F_0$, kuri kinta didumo $p_{i0}$ šuoliukais taškuose $a_i$, $\sum_i p_{i0} = 1$.

Apibrėžkime diskrečiuosius Kolmogorovo ir Smirnovo ir $\omega^2$ statistikų analogus

$$\tilde{D}_n = \max_{1\le i\le k}|\hat{F}_n(a_i) - F_0(a_i)| = \max_{1\le i\le k}|\hat{G}_n(t_i) - t_i|,$$

$$\tilde{\omega}^2 = \sum_{i=1}^{k}(\hat{F}_n(a_i) - F_0(a_i))^2\psi(F_0(a_i))p_{i0} = \sum_{i=1}^{k}(\hat{G}_n(t_i) - t_i)^2\psi(t_i)p_{i0},$$

čia $\hat{G}_n(t)$ yra empirinė pasiskirstymo funkcija, kintanti didumo $U_i/n$ šuoliukais taškuose $t_i = p_{10} + ... + p_{i0}$, $i = 1, ..., k$.

Remiantis empirinio proceso savybėmis (B priedas, 8.4.1 teorema)

$$\sqrt{n}\tilde{D}_n \xrightarrow{d} \max_{1\le i\le k}|B(t_i)|, \quad n\tilde{\omega}_n(\psi) \xrightarrow{d} \sum_{i=1}^{k} B^2(t_i)\psi(t_i)p_{i0}.$$

Statistikų $\tilde{D}_n, \tilde{\omega}_n^2$ kritines reikšmes galime rasti modeliuodami a. v. $(U_1, ..., U_k)^T \sim \mathcal{P}_k(n, \boldsymbol{p}_0)$, $\boldsymbol{p}_0 = (p_{10}, ..., p_{k0})^T$. Kai $n$ didelis asimptotines kritines reikšmes galime rasti remdamiesi tuo, kad a. v. $(B(t_1), ..., B(t_k))^T$ turi $k$-matį normalųjį skirstinį su nuliniu vidurkių vektoriumi ir kovariacijomis $\sigma_{ij} = t_i(1 - t_j), t_i \le t_j$ (B priedas, 8.2.3. skyrelis).

Jeigu galimų reikšmių skaičius $k$ yra didelis ir tikimybės $p_{i0}$ mažos, tai kriterijus konstruojame kaip ir tolydžiojo skirstinio atveju. Reikia pažymėti, kad ir tolydžiojo skirstinio atveju stebėjimo duomenys būna suapvalinti tam tikru tikslumu, t. y. faktiškai stebime diskretųjį a. d. su dideliu galimų reikšmių skaičiumi ir mažomis jų įgijimo tikimybėmis (žr. taip pat pratimus 4.6 – 4.10).

## 4.2. Kolmogorovo ir Smirnovo kriterijus

Kolmogorovo ir Smirnovo kriterijus hipotezei $H_0: \; F(x) \equiv F_0(x)$ tikrinti grindžiamas statistika

$$D_n = \sup_{x\in\mathbf{R}}|\hat{F}_n(x) - F_0(x)|. \quad (4.2.1)$$

**Dvipusė alternatyva:**

$$\bar{H} : \sup_{x \in \mathbf{R}} |F(x) - F_0(x)| > 0. \tag{4.2.2}$$

Taigi nuokrypis nuo hipotezės matuojamas tolygiąja metrika. Remiantis 4.1.1 teorema statistikos $D_n$ skirstinys, kai teisinga hipotezė, nepriklauso nuo $F_0$.

**Kolmogorovo ir Smirnovo statistikos skaičiavimas.** Statistikos $D_n$ realizaciją galime rasti remdamiesi pateikiama teorema.

**4.2.1 teorema.** *Tarkime, $X_{(1)} \leq \cdots \leq X_{(n)}$ yra pozicinės statistikos. Tada teisinga lygybė*

$$D_n = max(D_n^+, D_n^-\}; \tag{4.2.3}$$

*čia*

$$D_n^+ = \max_{1 \leq i \leq n} [\hat{F}_n(X_{(i)}) - F_0(X_{(i)})],$$

$$D_n^- = \max_{1 \leq i \leq n} [F_0(X_{(i)}) - \hat{F}_n(X_{(i-1)})].$$

*Jeigu $X_{(1)} < \cdots < X_{(n)}$, tada*

$$D_n^+ = \max_{1 \leq i \leq n} \left( \frac{i}{n} - F_0(X_{(i)}) \right), \quad D_n^- = \max_{1 \leq i \leq n} \left( F_0(X_{(i)}) - \frac{i-1}{n} \right). \tag{4.2.4}$$

**Įrodymas** Jeigu $X_{(i-1)} < X_{(i)}$, tai $F_0(x)$ nemažėja intervale $(X_{(i-1)}, X_{(i)}]$ ir $\hat{F}_n(x) = \hat{F}_n(X_{(i-1)})$ su visais $x \in (X_{(i-1)}, X_{(i)})$, todėl

$$\sup_{x \in (X_{(i-1)}, X_{(i)}]} [\hat{F}_n(x) - F_0(x)] = max[\hat{F}_n(X_{(i-1)}) - F_0(X_{(i-1)}), \hat{F}_n(X_{(i)}) - F_0(X_{(i)})],$$

$$\sup_{x \in (X_{(i-1)}, X_{(i)}]} [F_0(x) - \hat{F}_n(x)] = F_0(X_{(i)}) - \hat{F}_n(X_{(i-1)}).$$

Taigi

$$\sup_{x \in \mathbf{R}} [\hat{F}_n(x) - F_0(x)] = \max_{1 \leq i \leq n} [\hat{F}_n(X_{(i)}) - F_0(X_{(i)})] = D_n^+,$$

$$\sup_{x \in \mathbf{R}} [F_0(x) - \hat{F}_n(x)] = \max_{1 \leq i \leq n} [F_0(X_{(i)}) - \hat{F}_n(X_{(i-1)})] = D_n^-,$$

$$D_n = max\{ \sup_{\hat{F}_n(x) - F_0(x) \geq 0} |\hat{F}_n(x) - F_0(x)|, \sup_{\hat{F}_n(x) - F_0(x) < 0} |\hat{F}_n(x) - F_0(x)| \} =$$

$$max\{ \sup_{x \in \mathbf{R}} [\hat{F}_n(x) - F_0(x)], \sup_{x \in \mathbf{R}} [F_0(x) - \hat{F}_n(x)] \} = max(D_n^+, D_n^- \}.$$

Jeigu $X_{(1)} < \cdots < X_{(n)}$, tai $\hat{F}_n(X_{(i)}) = i/n$, todėl teisinga (4.2.4). ▲

**Kolmogorovo ir Smirnovo kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $D_n > D_\alpha(n)$; čia $D_\alpha(n)$ yra statistikos $D_n$ lygmens $\alpha$

kritinė reikšmė. Nedideliems $n$ kritinės reikšmės $D_\alpha(n)$ yra tabuliuotos [7],[17] t. y. galime rasti pasiskirstymo funkcijos $F_{D_n}$ reikšmes, kartu $P$ reikšmes

$$pv = 1 - F_{D_n}(D_n). \tag{4.2.5}$$

Kai n didelis, naudojamos asimptotinio skirstinio kritinės reikšmės.

**Statistikos $\sqrt{n}D_n$ asimptotinis skirstinys**

Statistikos $\sqrt{n}D_n$ asimptotinis skirstinys gaunamas naudojant sąryšį (4.1.6).

**4.2.2 teorema.** *Tarkime, $X_1, \dots, X_n$ yra paprastoji imtis, gauta stebint absoliučiai tolydųjį a. d. $X$ su pasiskirstymo funkcija $F_0(x)$. Jeigu $n \to \infty$, tai su visais $x \in \mathbf{R}$*

$$\mathbf{P}\{\sqrt{n}D_n \le x\} \to K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}. \tag{4.2.6}$$

**Įrodymas.** Teoremos rezultatas išplaukia iš sąryšio (4.1.6) ir 4 Brauno tilto savybės (B priedas, 8.5. skyrelis): su visais $x > 0$

$$\mathbf{P}\{\sup_{0\le t\le 1} |B_t| \ge x\} = 2\sum_{n=1}^{\infty}(-1)^{n-1} e^{-2n^2x^2}. \tag{4.2.7}$$

▲

Daugumoje matematinės statistikos programų paketų yra numatyta rasti funkcijos $K(x)$ reikšmes. Tada asimptotinė Kolmogorovo ir Smirnovo kriterijaus $P$ reikšmė yra

$$pv_a = 1 - K(\sqrt{nD_n}).$$

**Kritinės reikšmės paprasta aproksimacija.** Jeigu $n > 100$, tai knygoje [7] kritinei reikšmei $D_\alpha(n)$ rasti rekomenduojama naudoti apytiksles formules

$$D_\alpha(n) \approx \sqrt{\frac{1}{2n}(y - \frac{2y^2 - 4y - 1}{18n})} - \frac{1}{6n} \approx \sqrt{\frac{y}{2n}} - \frac{1}{6n}, \tag{4.2.8}$$

čia $y = -\ln(\alpha/2)$. Ši formulė gana tiksli ir kai $n$ mažesni. Pavyzdžiui, $D_{0,05}(20) = 0,2953$ (žr. [7], 6.2 lentelę; [17], IX lentelę). Taikydami tikslesnę aproksimaciją 4.2.8 gauname apytikslę reikšmę 0,29535, o taikydami grubesnę aproksimaciją – 0,29403.

**Vienpusės alternatyvos.** Jeigu alternatyva yra vienpusė, t. y.

$$\bar{H}_1 : \sup_{x\in\mathbf{R}}(F(x) - F_0(x)) > 0 \quad \text{arba} \quad \bar{H}_2 : \sup_{x\in\mathbf{R}}(F(x) - F_0(x)) < 0,$$

tai kriterijus grindžiamas statistikomis $D_n^+$ arba $D_n^-$, kurių skirstiniai dėl simetrijos yra vienodi. Hipotezė $H_0$ atmetama, kai

$$D_n^+ > D_\alpha^+(n) \quad \text{arba} \quad D_n^- > D_\alpha^+(n);$$

čia $D^+_\alpha(n)$ yra statistikos $D^+_n$ lygmens $\alpha$ kritinė reikšmė.

Smirnovas [28] rado tikslų ir asimptotinį statistikos $D^+_n$ skirstinį: su visais $x \in [0,\, 1)$

$$\mathbf{P}\{D^+_n \le x\} = 1 - (1-x)^n - x \sum_{j=1}^{[n(1-x)]} C^j_n \left(1 - x - \frac{j}{n}\right)^{n-j} \left(x + \frac{j}{n}\right)^{j-1}. \tag{4.2.9}$$

Taigi alternatyvų $\bar{H}_1$ ir $\bar{H}_2$ atveju $P$ reikšmės yra $pv = 1 - F_{D^+_n}(D^+_n)$ ir $pv = 1 - F_{D^-_n}(D^-_n)$. Remiantis empirinio proceso invariantiškumo principu (B priedas, 8.4.1 teorema)

$$\sqrt{n} D^+_n \xrightarrow{d} \sup_{0 \le t \le 1} B(t).$$

Pagal 4 Brauno tilto savybę (B priedas, 8.5. skyrelis) gauname, kad su visais $x > 0$

$$\mathbf{P}\{\sqrt{n} D^+_n \le x\} \to K^+(x) = 1 - P_1(x) = 1 - e^{-2x^2}, \quad \text{kai} \quad n \to \infty, \tag{4.2.10}$$

todėl alternatyvų $\bar{H}_1$ ir $\bar{H}_2$ atveju asimptotinės $P$ reikšmės yra $pv_a = 1 - K^+(\sqrt{n} D^+_n)$ ir $pv_a = 1 - K^+(-\sqrt{n} D^-_n)$.

**4.2.1 pavyzdys.** Cheminė medžiaga sufasuota pakeliais, kurių kiekvieno masė turėtų būti lygi 1 kg. Reikia patikrinti hipotezę, kad pakelių masė yra pasiskirsčiusi pagal normalųjį dėsnį $N(\mu, \sigma^2)$, kai vidurkis $\mu = 1$ kg ir vidutinis kvadratinis nuokrypis $\sigma = 25$ g. Atsitiktinai atrinktų 20 pakelių masės surašytos lentelėje didėjimo tvarka.

**3.2.1 lentelė.** Statistiniai duomenys

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $X_i$ | 0,9473 | 0,9655 | 0,9703 | 0,9757 | 0,9775 | 0,9788 | 0,9861 |
| $i$ | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| $X_i$ | 0,9887 | 0,9964 | 0,9974 | 1,0002 | 1,0016 | 1,0077 | 1,0084 |
| $i$ | 15 | 16 | 17 | 18 | 19 | 20 | |
| $X_i$ | 1,0102 | 1,0132 | 1,0182 | 1,0225 | 1,0248 | 1,0306 | |

Randame $D_n = 0,1106$. Parinkime reikšmingumo lygmenį $\alpha = 0,05$. Kadangi $D_{0,05}(20) = 0,2941$ (žr. [17], IX lentelę), atmesti hipotezę nėra pagrindo. Asimptotinė $P$ reikšmė $pv_a = 1 - K(\sqrt{n} D_n) = 0,9673$.

## 4.3. Kramero ir Mizeso bei Anderseno ir Darlingo kriterijai

Turėjome, kad $\omega^2$, Kramero ir Mizeso, Anderseno ir Darlingo kriterijai paprastajai suderinamumo hipotezei $H_0 : F(x) \equiv F_0(x)$ tikrinti grindžiami statistikomis

$$\omega^2_n = \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 \psi(F_0(x)) dF_0(x) = \int_0^1 (\hat{G}_n(y) - y)^2 \psi(y) dy,$$

$$C_n = \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 dF_0(x) = \int_0^1 (\hat{G}_n(y) - y)^2 dy,$$

$$A_n = \int_{-\infty}^{\infty} \frac{(\hat{F}_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x) = \int_0^1 \frac{(\hat{G}_n(y) - y)^2}{y(1-y)} dy, \qquad (4.3.1)$$

kur $\psi(t)$ yra neneigiama funkcija apibrėžta intervale $[0, 1]$.

Nuokrypis nuo hipotezės matuojamas kvadratine metrika su svoriu:

$$\bar{H} : \int_{-\infty}^{\infty} (F(x) - F_0(x))^2 \psi(F_0(x)) dF_0(x) > 0. \qquad (4.3.2)$$

Kramero ir Mizeso bei Anderseno ir Darlingo kriterijų alternatyvos gaunamos imant atitinkamai $\psi(y) = 1$ bei $\psi(y) = 1/(y(1-y))$.

**$\omega^2$, Kramero ir Mizeso, Anderseno ir Darlingo statistikų realizacijų skaičiavimas.** Tarkime funkcijos $\psi(y) = 1$, $y\psi(y)$, $y^2\psi(y)$ yra integruojamos intervale $[0, 1]$. Pažymėkime $Y_i = F_0(X_{(i)})$, $i = 1, \ldots, n$, $Y_{(0)} = 0$, $Y_{(n+1)} = 1$,

$$g(t) = \int_0^t x\psi(x)dx, \quad h(t) = \int_0^t \psi(x)dx, \quad k = \int_0^1 (1-t)^2 \psi(t)dt.$$

**4.3.1 teorema.** *$\omega^2$, Kramero ir Mizeso bei Anderseno ir Darlingo statistikos gali būti užrašytos tokiu pavidalu:*

$$\omega_n^2 = \frac{2}{n} \sum_{i=1}^{n} [g(Y_{(i)}) - \frac{2i-1}{2n} h(Y_{(i)})] + k, \qquad (4.3.3)$$

$$nC_n = \frac{1}{12n} + \sum_{i=1}^{n} \left( Y_{(i)} - \frac{2i-1}{2n} \right)^2. \qquad (4.3.4)$$

$$nA_n = -n - \frac{1}{n} \left[ \sum_{i=1}^{n} (2i-1)[\ln Y_{(i)} + \ln(1 - Y_{(n-i+1)})] \right]. \qquad (4.3.5)$$

**Įrodymas.** Remdamiesi (4.3.1) ir įvestais žymėjimais, gauname

$$\omega_n^2 = \sum_{i=0}^{n} \int_{Y_{(i)}}^{Y_{(i+1)}} (\frac{i}{n} - y)^2 \psi(y)dy = \sum_{i=0}^{n} \frac{i^2}{n^2} [h(Y_{(i+1)}) - h(Y_{(i)})]$$

$$-2 \sum_{i=0}^{n} \frac{i}{n} [g(Y_{(i+1)}) - g(Y_{(i)})] + \int_0^1 y^2 \psi(y)dy$$

$$= \sum_{i=1}^{n} \frac{(i-1)^2 - i^2}{n^2} h(Y_{(i)}) + h(1) + 2 \sum_{i=1}^{n} \frac{i}{n} g(Y_{(i)}) - 2g(1)$$

$$+ \int_0^1 y^2 \psi(y)dy = \frac{2}{n} \sum_{i=1}^{n} [g(Y_{(i)}) - \frac{2i-1}{2n} h(Y_{(i)})] + k.$$

Jeigu $\psi(t) \equiv 1$, tai (4.3.1) yra Kramero ir Mizeso statistika. Randame

$$g(t) = \frac{t^2}{2}, \quad h(t) = t, \quad k = \frac{1}{3},$$

$$C_n = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{(i)}^2 - \frac{2i-1}{n}Y_{(i)}\right)+\frac{1}{3} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{(i)}^2 - 2\frac{2i-1}{2n}Y_{(i)} + \left(\frac{2i-1}{2n}\right)^2\right)$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{2i-1}{2n}\right)^2 + \frac{1}{3} = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left(Y_{(i)} - \frac{2i-1}{2n}\right)^2.$$

Jeigu $\psi(t) = 1/(t(1-t))$, tai funkcijos $\psi(t),\ t\psi(t),\ t^2\psi(t)$ nėra integruojamos intervale $[0,\ 1]$, todėl Anderseno ir Darlingo statistikos $A_n$ apibrėžti tiesiogiai pagal (4.3.3) negalime. Fiksuokime $0 < \varepsilon < Y_{(1)}$, $0 < \delta < 1-Y_{(n)}$ ir apibrėžkime statistiką $A_n$ kaip ribą

$$A_n = \lim_{\varepsilon,\delta\to 0}\int_{\varepsilon}^{1-\delta}(\hat{G}_n(y) - y)^2\frac{dy}{y(1-y)}.$$

Tada (4.3.3) lygybėje $g(t), h(t)$ ir $k$ reikia pakeisti į

$$g(t;\varepsilon,\delta) = \ln(1-\varepsilon) - \ln(1-t), \quad h(t;\varepsilon,\delta) = \ln t - \ln\varepsilon + \ln(1-\varepsilon) - \ln(1-t),$$

ir

$$k(\varepsilon,\delta) = \ln(1-\delta) - \ln\varepsilon - 1 + \delta + \varepsilon, \quad \varepsilon \le t \le 1-\delta.$$

Gauname

$$A_n = -\frac{2}{n}\sum_{i=1}^{n}\left[\frac{2i-1}{2n}\ln Y_{(i)} + (1-\frac{2i-1}{2n})\ln(1-Y_{(i)})\right]$$

$$+\lim_{\varepsilon,\delta\to 0}\{\ln(1-\delta) - \ln(1-\varepsilon) - 1 + \delta + \varepsilon\};$$

$$nA_n = -n - 2\sum_{i=1}^{n}\left[\frac{2i-1}{2n}\ln Y_{(i)} + (1-\frac{2i-1}{2n})\ln(1-Y_{(i)})\right]$$

$$= -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln Y_{(i)} + \ln(1-Y_{(n-i+1)})\right].$$

▲

**Kramero ir Mizeso bei Anderseno ir Darlingo kriterijai:** hipotezė $H_0$ atmetama, kai statistikos $nC_n$ arba $nA_n$ viršija jų atitinkamas kritines reikšmes.

Nedideliems $n$ daugelyje matematinės statistikos programų paketų (pvz., SAS) yra numatytas šių statistikų $P$ reikšmių, t. y. tikimybių, kad atitinkama statistika viršys stebėtąją realizaciją esant teisingai hipotezei, radimas. Palyginę $P$ reikšmę su pasirinktu reikšmingumo lygmeniu ir priimame sprendimą apie tikrinamąją hipotezę.

Kai $n$ yra didelis, kritinės reikšmės gaunamos naudojant asimptotinius skirstinius. Pagal (4.1.6)

$$nC_n \xrightarrow{d} C = \int_0^1 B^2(t)dt, \quad nA_n \xrightarrow{d} A = \int_0^1 \frac{B^2(t)}{t(1-t)}dt.$$

Remdamiesi tuo faktu, kad

$$B(x) = \frac{\sqrt{2}}{\pi}\sum_{k=1}^{\infty}\frac{\sin \pi k x}{k}Z_k;$$

čia $Z_1, Z_2, ...,$ yra vienodai pasiskirstę n. a. d., turintys standartinį normalųjį skirstinį, $Z_i \sim N(0,1)$, gauname

$$C = \frac{1}{\pi^2}\sum_{k=1}^{\infty}\frac{Z_k^2}{k^2}, \quad A = \sum_{k=1}^{\infty}\frac{Z_k^2}{k(k+1)}.$$

Atsitiktinių dydžių $C$ ir $A$ pasiskirstymo funkcijų išraiškos yra tokios (žr. [7]:

$$\mathbf{P}\{C \le x\} = a_1(x) = 1 - \frac{1}{\pi}\sum_{j=1}^{\infty}(-1)^{j+1}\int_{(2j-1)^2\pi^2}^{4j^2\pi^2}\sqrt{\frac{-\sqrt{y}}{\sin(\sqrt{y})}}\frac{e^{-xy/2}}{y}dy. \quad (4.3.6)$$

$$\mathbf{P}\{A \le x\} = a_2(x) =$$

$$\frac{\sqrt{2\pi}}{x}\sum_{j=0}^{\infty}(-1)^j\frac{j\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\int_0^{\infty}\exp\{\frac{x}{8(y^2+1)} - \frac{(4j+1)^2\pi^2(1+y^2)}{8x}\}dy. \quad (4.3.7)$$

**4.3.1 pavyzdys** (**4.2.1** pavyzdžio tęsinys). Pavyzdyje **4.2.1** suformuluotą uždavinį išspręsime taikydami Kramero ir Mizeso bei Anderseno ir Darlingo kriterijus. Gauname

$$nC_n = \frac{1}{240} + \sum_{i=1}^{20}(\Phi(z_i) - \frac{2i-1}{40})^2 = 0,0526, \quad z_i = \frac{X_i - 1}{0,025},$$

$$nA_n = -20 - 2\sum_{i=1}^{20}\left(\frac{2i-1}{40}\ln\Phi(z_i) + (1 - \frac{2i-1}{40})\ln(1-\Phi(z_i))\right) = 0,3971.$$

Asimptotinės $P$ reikšmės yra atitinkamai

$$\mathbf{P}\{nC_n > 0,0526\} \approx 1 - a_1(0,0526) \approx 0,86,$$

ir

$$\mathbf{P}\{nA_n > 0,3971\} \approx 1 - a_2(0,3971) \approx 0,85.$$

Anderseno ir Darlingo bei Kramero ir Mizeso kriterijų $P$ reikšmės yra beveik vienodos, tačiau gerokai skiriasi nuo Kolmogorovo ir Smirnovo kriterijaus $P$ reikšmės. Remdamiesi visais trimis ktriterijais darome tą pačią išvadą: atmesti hipotezę nėra pagrindo.

**4.3.1 pastaba.** Lemeško darbuose [18], [19], [20] skaitiniais metodais atliktas pateiktų kriterijų galios funkcijų palyginimas. Apskritai kalbant, kriterijai išrikiuoti tokia tvarka: *Pirsono chi kvadrato* $\succ$ *Anderseno ir Darlingo* $\succ$ *Kramero ir Mizeso* $\succ$ *Kolmogorovo ir Smirnovo*.

## 4.4. Modifikuotieji kriterijai

Praktiškai kur kas dažniau tenka tikrinti sudėtines suderinamumo hipotezes.

**Sudėtinė suderinamumo hipotezė:**

$$X \sim F(x) \in \mathcal{F}_0 = \{F_0(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbf{R}^m\},$$

čia $F_0(x;\boldsymbol{\theta})$ žinomo pavidalo pasiskirstymo funkcija, priklausanti nuo nežinomo baigtinės dimensijos parametro $\boldsymbol{\theta}$.

**Kriterijaus sudarymo idėja.** Kriterijų statistikos apibrėžiamos pagal analogiją su paprastosios hipotezės atveju. Modifikuotieji Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijai grindžiami statistikomis

$$D_n^{(mod)} = \sup_{y\in[0,\,1]} |\hat{G}_n(y) - y|, \quad C_n^{(mod)} = \int_0^1 (\hat{G}_n(y) - y)^2 dy,$$

$$A_n^{(mod)} = \int_0^1 \frac{(\hat{G}_n(y) - y)^2}{y(1-y)} dy, \tag{4.4.1}$$

čia

$$\hat{G}_n(y) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty,y]}(Y_i) \tag{4.4.2}$$

yra empirinė pasiskirstymo funkcija, sudaryta remiantis a. d. $Y_i = F_0(X_i,\hat{\boldsymbol{\theta}})$; čia $\hat{\boldsymbol{\theta}}$ yra DT (ar kitoks) parametro įvertinys.

Nors modifikuotųjų kriterijų statistikos turi lygiai tokį pat pavidalą kaip ir statistikos $D_n, C_n, A_n$ paprastosios hipotezės atveju, tačiau jų skirstiniai yra ne tie patys. Atsitiktiniai dydžiai $Y_i = F_0(X_i,\hat{\boldsymbol{\theta}}), i = 1,...,n$, nėra vienodai pasiskirstę n. a. d., turintys tolygųjį pasiskirstymą $U(0,\,1)$. Šie a. d. yra priklausomi, nes visi jie apibrėžiami naudojant įvertinį $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1,...,X_n)$. Todėl minėtų statistikų skirstiniai, apskritai kalbant, turėtų priklausyti ir nuo pasiskirstymo funkcijos $F_0$ pavidalo, ir nuo parametro $\boldsymbol{\theta}$.

Tam tikrais specialiais atvejais modifikuotųjų statistikų skirstiniai esant teisingai hipotezei nuo nežinomo parametro nepriklauso, o priklauso tik nuo pasiskirstymo funkcijos $F_0$.

**4.4.1 teorema.** *Tarkime, parametras vertinamas DT metodu. Tada tikimybinių šeimų, priklausančių tik nuo poslinkio ir mastelio parametrų*

$$\{F_0(x;\boldsymbol{\theta}) = G((x-\mu)/\sigma), \quad \mu \in \mathbf{R}, \quad \sigma > 0\},$$

*arba priklausančių tik nuo laipsnio ir mastelio parametrų*

$$\{F_0(x;\boldsymbol{\theta}) = G((\frac{x}{\theta})^\nu), \quad \theta > 0, \quad \nu > 0\}$$

*atvejais, modifikuotųjų statistikų skirstiniai nepriklauso nuo nežinomų parametrų, bet gali būti skirtingi skirtingoms funkcijoms G.*

**Įrodymas.** Pakartoję 3.4.1 teoremos įrodymą gauname, kad a. d. $Y_i$ skirstinys nepriklauso nuo nežinomų parametrų. Taigi ir $\hat{G}_n(y)$ bei modifikuotųjų

Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijų statistikų $D_n^{(mod)}$, $C_n^{(mod)}$, $A_n^{(mod)}$ skirstiniai nuo nežinomų parametrų nepriklauso, tačiau gali priklausyti nuo funkcijos $G$ pavidalo.

Skirstinių šeimos, priklausančios tik nuo laipsnio ir mastelio parametrų, nagrinėjamos analogiškai. ▲

**4.4.1 pastaba.** Remiantis 4.4.1 teorema galima sukonstruoti modifikuotuosius suderinamumo kriterijus dėl stebimojo a. d. skirstinio priklausymo eksponentinių, normaliųjų, logistinių, ekstremalių reikšmių, Koši skirstinių šeimoms (priklauso tik nuo poslinkio ir mastelio parametrų), bei lognormaliųjų, loglogistinių, Veibulo skirstinių šeimoms (priklauso tik nuo laipsnio ir mastelio parametrų). Reikia pažymėti, kad pastarosioms šeimoms atskiri kriterijai nereikalingi, nes, perėjus prie a. d. $\ln X_i$, jos suvedamos į skirstinių šeimas, priklausančias tik nuo poslinkio ir mastelio parametrų.

**4.4.2 pastaba.** Netgi kai $n$ yra didelis, modifikuotųjų kriterijų kritinės reikšmės gali gerokai skirtis nuo kritinių reikšmių, gaunamų tikrinant paprastąją hipotezę. Pavyzdžiui, modifikuotojo Kolmogorovo ir Smirnovo kriterijaus (normalusis skirstinys) $0,01$ kritinė reikšmė atitinka to paties kriterijaus tikrinant paprastąją hipotezę $0,2$ kritinę reikšmę. Daugelyje matematinės statistikos knygų ir kai kuriuose programų paketuose kriterijai sudėtinei suderinamumo hipotezei tikrinti nėra korektiški: naudojamos modifikuotosios statistikos, o kritinės reikšmės (arba $P$ reikšmės) imamos tos, kurios gaunamos tikrinant paprastąją hipotezę. Gaunamos išvados gali būti klaidingos: remdamiesi tokiu nekorektišku kriterijumi galime gauti išvadą, kad duomenys neprieštarauja iškeltai hipotezei, o remiantis modifikuotuoju kriterijumi hipotezę reikėtų atmesti.

Matematinės statistikos programų pakete SAS paprastosios ir sudėtinės hipotezių atvejai yra atskirti ir pateikiamos modifikuotųjų kriterijų $P$ reikšmės dažniausiai naudojamų skirstinių šeimoms. Modifikuotųjų kriterijų kritinių reikšmių lentelės pateiktos knygoje [11].

**Modifikuotieji Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijai:** sudėtinė suderinamumo hipotezė $H_0$ atmetama reikšmingumo lygmens kriterijumi, kai statistikos $D_n^{(mod)}, C_n^{(mod)}, A_n^{(mod)}$ viršija atitinkamas šių statistikų $\alpha$ kritines reikšmes.

Kai imtis yra didelė, asimptotinės P reikšmės randamos remiantis modifikuotųjų kriterijų statistikų ribiniais skirstiniais.

**Modifikuotųjų kriterijų statistikų asimptotiniai skirstiniai**

Modifikuotųjų kriterijų statistikos yra funkcionalai nuo empirinio proceso

$$\xi_n(y) = \sqrt{n}(\hat{G}_n(y) - y), \quad y \in (0,1), \tag{4.4.3}$$

todėl reikia žinoti empirinio proceso $\xi_n(y)$ asimptotinį elgesį.

Įrodyta, kad poslinkio ir mastelio bei laipsnio ir mastelio šeimų atveju $\xi_n(y)$ asimptotinis skirstinys yra toks [22]: $\xi_n(y) \xrightarrow{d} \xi(y)$, $y \in (0,\ 1$, čia $\xi(y)$ yra Gauso procesas su nuliniu vidurkiu ir koreliacine funkcija

$$K(x,y) = x \wedge y - xy - \frac{1}{a} K_1(x,y).$$

Lyginant su asimptotiniu skirstiniu paprastosios hipotezės atveju skirtumas yra tas, kad koreliacinėje funkcijoje atsiranda pataisa (trečiasis dėmuo), priklausanti nuo funkcijos $G$. Funkcijos $K_1(x,y)$ pavidalas yra toks:

$$K_1(x,y) = c_2 w_1(x) w_2(y) + c_1 w_2(x) w_2(y) - c_3 [w_1(x) w_2(y) + w_2(x) w_1(y)];$$

čia

a) poslinkio ir mastelio parametrų šeimoms

$$w_1(x) = g(G^{-1}(x)), \quad w_2(x) = G^{-1}(x) g(G^{-1}(x)),$$

$$c_1 = \int_{-\infty}^{\infty} \frac{(g'(x))^2}{g(x)} dx, \quad c_2 = \int_{-\infty}^{\infty} x^2 \frac{(g'(x))^2}{g(x)} dx - 1,$$

$$c_3 = \int_{-\infty}^{\infty} x \frac{(g'(x))^2}{g(x)} dx, \quad a = c_1 c_2 - c_3^2;$$

b) laipsnio ir mastelio parametrų šeimoms

$$w_1(x) = G^{-1}(x) g(G^{-1}(x)), \quad w_2(x) = G^{-1}(x) g(G^{-1}(x)) \ln G^{-1}(x),$$

$$c_1 = \int_{-\infty}^{\infty} \left(1 + x \frac{g'(x)}{g(x)}\right)^2 g(x) dx, \quad c_2 = \int_{-\infty}^{\infty} \left(1 + x \ln x \, \frac{g'(x)}{g(x)} + \ln x\right)^2 g(x) dx,$$

$$c_3 = \int_{-\infty}^{\infty} \left(1 + x \frac{g'(x)}{g(x)}\right) \left(1 + x \ln x \, \frac{g'(x)}{g(x)} + \ln x\right) g(x) dx, \quad a = c_1 c_2 - c_3^2.$$

Remiantis invariantiškumo principu (žr. B priedą, 8.4.1 teoremą) modifikuotųjų Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijų statistikos turi tokias ribas:

$$\sqrt{n} D_n^{(mod)} \xrightarrow{d} \sup_{0 \le t \le 1} |\xi(t)|, \quad n C_n^{(mod)} \xrightarrow{d} \int_0^1 \xi^2(t) dt, \quad n A_n^{(mod)} \xrightarrow{d} \int_0^1 \frac{\xi^2(t)}{t(1-t)} dt. \tag{4.4.4}$$

**4.4.1 pavyzdys.** (**2.3.2** pavyzdžio tęsinys.) Pagal **2.3.2** pratimo duomenis patikrinsime hipotezę, kad stebimo a. d. skirstinys yra a)normalusis; b) lognormalusis.

a) Apskaičiuojame modifikuotųjų Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijų statistikų reikšmes:

$$\bar{X} = 12,0184, \quad s = 10,03248, \quad D_{49}^{+} = 0,1806, \quad D_{49}^{-} = 0,1296, \quad D_{49}^{(mod)} = 0,1806;$$

$$n C_n^{(mod)} = \frac{1}{588} + \sum_{i=1}^{49} (\Phi(Y_{(i)}) - \frac{2i-1}{98})^2 = 0,3241, \ \ Y_{(i)} = \frac{X_i - 12,0184}{10,03248};$$

$$n A_n^{(mod)} = -49 - 2 \sum_{i=1}^{49} \left( \frac{2i-1}{49} \ln \Phi(Y_{(i)}) + (1 - \frac{2i-1}{49}) \ln(1 - \Phi(Y_{(i)})) \right) = 1,8994.$$

Atlikdami skaičiavimus SAS programų paketu gauname atitinkamas P reikšmes

$$pv < 0,01, \quad pv < 0,005, \quad pv < 0,005.$$

Normalumo hipotezė atmetama.

Jeigu atliekant skaičiavimus tartume, kad buvo tikrinama paprasta normalumo hipotezė $H_0 : X \sim N(12,0184,\ 10,03248)$, kai nežinomi parametrai pakeisti jų įverčiais, tai atitinkamai gautume tokias P reikšmes: $0,0724; 0,1193; 0,1049$. Normalumo hipotezė neatmetama, jeigu kriterijaus reikšmingumo lygmuo mažesnis už $0,0724$. Reikia pažymėti, kad kai kuriuose matematinės statistikos paketuose sudėtinės suderinamumo hipotezės tikrinamos būtent tokiu būdu. Pavyzdžiui, pagal šiuos duomenis tikrindami normalumo hipotezę Kolmogorovo ir Smirnovo kriterijumi SPSS paketu gauname P reikšmę $0,0724$.

b) Perėję prie logaritmų gauname tokias modifikuotųjų Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijų statistikų reikšmes:

$$\bar{X} = 2,1029, \quad s = 0,9675, D_{49}^{(mod)} = 0,1033,$$

$$nC_n^{(mod)} = 0,0793, \quad nA_n^{(mod)} = 0,5505.$$

Atlikdami skaičiavimus SAS programų paketu gauname atitinkamas P reikšmes:

$$pv = 0,6723, \quad pv = 0,2141, \quad pv = 0,1517.$$

Turimi duomenys neprieštarauja prielaidai, kad buvo stebėtas lognormalusis a. d.

**4.4.3 pastaba.** Lemeško ir kt. darbuose [18], [19], [20] skaitiniais metodais atlikta sudėtinių suderinamumo hipotezių tikrinimo kriterijų galios priklausomybės nuo alternatyvų analizė. Pakankamai plačiai alternatyvų klasei daroma išvada, kad kriterijus galima išrikiuoti tokia tvarka: *modifikuotas Anderseno ir Darlingo* $\succ$ *modifikuotas chi kvadrato* $\succ$ *modifikuotas Kramero ir Mizeso* $\succ$ *Pirsono chi kvadrato* $\succ$ *modifikuotas Kolmogorovo ir Smirnovo.*

## 4.5. Dviejų imčių kriterijai

### 4.5.1. Dviejų imčių Kolmogorovo ir Smirnovo kriterijus

Tarkime, turime dvi nepriklausomas paprastąsias imtis

$$\mathbf{X} = (X_1, ..., X_m)^T \quad \text{ir} \quad \boldsymbol{Y} = (Y_1, ..., Y_n)^T,$$

gautas stebint absoliučiai tolydžius a. d. $X$ ir $Y$ su pasiskirstymo funkcijomis $F_1$ ir $F_2$. Pažymėkime $\hat{F}_{1m}(x)$ ir $\hat{F}_{2n}(x)$ empirines pasiskirstymo funkcijas, sudarytas remiantis imtimis $\mathbf{X}$ ir $\boldsymbol{Y}$.

Reikia patikrinti homogeniškumo hipotezę

$$H_0 : F_1(x) = F_2(x), \ \ \forall x \in \mathbf{R}, \tag{4.5.1}$$

kai alternatyva yra dvipusė

$$\bar{H} : \sup_{x \in \mathbf{R}} |F_1(x) - F_2(x)| > 0, \tag{4.5.2}$$

arba vienpusė

$$\bar{H}^+ : \sup_{x \in \mathbf{R}} (F_1(x) - F_2(x)) > 0, \quad \bar{H}^- : \inf_{x \in \mathbf{R}} (F_1(x) - F_2(x)) < 0. \tag{4.5.3}$$

*Dviejų imčių Kolmogorovo ir Smirnovo kriterijus* hipotezei $H_0$, kai alternatyva dvipusė $\bar{H}$, tikrinti grindžiamas statistika

$$D_{m,n} = \sup_{|x|<\infty} |\hat{F}_{1m}(x) - \hat{F}_{2n}(x)|. \tag{4.5.4}$$

Analogiškai 4.1.1 teoremai įsitikiname, kad esant teisingai hipotezei $H_0$ statistikos $D_{m,n}$ skirstinys nepriklauso nuo stebimų a. d. pasiskirstymo funkcijos, o priklauso tik nuo imčių didumų $m$ ir $n$:

$$D_{m,n} = \sup_{x\in\mathbf{R}} |\hat{F}_{1m}(x) - \hat{F}_{2n}(x)| = \sup_{0\leq y\leq 1} |\hat{G}_{1m}(y) - \hat{G}_{2n}(y)|,$$

čia $\hat{G}_{1m}$ ir $\hat{G}_{2n}$ yra empirinės pasiskirstymo funkcijos, sukonstruotos pagal imtis

$$(U_{11}, ..., U_{1m})^T \quad \text{ir} \quad (U_{21}, ..., U_{2n})^T,$$

gautas stebint nepriklausomus tolygiai pasiskirsčiusius intervale $[0,\ 1]$ atsitiktinius dydžius $U_1$ ir $U_2$

$$U_{1i} = F_1(X_i), \quad U_{2j} = F_2(Y_j).$$

Kai alternatyvos vienpusės, kriterijai grindžiami statistikomis

$$D^+_{m,n} = \sup_{x\in\mathbf{R}}(\hat{F}_{1m}(x) - \hat{F}_{2n}(x)),\ D^-_{m,n} = -\inf_{x\in\mathbf{R}}(\hat{F}_{1m}(x) - \hat{F}_{2n}(x)). \tag{4.5.5}$$

Kadangi funkcijos $\hat{F}_{1m}(x)$ ir $\hat{F}_{2n}(x)$ yra laiptinės, tai supremumas pasiekiamas šių funkcijų šuoliukų taškuose. Analogiškai vienos imties Kolmogorovo ir Smirnovo statistikos atvejui $D_{m,n}$ gali būti apskaičiuota šitaip:

$$D_{m,n} = \max(D^+_{m,n},\ D^-_{m,n});$$

$$D^+_{m,n} = \max_{1\leq r\leq m}\left(\frac{r}{m} - \hat{F}_{2n}(X_{(r)})\right) = \max_{1\leq s\leq n}\left(\hat{F}_{1m}(Y_{(s)}) - \frac{s-1}{n})\right),$$

$$D^-_{m,n} = \max_{1\leq r\leq m}\left(\hat{F}_{2n}(X_{(r)}) - \frac{r-1}{m}\right) = \max_{1\leq s\leq n}\left(\frac{s}{n} - \hat{F}_{1m}(Y_{(s)})\right).$$

**Dviejų imčių Kolmogorovo ir Smirnovo kriterijus:** hipotezė $H_0$ atmetama lygmens $\alpha$ kriterijumi, kai

$$D_{m,n} \geq D_\alpha(m,\ n);$$

čia $D_\alpha(m,\ n)$ yra statistikos $D_{m,n}$ lygmens $\alpha$ kritinė reikšmė, t. y.

$$\mathbf{P}\{D_{m,n} \geq D_\alpha(m,\ n)\} \leq \alpha,\ \ \mathbf{P}\{D_{m,n} < D_\alpha(m,\ n)\} > 1-\alpha.$$

**4.5.1 pastaba.** Statistika $D_{m,n}$ įgyja reikšmes pavidalo $l/k$; čia $k = k(m,\ n)$ yra skaičių $m$ ir $n$ bendras mažiausias kartotinis, o $l$ – sveikasis skaičius. Todėl paprastai reikšmingumo lygmens $\alpha$ kriterijus bus randomizuotas. Dažniausiai naudojami nerandomizuoti kriterijai gaunami šiek tiek sumažinus reikšmingumo lygmenį $\alpha$. Tiksliau, kritine reikšme $D_\alpha(m,\ n)$ imamas mažiausias pavidalo $l/k$ skaičius, kuris tenkina sąlygą

$$\mathbf{P}\{D_{m,\,n} > D_\alpha(m,\ n)|H\} = \alpha' \leq \alpha.$$

Apytiksles $D_\alpha(m,\ n)$ reikšmes galima rasti naudodami aproksimaciją (žr. [7])

$$D_\alpha(m,n) \approx \frac{1}{k(m,n)} + D_\alpha(\nu) + \frac{1}{\nu} -$$
$$\frac{1}{\nu}\left[\frac{n-m}{6(m+n)} + \frac{1}{2}\frac{m-d(m,n)}{m+n+d(m,n)}\right],$$

čia $k(m,n)$ ir $d(m,n)$ yra skaičių $m$ ir $n$ bendras mažiausias kartotinis ir bendras didžiausias daliklis, $\nu = mn/(m+n)$, $m \leq n$.

Kai alternatyvos vienpusės, hipotezė $H_0$ atmetama, kai

$$D^+_{m,n} > D_{2\alpha}(m,n) \text{ arba } D^-_{m,n} > D_{2\alpha}(m,n). \tag{4.5.6}$$

Pastarieji kriterijai yra apytikslūs. Tačiau, kai reikšmingumo lygmuo $\alpha < 0,1$, aproksimavimo tikslumas praktiškai pakankamas (žr. [7]).

Kai $m$ ir $n$ maži, kritinės reikšmės yra tabuliuotos (žr. [7], [17]) arba jų skaičiavimas numatytas matematinės statistikos programų paketuose.

Kai $m$ ir $n$ dideli, asimptotinės kritinės reikšmės randamos naudojant asimptotinius statistikų $D^+_{m,n}$ ir $D_{m,n}$ skirstinius.

**4.5.1 teorema.** *Tarkime, kad $m/(m+n) \to p \in (0,\,1)$, $m,n \to \infty$. Jeigu hipotezė $H_0$ teisinga, tai*

$$\mathbf{P}\left\{\sqrt{\frac{mn}{m+n}}D_{m,n} \leq x\right\} \to 1 - 2\sum_{n=1}^{\infty}(-1)^{n-1}e^{-2n^2x^2}, \tag{4.5.7}$$

$$\mathbf{P}\left\{\sqrt{\frac{mn}{m+n}}D^+_{m,n} \leq x\right\} \to 1 - e^{-2x^2}, \quad \textit{kai} \quad n \to \infty. \tag{4.5.8}$$

**Įrodymas.** Kai hipotezė $H_0$ teisinga, remdamiesi B priedo 8.4.1 teorema gauname

$$\sqrt{\frac{mn}{m+n}}(\hat{F}_{1m} - \hat{F}_{2n}) =$$
$$\sqrt{\frac{n}{m+n}}\sqrt{m}\,[\hat{F}_{1m} - F_1] - \sqrt{\frac{m}{m+n}}\sqrt{n}\,[\hat{F}_{2n} - F_2] \xrightarrow{d} B =$$
$$\sqrt{1-p}\,B_1 - \sqrt{p}\,B_2;$$

čia $B_1$ ir $B_2$ yra nepriklausomi Brauno tiltai. Stochastinis procesas $B$ yra Gauso, nes jis yra tiesinė Gauso procesų funkcija. Kadangi

$$\mathbf{E}B(t) = 0, \quad \mathbf{cov}(B(s), B(t)) = (1-\gamma)\mathbf{cov}(B_1(s), B_1(t)) +$$

$$\gamma \mathbf{cov}(B_2(s), B_2(t)) = s \wedge t - st,$$

tai $B$ taip pat yra Brauno tiltas. Taigi

$$\sqrt{\frac{mn}{m+n}} D_{m,n} \xrightarrow{d} \sup_{0 \le t \le 1} |B(t)|, \quad \sqrt{\frac{mn}{m+n}} D^+_{m,n} \xrightarrow{d} \sup_{0 \le t \le 1} B(t).$$

▲

Remdamiesi šia teorema gauname

$$pv_a = 1 - \mathbf{P}\{\sup_{0 \le t \le 1} |B(t)| < \sqrt{\frac{mn}{m+n}} D_{m,n}\} \xrightarrow{d} 1 - K\left(\sqrt{\frac{mn}{m+n}} D_{m,n}\right),$$

čia $K(x)$ yra Kolmogorovo pasiskirstymo funkcija. Tikslesnes aproksimacijas galima rasti [7].

**4.5.1 pavyzdys.** Tiriamas fungicidų poveikis kavos medelių sergamumui. Lentelėje pateikti duomenys apie kavos medelių sergamumą (procentais) naudojant fungicidus ir jų nenaudojant.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fungicidai naudoti | $6,01$ | $2,48$ | $1,76$ | $5,10$ | $0,75$ | $7,13$ | $4,88$ |
| Fungicidai nenaudoti | $5,68$ | $5,68$ | $16,30$ | $21,46$ | $11,63$ | $44,20$ | $33,30$ |

Tikriname hipotezę, kad fungicidų naudojimas neturi įtakos kavos medelių sergamumui.

Tarpinius skaičiavimo rezultatus pateikiame lentelėje:

| $r$ | $X_{(r)}$ | $Y_{(r)}$ | $\frac{r}{n}$ | $\hat{F}_n(Y_{(r)})$ | $\hat{G}_n(X_{(r)})$ | $\frac{r}{n} - \hat{F}_n(Y_{(r)})$ | $\frac{r}{m} - \hat{G}_n(X_{(r)})$ |
|---|---|---|---|---|---|---|---|
| 1 | $0,75$ | $5,68$ | $1/7$ | $5/7$ | $0$ | $-4/7$ | $1/7$ |
| 2 | $1,76$ | $5,68$ | $2/7$ | $5/7$ | $0$ | $-3/7$ | $2/7$ |
| 3 | $2,48$ | $11,63$ | $3/7$ | $1$ | $0$ | $-4/7$ | $3/7$ |
| 4 | $4,88$ | $16,30$ | $4/7$ | $1$ | $0$ | $-3/7$ | $4/7$ |
| 5 | $5,10$ | $21,46$ | $5/7$ | $1$ | $0$ | $-2/7$ | $5/7$ |
| 6 | $6,01$ | $33,30$ | $6/7$ | $1$ | $2/7$ | $-1/7$ | $4/7$ |
| 7 | $7,13$ | $44,20$ | $1$ | $1$ | $2/7$ | $0$ | $5/7$ |

Gauname

$$D^+_{m,n} = \max_{1 \le r \le m} \left(\frac{r}{m} - \hat{G}_n(X_{(r)})\right) = \frac{5}{7},$$

$$D^-_{m,n} = \max_{1 \le r \le n} \left(\frac{r}{n} - \hat{F}_m(Y_{(r)})\right) = 0, \quad D_{m,n} = \frac{5}{7} \approx 0,714286.$$

Atlikdami skaičiavimus SPSS paketu gauname P reikšmę

$$pv = \mathbf{P}\{D_{m,n} \ge 0,714286\} = 0,05303.$$

Asimptotinė $P$ reikšmė yra $pv_a = 0,05623$.

Hipotezė neatmetama, jeigu kriterijaus reikšmingumo lygmuo yra $0,05$, tačiau atmetama, jeigu reikšmingumo lygmuo yra $0,1$. Pagal tokias mažas imtis sunku daryti galutinę išvadą. Toliau šis pavyzdys bus nagrinėjamas taikant Kramero ir Mizeso bei Vilkoksono kriterijus.

## 4.5.2. Dviejų imčių Kramero ir Mizeso kriterijus

Dviejų imčių Kramero ir Mizeso kriterijus yra grindžiamas statistika (žr. [1])

$$T_{m,n} = \frac{mn}{m+n} \int_{-\infty}^{+\infty} (\hat{F}_{1m}(x) - \hat{F}_{2n}(x))^2 d\hat{G}_{m+n}(x), \tag{4.5.9}$$

čia
$$\hat{G}_{m+n}(x) = \frac{m}{m+n}\hat{F}_{1m}(x) + \frac{n}{m+n}\hat{F}_{2n}(x)$$
yra empirinė pasiskirstymo funkcija, sukonstruota pagal didumo $m+n$ jungtinę imtį $(X_1, ..., X_m, Y_1, ..., Y_n)^T$.

Remiantis apibrėžimu (4.5.9) galima įrodyti (žr. 4.3 pratimą), kad statistika $T_{m,n}$ gali būti užrašyta tokiu pavidalu:
$$T_{m,n} = \frac{1}{mn(m+n)}[m\sum_{j=1}^{m}(R_{1j}-j)^2 + n\sum_{i=1}^{n}(R_{2i}-i)^2] - \frac{4mn-1}{6(m+n)}, \quad (4.5.10)$$
čia $R_{1j}$ ir $R_{2i}$ yra pozicijos,kurias užėmė pirmosios ir antrosios imties elementai jungtinėje imtyje.

Įrodyta, kad statistika $T_{m,n}$ turi ribinį skirstinį, kai $m, n \to \infty$, $m/n \to \lambda$, $0 < \lambda < \infty$. Šis ribinis skirstinys sutampa su statistikos $nC_n$ ribiniu skirstiniu (4.3.6):
$$\mathbf{P}\{T_{m,n} \le x\} \to \mathbf{P}\{C \le x\} = a_1(x).$$

Statistikos $C$ vidurkis ir dispersija yra lygūs $1/6$ ir $1/45$ (žr. 4.1 pratimą), o statistikos $T_{m,n}$ atitinkami momentai yra (4.4 pratimas)
$$\mathbf{E}T_{m,n} = \frac{1}{6}(1+\frac{1}{m+n}), \quad \mathbf{V}T_{m,n} = \frac{1}{45}(1+\frac{1}{m+n})(\frac{m+n+1}{m+n} - \frac{3(m+n)}{4mn}).$$
Todėl vietoje $T_{m,n}$ rekomenduojama naudoti modifikuotą statistiką
$$T^*_{m,n} = \frac{T_{m,n} - \mathbf{E}T_{m,n}}{\sqrt{45\mathbf{V}T_{m,n}}} + \frac{1}{6}, \quad (4.5.11)$$
kurios pirmieji du momentai sutampa su atitinkamais a. d. $C$ momentais.

**Asimptotinis Kramero ir Mizeso dviejų imčių kriterijus**: homogeniškumo hipotezė atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai
$$T^*_{m,n} > t^*_\alpha(m,n); \quad (4.5.12)$$
čia $t^*_\alpha(m,n)$ yra kritinė reikšmė, randama iš sąlygos
$$1 - a_1(t^*_\alpha(m,n)) = \alpha.$$

Aproksimacija funkcija $a_1(x)$ yra gana tiksli ir su palyginti nedideliais $m, n$. Aproksimacijos tikslumo analizę galima rasti [7].

**4.5.2 pavyzdys.** ( **4.5.1** pavyzdžio tęsinys). Pritaikysime Kramero ir Mizeso kriterijų **4.5.1** pratimo duomenims. Gauname: $T_{7,7} = 0,7704$ ir modifikuotosios statistikos reikšmė yra $T^*_{7,7} = 0,7842$. Asimptotinė P reikšmė $pv_a = 1 - a_1(0,7842) \approx 0,008$. Homogeniškumo hipotezė atmestina. Šiame pavyzdyje Kramero ir Mizeso kriterijus pasirodė galingesnis už dviejų imčių Kolmogorovo ir Smirnovo kriterijų.

## 4.6. Pratimai

**4.1.** Raskite atsitiktinių dydžių $C = \int_0^1 B^2(t)dt$ ir $A = \int_0^1 B^2(t)/(t(1-t))dt$ pirmuosius du momentus.

**4.2.** Raskite statistikų $nC_n$ ir $nA_n$ pirmuosius du momentus ir palyginkite juos su **4.1** pratime gautais momentais.

**4.3.** Įrodykite, kad Kramero ir Mizeso dviejų imčių statistiką (4.5.9) galima užrašyti pavidalu (4.5.10).

**4.4.** Raskite statistikos $T_{m,n}$, apibrėžtos formulėmis (4.5.9), (4.5.10), pirmuosius du momentus.

**4.5.** Sukonstruokite tolydžiojo a. d. pasiskirstymo funkcijos $F(x)$ lygmens $Q$ pasikliovimo sritį pagal paprastąją didumo $n$ imtį.

**4.6.** Tarkime, $X$ yra diskretus a. d., kurio galimos reikšmės 0, 1, 2...., o jų įgijimo tikimybės $p_k = \mathbf{P}\{X = k\}, k = 0, 1, ....$ Įrodykite, kad a. d.

$$Z = \sum_{k=0}^{X-1} p_k + p_X Y$$

yra tolygiai pasiskirstęs intervale (0, 1), kai $Y \sim U(0,\ 1)$ ir nepriklauso nuo $X$.

**4.7** (**4.6** tęsinys). Tarkime, $X_1, ..., X_n$ yra paprastoji imtis diskrečiojo a. d. $X$, o $\hat{F}_n(x)$ – empirinė pasiskirstymo funkcija. Reikia patikrinti hipotezę $H : \mathbf{E}(\hat{F}_n(x)) = F_0(x), |x| < \infty$; čia $F_0(x)$ – visiškai nusakyta diskrečioji pasiskirstymo funkcija. Hipotezę $H$ pakeiskime hipoteze $H' : \mathbf{E}(\hat{G}_n(z)) = G(z), o < z < 1$. Čia $\hat{G}_n(z)$ – empirinė pasiskirstymo funkcija imties $Z_i = F_0(X_i) + p_{X_i} Y_k,\ \ i = 1, ..., n,\ \ k = 1, ..., n;$ $Y_1, ..., Y_n$ – nepriklausanti nuo $X_1, ..., X_n$ paprastoji a. d. $Y \sim U(0,\ 1)$ imtis, o $G(z)$ – tolygiojo skirstinio $U(0,\ 1)$ pasiskirstymo funkcija. Taip gauname randomizuotą, pavyzdžiui, Kolmogorovo ir Smirnovo kriterijaus analogą diskretiesiems skirstiniams.

**4.8.** (**4.7** tęsinys). Remdamiesi **4.7** pratime aptartu kriterijumi atlikite **2.5** pratimo užduotį.

**4.9.** (**4.7** tęsinys). Remdamiesi **4.7** pratime aptartu kriterijumi atlikite **2.7** pratimo užduotį.

**4.10.** Remdamiesi Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijais patikrinkite hipotezę, kad **2.16** pratimo imtis gauta stebint a) normalųjį a. d. su parametrais $\hat{\mu}$ ir $\hat{\sigma}^2$, b) normalųjį a. d.

**4.11.** Remdamiesi Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijais patikrinkite hipotezę, kad **2.17** pratimo imtis gauta stebint a) lognormalųjį a. d. su parametrais $\hat{\mu}$ ir $\hat{\sigma}$, b) lognormalųjį a. d.

**4.12.** Kontroliuojant staklių stabilumą, kiekvieną valandą paimama 20 gaminių ir remiantis jų tam tikro parametro matavimo rezultatais apskaičiuojamas nepaslinktasis dispersijos įvertinys $s^2$. Lentelėje pateikta 47 įvertinių realizacijos.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0,1225 | 0,1764 | 0,1024 | 0,1681 | 0,0841 | 0,0729 | 0,1444 | 0,0900 |
| 0,0961 | 0,1369 | 0,1521 | 0,1089 | 0,1296 | 0,1225 | 0,1156 | 0,1681 |
| 0,0676 | 0,0784 | 0,1024 | 0,1156 | 0,1024 | 0,0676 | 0,1225 | 0,1521 |
| 0,1369 | 0,1444 | 0,1521 | 0,1024 | 0,1089 | 0,1600 | 0,0961 | 0,1600 |
| 0,1024 | 0,1369 | 0,1089 | 0,1681 | 0,1296 | 0,1521 | 0,1600 | 0,0576 |
| 0,0784 | 0,1089 | 0,1056 | 0,1444 | 0,1296 | 0,1024 | 0,1369 | |

Remdamiesi Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijais, patikrinkite hipotezę, kad prietaisas buvo stabilus (pagal matuojamo parametro reikšmių nukrypimus). Laikykite, kad tokiu atveju matuojamasis parametras pasiskirstęs pagal normalųjį dėsnį su su dispersija $\sigma^2 = 0,1090$.

**4.13.** Sumodeliuokite didumo 50 paprastąsias imtis a) normaliojo a. d. $N(3,\ 4)$; b) lognormaliojo a. d. $LN(1,\ 2)$; c) Erlango a. d. $G(3,\ 4)$; d) Koši a. d. $K(0,\ 2)$. Remdamiesi Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo kriterijais patikrinkite hipotezes, kad buvo sumodeliuoti būtent minėtieji a. d.

**4.14.** Lentelėje pateikti dviejų eksperimentų su musėmis rezultatai. Pirmame eksperimente tam tikrais nuodais musės veikiamos 30 sekundžių, antrajame – 60 sekundžių. Paralyžiuojantį nuodų poveikį apibūdina reakcijos laikas ($X_{1i}$ pirmame ir $X_{2i}$ antrame eksperimente), praėjęs nuo musės sąlyčio su nuodais iki to momento, kai musė nebegali stovėti.

| $i$ | $X_{1i}$ | $i$ | $X_{1i}$ | $i$ | $X_{2i}$ | $i$ | $X_{2i}$ |
|---|---|---|---|---|---|---|---|
| 1 | 3,1 | 9 | 53,1 | 1 | 3,3 | 9 | 56,7 |
| 2 | 9,4 | 10 | 59,4 | 2 | 10,0 | 10 | 63,3 |
| 3 | 15,6 | 11 | 65,6 | 3 | 10,7 | 11 | 70,0 |
| 4 | 21,9 | 12 | 71,9 | 4 | 23,3 | 12 | 76,7 |
| 5 | 28,1 | 13 | 78,1 | 5 | 30,0 | 13 | 83,3 |
| 6 | 34,4 | 14 | 84,4 | 6 | 36,7 | 14 | 90,0 |
| 7 | 40,6 | 15 | 90,6 | 7 | 43,3 | 15 | 96,7 |
| 8 | 46,9 | 16 | 96,9 | 8 | 50,0 | | |

Remdamiesi Kolmogorovo ir Smirnovo bei Kramero ir Mizeso dviejų imčių kriterijais, patikrinkite hipotezę, kad imtys gautos stebint tą patį atsitiktinį dydį.

**4.15.** Sudalinkite **2.16** pratimo duomenis į dvi imtis (pirmieji 5 ir likusieji 5 stulpeliai). Remdamiesi Kolmogorovo ir Smirnovo bei Kramero ir Mizeso dviejų imčių kriterijais, patikrinkite hipotezę, kad imtys gautos stebint tą patį atsitiktinį dydį.

**4.16.** Sudalinkite **2.17** pratimo duomenis į dvi imtis (pirmieji 5 ir likusieji 10 stulpelių). Remdamiesi Kolmogorovo ir Smirnovo bei Kramero ir Mizeso dviejų imčių kriterijais patikrinkite hipotezę, kad imtys gautos stebint tą patį atsitiktinį dydį.

## 4.7. Atsakymai ir nurodymai

**4.1.** $\mathbf{E}C = 1/6, \mathbf{V}C = 1/45$; $\mathbf{E}A = 1, \mathbf{V}A = 2\pi^2/3 - 6$. **Nurodymas.** $\mathbf{E}C = \int_0^1 \mathbf{E}(B^2(t))dt$, $\mathbf{E}(C^2) = 2\int_0^1\int_0^t \mathbf{E}(B^2(t)B^2(s))dsdt$. Pasinaudokite tuo, kad $B(t) \sim N(0, t(1-t)), 0 \le t \le 1; (B(s), B(t))^T \sim N_2(\mathbf{0}, \mathbf{\Sigma}), \sigma_{11} = s(1-s), \sigma_{22} = t(1-t), \sigma_{12} = s(1-t), 0 \le s \le t \le 1$. **4.2.** $\mathbf{E}(nC_n) = 1/6, \mathbf{V}(nC_n) = 1/45 - 1/(60n)$; $\mathbf{E}(nA_n) = 1, \mathbf{V}(nA_n) = 2\pi^2/3 - 6 + (10 - \pi^2)/n$. **Nurodymas.** $\mathbf{E}(nC_n) = \int_0^1 n\mathbf{E}(\hat{G}_n(y) - y)^2dy$, $\mathbf{E}(nC_n)^2) = 2\int_0^1\int_0^y n^2\mathbf{E}[(\hat{G}_n(x) - x)^2(\hat{G}_n(y) - y)^2]dxdy$. Atsitiktinis dydis $n\hat{G}_n(y) \sim B(n, y), 0 < y < 1$; atsitiktinis vektorius $(n\hat{G}_n(x), n(\hat{G}_n(y) - \hat{G}_n(x)), n(1 - \hat{G}_n(y)))^T \sim \mathcal{P}_3(n, (x, y - x, 1 - y)), 0 \le x \le y \le 1$. Randame po integralų ženklais parašytų momentų išraiškas ir jas integruojame. **4.5.** $\underline{F(x)} = \max(0, \hat{F}(x) - D_\alpha(n))$, $\overline{F(x)} = \min(\hat{F}(x) + D_\alpha(n), 1)$. **4.8.** Atlikę 4.7 pratime nurodytą randomizaciją, gauname statistikų realizacijas $D_n = 0,0196, C_n = 0,0603, A_n = 0,3901$. Atitinkamos P reikšmės $> 0,25$. Hipotezė neatmetama. **4.9.** Atlikę 4.7 pratime nurodytą randomizaciją, gauname statistikų realizacijas $D_n = 0,0305, C_n = 0,0814, A_n = 0,5955$. Atitinkamos P reikšmės $> 0,25$. Hipotezė neatmetama. **4.10.** a) Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo statistikų realizacijos yra 0,0828, 0,1139, 0,7948, o atitinkamos P reikšmės 0,5018, 0,5244, 0,4834. Hipotezė neatmetama. b) Taikydami modifikuotuosius kriterijus SAS paketu gauname tas pačias statistikų realizacijas, o P reikšmės yra 0,0902, 0,0762, 0,0399. Hipotezės teisingumas kelia abejonių. **4.11.** a) Kolmogorovo ir Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo statistikų realizacijos yra 0,0573, 0,0464, 0,3411, o atitinkamos P reikšmės 0,6927, 0,8911, 0,8978. Hipotezė neatmetama. b) Taikydami modifikuotuosius kriterijus SAS paketu gauname, kad P reikšmė pirmu atveju $> 0,15$, o kitais dviem atvejais $> 0,25$. Atsakymas nepakinta. **4.12.** Kolmogorovo ir -Smirnovo, Kramero ir Mizeso bei Anderseno ir Darlingo statistikų realizacijos yra 0,2547, 0,8637, 4,3106, o atitinkamos P reikšmės 0,0041, 0,0049, 0,0065. Hipotezė atmestina. **4.14.** Statistika $D_{m,n}$ įgijo reikšmę 0,075. Kritinė reikšmė $D_{0,05}(15,\ 16) = 0,475$; apytikslė reikšmė pagal pateiktą aproksimaciją $D_{0,05}(15,\ 16) \approx 0,436$. Asimptotinė P reikšmė $pv_a = 1 - K(0,2087) \approx 1$. Hipotezė neatmetama. Kramero ir Mizeso statistikos reikšmė 0,0032 ir $pv_a = 1 - a_1(0,0032) \approx 1$. Hipotezė neatmetama. **4.15.** Statistika $D_{m,n}$ įgijo reikšmę 0,16, $pv = 0,471$ ir $pv_a = 1 - K(1,1) = 0,5441$. Kramero ir Mizeso statistikos reikšmė 0,2511 ir $pv_a = 1 - a_1(0,2511) = 0,187$. Hipotezė neatmetama. **4.16.** Statistika $D_{m,n}$ įgijo reikšmę 0,13, $pv = 0,5227$ ir $pv_a = 0,6262$. Kramero ir Mizeso statistikos reikšmė 0,1668 ir $pv_a = 1 - a_1(0,1668) = 0,3425$. Hipotezė neatmetama.

# 5 skyrius

# Ranginiai kriterijai

## 5.1. Įvadas

Ankstesniuose skyriuose aptarėme du neparametrinių kriterijų sudarymo metodus. 2 skyriaus chi-kvadrato tipo kriterijų statistikos priklauso tik nuo imties elementų patekimo į tam tikras aibes dažnių. Ketvirto skyriaus kriterijai grindžiami funkcionalais nuo empirinės ir hipotetinės teorinės pasiskirstymo funkcijų skirtumo. Jų nepriklausomumas nuo skirstinio pavidalo grindžiamas integraline transformacija (žr. 4.1.1 teoremą), kuria absoliučiai tolydieji a. d. keičiami tolygiai intervale [0, 1] pasiskirsčiusiais atsitiktiniais dydžiais.

Šiame skyriuje aptarsime dar vieną nesusijusių su skirstinio pavidalu kriterijų sudarymo metodą. Pateikiamų kriterijų statistikos priklauso tik nuo stebėjimo rezultatų tarpusavio padėties variacinėje eilutėje, o ne tiesiogiai nuo jų pačių.

## 5.2. Rangai ir jų skirstiniai

Tarkime, $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis absoliučiai tolydžiojo a. d. $X$, o $X_{(1)} < \cdots < X_{(n)}$ yra pozicinės statistikos, gautos iš šios imties.

**5.2.1 apibrėžimas.** Imties elemento $X_i$ rangu $R_i$ vadinamas to elemento eilės numeris variacinėje eilutėje $(X_{(1)}, ..., X_{(n)})$, t. y.

$$rangas(X_i) = R_i = j, \quad \text{jeigu} \quad X_i = X_{(j)}.$$

Pavyzdžiui, jeigu $(X_1, ..., X_5)^T = (63,\ 32,\ 41,\ 25,\ 38)^T$, tai

$$(X_{(1)}, ..., X_{(5)})^T = (25,\ 32,\ 38,\ 41,\ 63)^T$$

$$(R_1, ..., R_5)^T = (5,\ 2,\ 4,\ 1,\ 3)^T.$$

Rangai įgyja reikšmes $1, 2, \ldots, n$, todėl jų suma yra konstanta:

$$R_1 + \cdots + R_n = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Ieškosime a. v., sudaryto iš rangų

$$(R_{i_1}, \ldots, R_{i_k})^T, \quad 1 \le i_1 < \cdots < i_k \le n$$

tikimybinio skirstinio. Kadangi a. d. $X_1, \ldots, X_n$ yra nepriklausomi ir vienodai pasiskirstę, šis rangų vektorius įgyja $n(n-1)\ldots(n-k+1) = n!/(n-k)!$ skirtingų reikšmių su vienodomis tikimybėmis. Taigi

$$\mathbf{P}\{(R_{i_1}, ..., R_{i_k}) = (j_1, ..., j_k)\} = \frac{(n-k)!}{n!} \tag{5.2.1}$$

su kiekvienu rinkiniu $(j_1, ..., j_k)$, sudarytu iš $k$ skirtingų aibės $\{1, \ldots, n\}$ elementų.

Atskirais atvejais

$$\mathbf{P}\{R_i = j\} = \frac{1}{n}, \quad \mathbf{P}\{(R_{i_1}, R_{i_2}) = (j_1, j_2)\} = \frac{1}{n(n-1)},$$

$$\mathbf{P}\{(R_1, ..., R_n) = (j_1, ..., j_n)\} = \frac{1}{n!}. \tag{5.2.2}$$

**5.2.1 teorema.** *Jeigu a. d. $X$ skirstinys absoliučiai tolydus, tai*

$$\mathbf{E}R_i = \frac{n+1}{2}, \quad \mathbf{V}R_i = \frac{n^2-1}{12}, \quad \mathbf{cov}(R_i, R_j) = -\frac{n+1}{12}, \quad i \ne j. \tag{5.2.3}$$

**Įrodymas.** Gauname

$$\mathbf{E}R_i = \sum_{j=1}^{n} j\,\mathbf{P}\{R_i = j\} = \frac{1}{n}(1 + ... + n) = \frac{n+1}{2},$$

$$\mathbf{V}R_i = \mathbf{E}(R_i^2) - (\mathbf{E}R_i)^2$$

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

Jeigu $i \ne j$, tai

$$\mathbf{cov}(R_i, R_j) = \mathbf{E}(R_iR_j) - \mathbf{E}R_i\mathbf{E}R_j$$

$$= \sum\sum\nolimits_{k \ne l} kl \frac{1}{n(n-1)} - \frac{(n+1)^2}{4} = [(\sum_{k=1}^{n} k)^2 - \sum_{k=1}^{n} k^2] \frac{1}{n(n-1)} - \frac{(n+1)^2}{4}$$

$$= \frac{n(n+1)^2}{4(n-1)} - \frac{(n+1)(2n+1)}{6(n-1)} - \frac{(n+1)^2}{4} = \frac{(n+1)}{2(n-1)} \left[ \frac{n(n+1)}{2} - \frac{2n+1}{3} - \frac{n^2-1}{2} \right]$$

$$= \frac{(n+1)}{2(n-1)} \cdot \frac{-n+1}{6} = -\frac{n+1}{12}.$$

▲

**Sutampančios reikšmės.** Apibendrinsime rango sąvoką tuo atveju, kai a.d. $X_1$ skirstinys nebūtinai absoliučiai tolydus. Jei variacinėje eilutėje $X_{(1)} \leq \cdots \leq X_{(n)}$ yra grupė iš $t$ sutampančių elementų:

$$X_{(j-1)} < X_{(j)} = X_{(j+1)} = \cdots = X_{(j+t-1)} < X_{(j+t)}, \tag{5.2.4}$$

ir $X_i = X_{(j)}$, tai imties elemento $X_i$ rangą apibrėžiame kaip jo ir su juo sutampančių stebinių pozicijų variacinėje eilutėje vidurkį:

$$R_i = \frac{j + (j+1) + \cdots + (j+t-1)}{t} = j + \frac{t-1}{2}. \tag{5.2.5}$$

Pavyzdžiui, jei turime imties realizaciją 6, 7, 5, 10, 7, 6, 7, tai variacinė eilutė yra 5, 6, 6, 7, 7, 7, 10 ir rangai:

$$R_1 = \frac{2+3}{2} = 2,5;\ R_2 = \frac{4+5+6}{3} = 5;\ R_3 = 1;\ R_4 = 7;$$
$$R_5 = 5;\ R_6 = 2,5;\ R_7 = 5.$$

Tarkime, kad $X_{i_1} = \cdots = X_{i_t} = X_{(j)}$; čia $j$ yra numeris, su kuriuo patenkinama (5.2.4). Tada rangų suma

$$R_{i_1} + \cdots + R_{i_t} = j + (j+1) + \cdots + (j+t-1)$$

yra tokia pati, kaip ir absoliučiai tolydžiu atveju, kai pozicinės statistikos $X_{(j)}$, $X_{(j+1)}$, ..., $X_{(j+t-1)}$ įgyja skirtingas reikšmes. Taigi visų rangų suma kiek absoliučiai tolydžių, tiek kitokių skirstinių yra tokia pati:

$$R_1 + \cdots + R_n = n(n+1)/2.$$

Pažymėkime $k$ atsitiktinį grupių su vienodais elementais skaičių, $t_l$ – elementų skaičių $l$-oje grupėje ir

$$T = \sum_{l=1}^{k} t_l(t_l^2 - 1).$$

Jeigu $t_1 = \cdots = t_n = 1$, tai $T = 0$.

**5.2.2 teorema.** *Rangų vidurkiai, dispersijos ir kovariacijos yra*

$$\mathbf{E}R_i = \frac{n+1}{2}, \quad \mathbf{V}R_i = \frac{n^2-1}{12} - \frac{\mathbf{E}T}{12n},$$
$$\mathbf{cov}(R_i, R_j) = -\frac{n+1}{12} + \frac{\mathbf{E}T}{12n(n-1)}, \quad (i \neq j). \tag{5.2.6}$$

**Įrodymas.** Atsitiktiniai dydžiai $X_i$ yra nepriklausomi ir vienodai pasiskirstę, todėl a.d. $R_1, \ldots, R_n$ yra vienodai pasiskirstę ir

$$\mathbf{E}R_i = \frac{1}{n}\mathbf{E}(\sum_{j=1}^{n} R_j) = \frac{n+1}{2}, \quad i = 1, \ldots, n.$$

Remdamiesi (5.2.5) ir pažymėję $l$-osios grupės pirmojo nario poziciją variacinėje eilutėje raide $j$, randame rangų kvadratų sumą

$$\sum_{i=1}^{n} R_i^2 = \sum_{l=1}^{k} t_l(j_l + \frac{t_l - 1}{2})^2 = \sum_{l=1}^{k} t_l[j_l^2 + j_l(t_l-1) + (t_l-1)^2/4].$$

Gauname

$$1^2 + \cdots + n^2 = \sum_{l=1}^{k}[j_l^2 + \cdots + (j_l + t_l - 1)^2] = \sum_{l=1}^{k}[t_l j_l^2 + 2j_l(1 + \cdots + (t_l - 1))$$
$$+(1^2 + \cdots + (t_l-1)^2)] = \sum_{l=1}^{k} t_l[(j_l^2 + j_l(t_l-1) + \frac{(t_l-1)(2t_l-1)}{6}] = \sum_{i=1}^{n} R_i^2 + \frac{T}{12}.$$

Išreiškę $\sum\limits_{i=1}^{n} R_i^2$ ir pasinaudoję lygybe $1^2 + \cdots + n^2 = n(n+1)(2n+1)/6$, gauname

$$\sum_{i=1}^{n} R_i^2 = \frac{n(n+1)(2n+1)}{6} - \frac{T}{12}.$$

A. d. $R_i$ vienodai pasiskirstę, todėl

$$\mathbf{V}(R_i) = \mathbf{E}(R_i^2) - (\mathbf{E}R_i)^2 = \frac{(n+1)(2n+1)}{6} - \frac{\mathbf{E}T}{12n} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12} - \frac{\mathbf{E}T}{12n}.$$

Rikia pažymėti, kad

$$\sum_{j=1}^{n}\sum_{l=1}^{n} R_j R_l = (\frac{n(n+1)}{2})^2,$$

todėl

$$\mathbf{E}\sum\sum\nolimits_{j\neq l} R_j R_l = (\frac{n(n+1)}{2})^2 - \sum_{j=1}^{n} \mathbf{E}R_j^2 = (\frac{n(n+1)}{2})^2 - \frac{n(n+1)(2n+1)}{6} + \frac{\mathbf{E}T}{12}$$

$$= \frac{n(n^2-1)(3n+2)}{12} + \frac{\mathbf{E}T}{12}, \quad \mathbf{E}R_j R_l = \frac{(n+1)(3n+2)}{12} + \frac{\mathbf{E}T}{12n(n-1)},$$

$$\mathbf{cov}(R_j, R_l) = \frac{(n+1)(3n+2)}{12} + \frac{\mathbf{E}T}{12n(n-1)} - \frac{(n+1)^2}{4} = -\frac{n+1}{12} + \frac{\mathbf{E}T}{12n(n-1)}.$$

▲

**5.2.1 pastaba.** Jeigu imtis $(X_1, \ldots, X_n)^T$ yra simetriškas a. v., t. y. $(X_{i_1}, \ldots, X_{i_n})^T$ turi tą patį skirstinį su visomis $(1, \ldots, n)$ perstatomis $(i_1, \ldots, i_n)$, tai rangų skirstiniai išlieka tie patys kaip ir paprastosios imties.

## 5.3. Ranginiai nepriklausomumo kriterijai

### 5.3.1. Spirmeno nepriklausomumo kriterijus

Tarkime, kad

$$(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$$

yra paprastoji imtis a. v. $(X, Y)^T$ su pasiskirstymo funkcija $F = F(x, y) \in \mathcal{F}$; čia $\mathcal{F}$ yra neparametrinė absoliučiai tolydžių dvimačių pasiskirstymo funkcijų šeima.

**Nepriklausomumo hipotezė** (dviejų a. d.):

$$H_0 : F \in \mathcal{F}_0;$$

čia $\mathcal{F}_0 \subset \mathcal{F}$ yra šeima dvimačių pasiskirstymo funkcijų, kurios lygios marginaliųjų pasiskirstymo funkcijų sandaugai:

$$F(x, y) = F_1(x) F_2(y) \quad \text{su visais} \quad (x, y) \in \mathbf{R}^2;$$

čia $F_1(x) = \mathbf{P}\{X \leq x\}$ ir $F_2(y) = \mathbf{P}\{Y \leq y\}$.

Pažymėkime $R_{11}, \ldots, R_{1n}$ ir $R_{21}, \ldots, R_{2n}$ atitinkamai imčių $X_1, \ldots, X_n$ ir $Y_1, \ldots, Y_n$ narių rangus. Šiuos rangus ir naudosime nepriklausomumo hipotezei tikrinti.

**4.3.1 apibrėžimas.** Empirinis pirmosios ir antrosios imties rangų koreliacijos koeficientas

$$r_S = \frac{\sum_{j=1}^{n}(R_{1j}-\bar{R}_1)(R_{2j}-\bar{R}_2)}{[\sum_{j=1}^{n}(R_{1j}-\bar{R}_1)^2\sum_{j=1}^{n}(R_{2j}-\bar{R}_2)^2]^{1/2}}, \quad \bar{R}_i = \frac{1}{n}\sum_{j=1}^{n}R_{ij} = \frac{n+1}{2}, \tag{5.3.1}$$

vadinamas *Spirmeno ranginiu koreliacijos koeficientu.*

Koeficiento $r_S$ reikšmė nepakinta, jei stebėjimai $(X_i, Y_i)^T$, $i = 1, 2, ..., n$, išdėstomi taip, kad $Y_i$ sudarytų didėjančią seką ir tada duomenys pakeičiami jų rangais, nes išlieka tos pačios rangų poros, tik surašytos kita tvarka. Tada vietoje eilutės $(R_{21}, \dots, R_{2n})$ gaunama eilutė $(1, 2, ..., n)$, o vietoje $(R_{11}, \dots, R_{1n})$ – eilutė, kurios elementus pažymėsime $R_1, ..., R_n$. Gauname:

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(i - \frac{n+1}{2})}{[\sum_{i=1}^{n}(R_i - \frac{n+1}{2})^2\sum_{i=1}^{n}(i - \frac{n+1}{2})^2]^{1/2}}.$$

Kadangi

$$\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)^2 = \sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)^2 = n\mathbf{V}(R_i) = \frac{n(n^2-1)}{12},$$

$$\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(i - \frac{n+1}{2}\right)$$

$$= \sum_{i=1}^{n} iR_i - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12} - \frac{1}{2}\sum_{i=1}^{n}(R_i - i)^2,$$

tai Spirmeno koreliacijos koeficientą galima užrašyti patogesne skaičiuoti forma.

**Spirmeno ranginis koreliacijos koeficientas:**

$$r_S = \frac{\frac{1}{n}\sum_{i=1}^{n} iR_i - (\frac{n+1}{2})^2}{(n^2-1)/12} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - i)^2. \tag{5.3.2}$$

Kai nepriklausomumo hipotezė teisinga, a. v. $(R_1, \dots, R_n)^T$ skirstinys sutampa su a. v. $(R_{11}, \dots, R_{1n})^T$ skirstiniu. Taigi jis nepriklauso nuo jokių nežinomų parametrų, o priklauso tik nuo imties didumo $n$.

**Spirmeno nepriklausomumo kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$r_s \le c_1 \quad \text{arba} \quad r_s \ge c_2; \tag{5.3.3}$$

čia $c_1$ – minimali, o $c_2$ – maksimali $r_S$ reikšmės, tenkinančios nelygybes

$$\mathbf{P}\{r_s \le c_1\} \le \alpha/2, \quad \mathbf{P}\{r_s \ge c_2\} \le \alpha/2.$$

Skaičiuoti nedidelių $n$ a. d. $r_S$ įgyjamų reikšmių tikimybes, kartu kritines reikšmes nesudėtinga, nes $r_S$ yra a. v. $(R_1, \ldots, R_n)^T$, kurio skirstinys pateiktas (5.2.2) formulėse, funkcija.

Jei $n$ didelis, tai naudojantis CRT a. d. $r_S$ skirstinys aproksimuojamas normaliuoju. Rasime pirmuosius du $r_S$ momentus. Kai nepriklausomumo hipotezė teisinga, tai

$$\mathbf{E}\sum_i i\left(R_i - \frac{n+1}{2}\right) = 0, \quad \mathbf{V}(\sum_i iR_i) = \mathbf{V}(R_1)\sum_i i^2 + \mathbf{cov}(R_1, R_2)\sum_{i\neq j} ij$$

$$= \frac{n^2-1}{12}\frac{n(n+1)(2n+1)}{6} - \frac{n+1}{12}\left[\left(\frac{n(n+1)}{2}\right)^2 - \frac{n(n+1)(2n+1)}{6}\right]$$

$$= \frac{n^2(n+1)^2(n-1)}{144}.$$

Remdamiesi (5.3.2) gauname

$$\mathbf{E}r_S = 0, \quad \mathbf{V}\, r_S = \frac{1}{n-1}.$$

Kai $n$ didelis, a. d. $r_S$ skirstinys aproksimuojamas normaliuoju:

$$Z_n = \sqrt{n-1}\, r_S \xrightarrow{d} Z \sim N(0,\, 1). \tag{5.3.4}$$

Naudodamiesi šia aproksimacija galime sudqaruti asimptotinį kriterijų.

**Asimptotinis Spirmeno nepriklausomumo kriterijus:** nepriklausomumo hipotezė atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai

$$|Z_n| > z_{\alpha/2}. \tag{5.3.5}$$

Vidutinio didumo imtims rekomenduojama taikyti aproksimaciją Stjudento skirstiniu, t. y. statistikos

$$t_n = \sqrt{n-2}\frac{r_S}{\sqrt{1-r_S^2}}$$

skirstinys aproksimuojamas Stjudento skirstiniu $S(n-2)$.

**Asimptotinis Spirmeno nepriklausomumo kriterijus grindžiamas Stjudento skirstiniu:** nepriklausomumo hipotezė atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai

$$|t_n| > t_{\alpha/2}(n-2). \tag{5.3.6}$$

**Modifikacija, kai yra sutampančių reikšmių.** Tarkime, kad yra vienodų rangų. Remiantis formule (5.3.1) Spirmeno ranginį koreliacijos koeficientą galima skaičiuoti taip:

$$r_S = \frac{\sum_{j=1}^n (R_{1j} - \frac{n+1}{2})(R_{2j} - \frac{n+1}{2})}{[\sum_{j=1}^n (R_{1j} - \frac{n+1}{2})^2 \sum_{j=1}^n (R_{2j} - \frac{n+1}{2})^2]^{1/2}} =$$

$$\frac{\sum_{j=1}^{n} R_{1j}R_{2j} - n(\frac{n+1}{2})^2}{[(\sum_{j=1}^{n} R_{1j}^2 - n(\frac{n+1}{2})^2)(\sum_{j=1}^{n} R_{2j}^2 - n(\frac{n+1}{2}))^2]^{1/2})}.$$

Pagal (5.2.6). kai yra teisinga hipotezė, skaitiklio vidurkis lygus nuliui, o dispersija yra

$$\sum_{j=1}^{n} \mathbf{V}(R_{1j})\mathbf{V}(R_{2j}) = n(\frac{n^2-1}{12})^2(1 - \frac{\mathbf{E}T_X}{n^3-n})(1 - \frac{\mathbf{E}T_Y}{n^3-n});$$

čia

$$T_X = \sum_{l=1}^{k} t_l(t_l^2 - 1),$$

$k$ yra atsitiktinis skaičius sutampančių grupių imtyje $X_1, ..., X_n$; $t_l$ – elementų skaičius $l$-oje grupėje. $T_Y$ apibrėžiamas analogiškai.

Remdamiesi teoremos 5.2.2 įrodymu, gauname

$$\sum_{j=1}^{n} R_{1j}^2 - n\left(\frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{12} - \frac{T_X}{12}.$$

Taigi vardiklio kvadratas yra toks:

$$\left(\frac{n(n^2-1)}{12}\right)^2 (1 - \frac{T_X}{n^3-n})(1 - \frac{T_Y}{n^3-n}).$$

Vadinasi, jeigu $n$ yra didelis ir nepriklausomumo hipotezė teisinga, tai ir esant kai kuriems vienodiems rangams pritaikoma tokia pati aproksimacija ir gaunami asimptotiniai kriterijai (5.3.5) ir (5.3.6).

**5.3.1 pavyzdys.** Lentelėje pateikiami $n = 50$ moksleivių matematikos $X_i$ ir kalbų $Y_i$ žinių tikrinimo testų rezultatai.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 12 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 59 | 63 | 72 | 55 | 50 | 46 | 67 | 61 | 67 | 53 | 39 | 41 | 62 | 51 | 64 | 52 | 54 |
| $Y_i$ | 50 | 55 | 53 | 54 | 59 | 52 | 57 | 58 | 57 | 60 | 49 | 59 | 59 | 50 | 66 | 51 | 59 |

| $i$ | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 59 | 64 | 32 | 48 | 65 | 62 | 53 | 65 | 58 | 51 | 53 | 64 | 64 | 61 | 65 | 40 | 52 |
| $Y_i$ | 60 | 58 | 57 | 52 | 57 | 52 | 58 | 58 | 64 | 55 | 54 | 56 | 57 | 59 | 62 | 54 | 55 |

| $i$ | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 38 | 56 | 49 | 60 | 52 | 65 | 68 | 58 | 47 | 39 | 59 | 60 | 42 | 51 | 52 | 65 |
| $Y_i$ | 51 | 64 | 55 | 50 | 50 | 54 | 59 | 59 | 57 | 42 | 49 | 50 | 37 | 46 | 48 | 60 |

Ar šie duomenys neprieštarauja prielaidai, kad dviejų testų rezultatai yra nepriklausomi?

Matematikos testo rezultatų rangus pažymėkime $R_{1i}$, o kalbų testų – $R_{2i}$. Jeigu yra sutampančių rezultatų, tai rangus priskiriame pagal (5.2.5). Gautas rangų reikšmes pateikiame lentelėje.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 12 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R_{1i}$ | 29 | 37 | 50 | 24 | 12 | 8 | 47,5 | 33.5 | 47,5 | 21 | 3,5 | 6 | 35,5 | 14 |
| $R_{2i}$ | 9 | 23,5 | 17 | 19,5 | 40 | 15 | 29,5 | 34,5 | 29,5 | 45 | 5,5 | 40 | 40 | 9 |

| $i$ | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R_{1i}$ | 39,5 | 17,5 | 23 | 29 | 39,5 | 1 | 10 | 44 | 35,5 | 21 | 44 | 26,5 | 14 |
| $R_{2i}$ | 50 | 12,5 | 40 | 45 | 34,5 | 29,5 | 15 | 29,5 | 15 | 34,5 | 34,5 | 48,5 | 23,5 |

| $i$ | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R_{1i}$ | 21 | 39,5 | 39,5 | 33.5 | 44 | 5 | 17,5 | 2 | 25 | 11 | 31,5 | 17,5 |
| $R_{2i}$ | 19,5 | 26 | 29,5 | 40 | 47 | 19,5 | 23,5 | 12,5 | 48,5 | 23,5 | 9 | 9 |

| $i$ | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R_{1i}$ | 44 | 49 | 26,5 | 9 | 3,5 | 29 | 31,5 | 7 | 14 | 17,5 | 44 |
| $R_{2i}$ | 19,5 | 40 | 40 | 29,5 | 2 | 5,5 | 9 | 1 | 3 | 4 | 45 |

Spirmeno ranginis koreliacijos koeficientas įgijo reikšmę

$$r_S = \frac{\sum_{j=1}^{50} R_{1j}R_{2j} - 50(\frac{51}{2})^2}{[(\sum_{j=1}^{50} R_{1j}^2 - 50(\frac{51}{2})^2)(\sum_{j=1}^{50} R_{1j}^2 - 50(\frac{51}{2})^2)]^{1/2}} = 0,39003.$$

Pagal asimptotinį Spirmeno kriterijų, grindžiamą normaliąja aproksimacija, gauname:

$$Z_n = \sqrt{n-1}\, r_S = \sqrt{49}\ 0,39003 = 2,73024.$$

Asimptotinė $P$ reikšmė

$$pv_a = 2(1 - \Phi(2,73024)) = 0,00633.$$

Pagal asimptotinį Spirmeno kriterijų, grindžiamą aproksimacija Stjudento skirstiniu, gauname:

$$t_n = \sqrt{50-2}\frac{039003}{\sqrt{1-039003^2}} = 2,93462, \quad pv_a = 0,00511.$$

Nepriklausomumo hipotezė atmetama, nes P reikšmės yra mažos. Kadangi $r_S > 0$, tai galima daryti išvadą, kad yra teigiama matematikos ir kalbų žinių priklausomybė.

## 5.3.2. Kendalo nepriklausomumo kriterijus

**5.3.2 apibrėžimas.** Kiekvieną atvejį, kai didesnis skaičius parašytas prieš mažesnį, vadiname *inversija.* Pavyzdžiui, kėlinyje (5, 3, 1, 4, 2) yra $4+2+1 = 7$ inversijos.

Inversijų skaičius $I_n$ kėlinyje $(R_1, ..., R_n)$ įgyja reikšmes nuo 0 (skaičiai išdėstyti didėjančia tvarka) iki $N_n = n(n-1)/2$ (skaičiai išdėstyti mažėjančia tvarka).

**5.3.1 teorema.** *Kai nepriklausomumo hipotezė teisinga, tai inversijų skaičiaus $I_n$ charakteristinė funkcija yra*

$$\varphi_n(t) = \mathbf{E}e^{itI_n} = \frac{1}{n!}\prod_{j=1}^{n}\frac{e^{itj}-1}{e^{it}-1} = \prod_{j=1}^{n}\left(\frac{1+e^{it}+\cdots+e^{it(j-1)}}{j}\right). \qquad (5.3.7)$$

**Įrodymas.** Turime:

$$\varphi_n(t) = \mathbf{E}e^{itI_n} = \sum_{k=0}^{N_n} e^{itk}\frac{\nu_n(k)}{n!}; \qquad (5.3.8)$$

čia $\nu_n(k)$ yra aibės $\{1, 2, ..., n\}$ kėlinių $\{i_1, ..., i_n\}$, turinčių $k$ inversijų, skaičius.

Skaičių $\nu_n(k)$ išreiškiame skaičiais $\nu_{n-1}(l)$, $k-(n-1) \le l \le \min(k,\ n-1)$.

Reikia pažymėti, kad $n!$ eilutės $\{1, 2, ..., n\}$ kėlinių galime gauti iš $(n-1)!$ skaičių $\{1, 2...., n-1\}$ kėlinių įterpiant "$n$" į visas galimas vietas.

Pavyzdžiui, iš 2! kėlinių (1, 2) ir (2, 1) gauname 3! kėlinių įdedant 3 į visas galimas vietas: (1, 2, 3), (1, 3, 2), (3, 1, 2), (2, 1, 3), (2, 3, 1), (3, 2, 1).

Eilutės $(1, 2, \ldots, n)$ kėlinys, turintis $k$ inversijų, gaunamas iš eilutės $(1, 2, \ldots, n-1)$ kėlinių, turinčių $k-l$ inversijų ir įstatant $n$ į tokią vietą, už kurios yra

$l$ skaičių ($l = 0, \dots, n-1$, jei $k \geq n-1$; $l = 0, \dots, k$, jei $0 \leq k < n-1$). Taigi teisingos lygybės

$$\nu_n(k) = \nu_{n-1}(k) + \nu_{n-1}(k-1) + \dots + \nu_{n-1}(k-(n-1)), \quad k \geq n-1,$$

$$\nu_n(k) = \nu_{n-1}(k) + \nu_{n-1}(k-1) + \dots + \nu_{n-1}(0), \quad 0 \leq k < n-1. \tag{5.3.9}$$

Su kiekvienu $l = 0, 1, \dots, N_{n-1}$ narys $\nu_{n-1}(l)$ įeina į $\nu_n(k)$ išraišką formulėje (5.3.6), kai $k = l, \dots, l+n-1$, nes įstatant $n$ inversijų skaičius gali padidėti $0, 1, \dots, n-1$ inversija. Taigi

$$\varphi_n(t) = \sum_{k=0}^{N_n} e^{itk} \frac{\nu_n(k)}{n!} = \sum_{l=0}^{N_{n-1}} \frac{\nu_{n-1}(l)}{n!} \sum_{k=l}^{l+n-1} e^{itk}$$

$$= \frac{\varphi_{n-1}(t)}{n}[1 + e^{it} + \dots + e^{it(n-1)}] = \frac{e^{itn}-1}{n(e^{it}-1)}\varphi_{n-1}(t).$$

Pakartotinai taikydami šią lygybę ir atsižvelgę į tai, kad

$$\varphi_1(t) = \nu_1(0) = 1, \quad \varphi_2(t) = \frac{\nu_2(0)}{2} + \frac{\nu_2(1)}{2}e^{it} = \frac{e^{2it}-1}{2(e^{it}-1)},$$

gauname (5.3.5). ▲

**5.3.3 apibrėžimas.** Atsitiktinis dydis

$$r_K = 1 - \frac{4I_n}{n(n-1)}, \quad -1 \leq r_K \leq 1. \tag{5.3.10}$$

yra vadinamas *Kendalo ranginiu koreliacijos koeficientu.*

**5.3.2 teorema.** *Kai nepriklausomumo hipotezė teisinga, tai*

$$\mathbf{E}r_K = 0, \quad \boldsymbol{V}r_K = \frac{2(2n+5)}{9n(n-1)},$$

*ir*

$$\frac{r_K}{\sqrt{\boldsymbol{V}r_K}} \xrightarrow{d} Z \sim N(0,1), \quad \textit{kai} \quad n \to \infty. \tag{5.3.11}$$

**Įrodymas.** Remdamiesi $I_n$ charakteristinės funkcijos išraiška (5.3.5), gauname, kad $I_n$ yra n. a. d. suma:

$$I_n = U_1 + U_2 + \dots + U_n; \tag{5.3.12}$$

čia $U_1 = 0$, o a. d. $U_j$ įgyja reikšmes $0,\ 1, \dots, j-1$ su vienodomis tikimybėmis $1/j$, $\ j = 2.\dots, n$, nes a. d. $U_j$ charakteristinė funkcija yra

$$\psi(t) = \mathbf{E}e^{itU_j} = \sum_{k=0}^{j-1} e^{itk}\frac{1}{j}.$$

Gauname

$$\mathbf{E}U_j = \frac{1}{j}\sum_{i=1}^{j-1} i = \frac{1}{j}\frac{j(j-1)}{2} = \frac{j-1}{2},$$

$$\mathbf{E}U_j^2 = \frac{(j-1)(2j-1)}{6}, \quad \boldsymbol{V}U_j = \frac{j^2-1}{12},$$

todėl

$$\mathbf{E}I_n = \sum_{j=1}^{n}\frac{j-1}{2} = \frac{n(n-1)}{4}, \quad \mathbf{E}R_K = 0;$$

$$\boldsymbol{V}I_n = \sum_{j=1}^{n}\frac{j^2-1}{12} = \frac{n(n-1)(2n+5)}{72},$$

$$\boldsymbol{V}r_K = \frac{16\boldsymbol{V}I_n}{n^2(n-1)^2} = \frac{2(2n+5)}{9n(n-1)}.$$

Įrodysime, kad inversijų skaičiui galioja CRT. Tuo tikslu pakanka patikrinti Lindebergo sąlygą

$$\frac{1}{\boldsymbol{V}I_n}\sum_{j=1}^{n}\int_{\frac{|x-\mathbf{E}U_j|}{\sqrt{\boldsymbol{V}_{I_n}}}>\epsilon}(x-\mathbf{E}U_j)^2 dF_{U_j}(x) \to 0, \quad \text{kai} \quad n\to\infty.$$

Atsitiktiniai dydžiai $U_j$ įgyja reikšmes nuo 1 iki $j-1$, $j=1,2,...,n$, $\sqrt{\boldsymbol{V}I_n} = O(n^{3/2})$, taigi

$$\frac{|x-\mathbf{E}U_j|}{\sqrt{\boldsymbol{V}I_n}} \le \frac{n-1}{O(n^{3/2})} < \varepsilon,$$

nes su pakankamai dideliais $n$ visos integravimo sritys yrs tuščios aibės. Lindebergo sąlyga patenkinta, todėl a. d. $I_n$ galioja CRT, ji galioja ir a. d. $r_K$, kuris yra tiesinė $I_n$ funkcija. ▲

Kai $n$ nedideli, statistikos $r_K$ kritinės reikšmės tabuliuotos arba jų skaičiavimas numatytas daugumoje matematinės statistikos paketų.

**Kendalo nepriklausomumo kriterijus:** nepriklausomumo hipotezė atmetama $\alpha$ lygmens kriterijumi, kai

$$r_K \le c_1 \quad \text{or} \quad r_K \ge c_2; \tag{5.3.13}$$

čia $c_1$ – minimali ir $c_2$ – maksimali statistikos $r_K$ reikšmės, tenkinančios sąlygas

$$\mathbf{P}\{r_K \le c_1\} \le \alpha/2. \quad \mathbf{P}\{r_K \ge c_2\} \le \alpha/2.$$

Kai $n$ yra didelis, taikoma aproksimacija normaliuoju skirstiniu.

**Asimptotinis Kendalo nepriklausomumo kriterijus:** nepriklausomumo hipotezė atmetama asimptotiniu $\alpha$ lygmens kriterijumi, kai

$$|\frac{r_K}{\sqrt{\boldsymbol{V}(r_K)}}| > z_{\alpha/2}. \tag{5.3.14}$$

**Sutampančios reikšmės.** Apibendrinsime Kendalo ranginio koreliacijos koeficiento apibrėžimą tuo atveju, kai imtyje yra sutampančių reikšmių.

Sakysime, kad poros $(X_i, Y_i)^T$ ir $(X_j, Y_j)^T$ yra *suderintos*, jeigu skirtumai $X_j - X_i$ ir $Y_j - Y_i$ yra vienodų ženklų: $(X_j - X_i)(Y_j - Y_i) > 0$. Poros yra *nesuderintos*, jei $(X_j - X_i)(Y_j - Y_i) < 0$. Tegu

$$A_{ij} = \begin{cases} 1, \text{ jei } (X_j - X_i)(Y_j - Y_i) > 0, \\ -1, \text{ jei } X_j - X_i)(Y_j - Y_i) < 0, \\ 0, \text{ jei } (X_j - X_i)(Y_j - Y_i) = 0. \end{cases}$$

**5.3.4 apibrėžimas.** Statistika

$$\tau_a = \frac{2}{n(n-1)} \sum_{i<j} A_{ij} \tag{5.3.15}$$

vadinama *Kendalo $\tau_a$ koreliacijos koeficientu.*

$\sum_{i<j} A_{ij}$ yra *suderintų ir nesuderintų porų skaičių skirtumas.*

**5.3.1 pastaba.** Jeigu sutampančių reikšmių nėra, tai $r_K = \tau_a$.

Iš tikrųjų šiuo atveju

$$h_{ij} = \begin{cases} 1, \text{ kai } R_i > R_j, \\ 0, \text{ kai } R_i < R_j, \end{cases}$$

ir inversijų skaičius užrašomas suma

$$I_n = \sum_{i<j} h_{ij}. \tag{5.3.16}$$

Skaičius $-1$ pasikartoja $\sum_{i<j} h_{ij}$ kartų, o skaičius 1 pasikartoja $n(n-1)/2 - \sum_{i<j} h_{ij}$ kartų sumoje $\sum_{i<j} A_{ij}$. Gauname

$$\sum_{i<j} A_{ij} = n(n-1)/2 - 2\sum_{i<j} h_{ij} = n(n-1)/2 - 2I_n,$$

taigi

$$\frac{2}{n(n-1)} \sum_{i<j} A_{ij} = 1 - \frac{4I_n}{n(n-1)}.$$

**5.3.2 pastaba.** Jeigu yra sutampančių reikšmių, tai Kendalo $\tau_a$ koreliacijos koeficientas gali nebūti lygus 1 netgi tada, kai $X_i = Y_i$ su visais $i$, nes ne visi sumos $\sum_{i<j} A_{ij}$ dėmenys lygūs 1. Ši suma turi $n(n-1)/2$ dėmenų.

Apibrėšime koeficiento modifikaciją.

Atsitiktinį dydį $A_{ij}$ užrašykime tokiu pavidalu $A_{ij} = U_{ij}V_{ij}$, čia

$$U_{ij} = \begin{cases} 1, \text{ jei } X_j - X_i > 0, \\ -1, \text{ jei } X_j - X_i < 0, \\ 0, \text{ jei } X_j - X_i = 0. \end{cases}, \quad V_{ij} = \begin{cases} 1, \text{ jei } Y_j - Y_i > 0, \\ -1, \text{ jei } Y_j - Y_i < 0, \\ 0, \text{ jei } Y_j - Y_i = 0. \end{cases}$$

**5.3.5 apibrėžimas.** Statistika

$$\tau_b = \frac{\sum_{i=1}^n \sum_{j=1}^n U_{ij}V_{ij}}{[(\sum_{i=1}^n \sum_{j=1}^n U_{ij}^2)(\sum_{i=1}^n \sum_{j=1}^n V_{ij}^2)]^{1/2}} \tag{5.3.17}$$

vadinama *Kendalo $\tau_b$ koreliacijos koeficientu.*

**5.3.3 pastaba.** Jeigu imtyse sutampančių reikšmių nėra, tai

$$\sum_{i=1}^n \sum_{j=1}^n U_{ij}^2 = \sum_{i=1}^n \sum_{j=1}^n V_{ij}^2 = n(n-1),$$

taigi $\tau_b = \tau_a = r_K$.

**5.3.4 pastaba.** Jeigu imtyse yra $k_X$ ir $k_Y$ sutampančių elementų grupių, o sutampančių elementų skaičiai $s$-oje ir $r$-oje grupėse yra atitinkamai $u_s$ ir $v_r$, tai Kendalo $\tau_b$ koreliacijos koeficientas gali būti užrašytas tokiu pavidalu

$$\tau_b = \frac{\sum_{i=1}^n \sum_{j=1}^n U_{ij} V_{ij}}{[(n(n-1) - \sum\limits_{s=1}^{k_X} u_s(u_s-1))(n(n-1) - \sum\limits_{r=1}^{k_Y} v_r(v_r-1))]^{1/2}}, \tag{5.3.18}$$

nes $s$-oje grupėje yra $u_s(u_s-1)$ porų, kurioms $U_{ij} = 0$. Todėl

$$\sum_{i=1}^n \sum_{j=1}^n U_{ij}^2 = n(n-1) - \sum_{s=1}^{k_X} u_s(u_s-1).$$

Analogišką lygybę gauname antrajai imčiai.

Taigi, jei sutampančių reikšmių yra, tai $|\tau_b| > |\tau_a|$, kadangi skaitikliai sutampa, o vardiklis $\tau_b$ išraiškoje yra mažesnis. Jeigu $X_i = Y_i$ su visais $i$, tai $U_{ij} = V_{ij}$ ir $\tau_b = 1$.

Asimptotinis nepriklausomumo kriterijus sudaromas aproksimuojant statistikos

$$S = \sum_{i<j} U_{ij} V_{ij}$$

skirstinį normaliuoju $N(0, V_S)$; čia

$$V_S = \frac{\nu_0 - \nu_u - \nu_v}{18} + \frac{\nu_{uv1}}{2n(n-1)} + \frac{\nu_{uv2}}{9n(n-1)(n-2)},$$

$$\nu_0 = n(n-1)(2n+5), \quad \nu_u = \sum_{s=1}^{k_X} u_s(u_s-1)(2u_s+5), \quad \nu_v = \sum_{r=1}^{k_Y} v_r(v_r-1)(2v_r+5),$$

$$\nu_{uv1} = \sum_{s=1}^{k_X} u_s(u_s-1) \sum_{r=1}^{k_Y} v_r(v_r-1),$$

$$\nu_{uv2} = \sum_{r=1}^{k_X} u_s(u_s-1)(u_s-2) \sum_{r=1}^{k_Y} v_r(v_r-1)(v_r-2).$$

**Asimptotinis Kendalo nepriklausomumo kriterijus:** nepriklausomumo hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|\frac{S}{\sqrt{VS}}| > z_{\alpha/2}. \tag{5.3.19}$$

**5.3.5 pastaba.** Dviejų a. d. koreliacija kartais vertinama sudarant vadinamąjį *Gudmano ir Kruskalo gama koeficientą*:

$$\gamma = \frac{\sum\limits_{A_{ij}=1} A_{ij} - \sum\limits_{A_{ij}=-1} A_{ij}}{\sum\limits_{A_{ij}\neq 0} A_{ij}}.$$

Matome, kad jis yra suderintų ir nesuderintų porų skaičių skirtumo ir nesutampančių porų skaičiaus santykis. Šis koeficientas taip pat sutampa su Kendalo koreliacijos koeficientu $r_K$, kai sutampančių stabėjimų imtyse nėra.

**5.3.2 pavyzdys.** (pavyzdžio **5.3.1** tęsinys.) Pagal pavyzdžio **5.3.1** duomenis patikrinsime nepriklausomumo hipotezę naudodami Kendalo asimptotinį kriterijų.

Sumuodami $U_{ij}V_{ij}$ gauname

$$\sum_{i=1}^n \sum_{j=1}^n U_{ij} V_{ij} = 2S = 630.$$

Pirmoje imtyje yra $k_X = 12$ sutampančių stebinių grupių: 6 grupės po 2 sutampančius stebėjimus, 3 – po 3, 2 – po 4, 1 grupė iš 5 sutampančių elementų. Antroje imtyje turime $k_Y = 11$ grupių: 3 grupes po 2 sutampančius elementus, 2 – po 3, 3 – po 4, 1 – iš 5, 1 – iš 6 ir 1 – iš 7 sutampančių elementų. Taigi

$$\sum_{s=1}^{k_X} u_s(u_s-1) = 6\cdot 2 + 3\cdot 6 + 2\cdot 12 + 1\cdot 20 = 74, \quad \sum_{r=1}^{k_Y} v_r(v_r-1) = 146,$$

$$\tau_b = \frac{630}{[(50\cdot 49 - 74)(50\cdot 49 - 74)]^{1/2}} = 0,26926.$$

$$\nu_0 = 50\cdot 49\cdot 105 = 257250, \quad \nu_u = 6\cdot 2\cdot 1\cdot 9 + 3\cdot 3\cdot 2\cdot 11 + 2\cdot 4\cdot 3\cdot 13 + 1\cdot 5\cdot 4\cdot 15 = 918, \quad \nu_v = 2262.$$

$$\nu_{uv1} = 74\cdot 146 = 10804, \quad \nu_{uv2} = 125\cdot 474 = 59724.$$

Statistikos $S$ dispersijos įvertis

$$V_S = \frac{257250 - 918 - 2262}{18} + \frac{10804}{2\cdot 50\cdot 49} + \frac{59724}{9\cdot 50\cdot 49\cdot 48} = 14117,26.$$

Gauname:

$$\frac{S}{\sqrt{V_S}} = 2,65116 \quad pv_a = 2(1-\Phi(2,65116)) = 0,00802$$

Kaip ir Spirmeno nepriklausomumo kriterijus, Kendalo nepriklausomumo kriterijus atmeta nepriklausomumo hipotezę, nes P reikšmė yra maža.

**5.3.6 pastaba.** Nors koeficientai $r_S$ ir $r_K$ apibrėžiami skirtingai, tačiau jie yra glaudžiai susiję. Tegu

$$I_n^* = \frac{3}{n+1}\sum_{i<j}(j-i)h_{ij}$$

yra svertinė inversijų suma, gauta atsižvelgiant į atstumus tarp rangų. Tada įrodoma (žr. 5.1 pratimą), kad

$$I_n^* = \frac{3}{2(n+1)}\sum_i (R_i - i)^2, \quad r_S = 1 - \frac{4I_n^*}{n(n-1)}.$$

Palyginę su (5.3.8) matome, kad $r_S$ ir $r_K$ skiriasi tik tuo, kad pirmame naudojama svertinė inversijų suma, o antrajame – tiesiog inversijų suma.

Pirsono koreliacijos koeficientas tarp $r_S$ ir $r_K$ (žr. 5.2 pratimą) yra

$$\rho(r_S,\ r_K) = \frac{2(n+1)}{\sqrt{2n(2n+5)}}.$$

Jis mažėja nuo 1, kai $n = 2$, iki 0,98, kai $n = 5$, paskui monotoniškai didėja iki 1, kai $n \to \infty$. Taigi šios statistikos labai mažai skiriasi ir jais grindžiami kriterijai asimptotiškai ekvivalentūs, o praktiškai ekvivalentūs ir su baigtiniais $n$.

### 5.3.3. Nepriklausomumo kriterijų ASE

Lygindami parametrinius kriterijus 1.6 skyrelyje įvedėme asimptotinio santykinio efektyvumo (ASE) sąvoką. Jis apibūdina kriterijaus galios elgesį hipotetinės parametro reikšmės aplinkoje. Tiksliau, didinant imtį, nagrinėjamas kriterijaus galios elgesys, kai alternatyvų seka tam tikru greičiu artėja prie hipotetinės parametro reikšmės.

Apie neparametrinio kriterijaus efektyvumą sprendžiame lygindami jį su parametriniu kriterijumi tai pačiai hipotezei tikrinti, dažniausiai esant normalumo prielaidai. Jeigu lygindami neparametrinį kriterijų su TG ar TGN parametriniu kriterijumi gauname ASE artimą 1, pirmenybę reikėtų teikti bendresniam neparametriniam kriterijui.

Rasime kriterijų, grindžiamų ranginiais koreliacijos koeficientais, ASE atžvilgiu kriterijaus, grindžiamo Pirsono empiriniu koreliacijos koeficientu, normaliojo skirstinio atveju.

Tarkime, nepriklausomi vienodai pasiskirstę a. v. $(X_i,\ Y_i)^T$, $i = 1, ..., n$, turi dvimatį normalųjį skirstinį su koreliacijos koeficientu $\rho$.

Normaliojo skirstinio atveju stebimi a. d. $X$ ir $Y$ yra nepriklausomi tada ir tik tada, kai Pirsono koreliacijos koeficientas $\rho = 0$. Pirsono nepriklausomumo kriterijus grindžiamas empiriniu koreliacijos koeficientu

$$\hat{\rho} = r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}. \tag{5.3.20}$$

Jeigu $\rho = 0$, tai a. d. $r$ tikimybių tankio funkcija yra

$$f_r(x) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})} (1 - x^2)^{\frac{n-4}{2}}, \quad -1 < x < 1.$$

Apibrėžę a. d.

$$U = \sqrt{n-2} \frac{r}{\sqrt{1 - r^2}} \tag{5.3.21}$$

įsitikiname, kad jo tankio funkcija yra

$$g(u) = \frac{1}{\sqrt{(n-2)\pi}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})} (1 + \frac{u^2}{n-2})^{-(n-1)/2},$$

t. y. a. d. $U$ turi Stjudento skirstinį su $n - 2$ laisvės laipsniais.

**Nepriklausomumo kriterijus grindžiamas empiriniu Pirsono koreliacijos koeficientu:** nepriklausomumo hipotezė, kai alternatyvos yra $H_1 : \rho > 0$, $H_2 : \rho < 0$ ir $H_3 : \rho \neq 0$, atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai atitinkamai teisingos nelygybės

$$U > t_\alpha(n-2), \quad U < -t_\alpha(n-2), \quad |U| > t_{\alpha/2}(n-2),$$

**5.3.3 teorema.** *Tarkime, nepriklausomi vienodai pasiskirstę a. v.* $(X_i,\ Y_i)^T$, $i = 1, ..., n$, *turi dvimatį normalųjį skirstinį su koreliacijos koeficientu* $\rho$. *Tada kriterijaus, grindžiamo Kendalo ranginiu koreliacijos koeficientu, ASE kriterijaus, grindžiamo Pirsono empiriniu koreliacijos koeficientu, atžvilgiu normaliojo skirstinio atveju yra*

$$e(r_K, r) = \frac{9}{\pi^2} \approx 0,912.$$

**Įrodymas.** Pažymėkime

$$\mu_1(\rho) = \mathbf{E}_\rho(r_K), \quad \sigma_1^2(0) = \lim_{n\to\infty} n\boldsymbol{V}_0(r_K),$$

$$\mu_2(\rho) = \lim_{n\to\infty} \mathbf{E}_\rho(r), \quad \sigma_2^2(0) = \lim_{n\to\infty} n\boldsymbol{V}_0(r).$$

Taikant formulę (1.6.1) (čia $\delta = 1/2$) reikia turėti $\mu_1'(0)$, $\mu_2'(0)$, $\sigma_1(0)$, $\sigma_2(0)$.

Kadangi $r_K = 1 - 4I_n/(n(n-1))$, tai ieškant $\mu_1(\rho)$ reikia rasti inversijų skaičiaus $I_n$ vidurkį, kai teisinga alternatyva. Inversijų skaičių $I_n$ galima užrašyti šitaip:

$$I_n = \sum_{i<j} \tilde{h}_{ij}, \quad \tilde{h}_{ij} = \frac{1}{2}\{1 - \operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j)\}.$$

Iš tikrųjų

$$\tilde{h}_{ij} = 1 \iff (X_i - X_j)(Y_i - Y_j) < 0 \iff$$

$$(R_{1i} - R_{1j})(R_{2i} - R_{2j}) < 0 \iff (k-l)(R_k - R_l) < 0;$$

čia $k = R_{2i}, l = R_{2j}$. Sąlyga $(k-l)(R_k - R_l) < 0$ ekvivalenti inversijos buvimui.

Sumoje yra $n(n-1)/2$ vienodai pasiskirsčiusių dėmenų, todėl

$$\mathbf{E}I_n = \frac{n(n-1)}{2}\mathbf{E}(\tilde{h}_{ij}),$$

$$\mu_1(\rho) = \mathbf{E}_\rho r_K = \mathbf{E}_\rho(1 - \frac{4I_n}{n(n-1)}) = 1 - 2\mathbf{E}_\rho(\tilde{h}_{ij}) = \mathbf{E}_\rho(\operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j)).$$

Kadangi n. a. v. $(X_i,\ Y_i)^T$, $i = 1, ..., n$, turi dvimatį normalųjį skirstinį su koreliacijos koeficientu $\rho$, tai vektorius $(U, V)^T$ su koordinatėmis $U = X_i - X_j$ ir $V = Y_i - Y_j$ taip pat turi dvimatį normalųjį skirstinį su nuliniu vidurkiu ir koreliacijos koeficientu $\rho$. Taigi

$$\mu_1(\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \operatorname{sign} u \operatorname{sign} v\ \phi(u,\ v;\ \rho)du\,dv$$

$$= 2\int_0^\infty \left[\int_0^\infty - \int_{-\infty}^0\right] \phi(u,\ v;\ \rho)dv\,du.$$

Integralai nepriklauso nuo $U$ ir $V$ dispersijų (atlikus keitimus $x = u/\sigma_1$ ir $y = v/\sigma_2$, integralų reikšmės nepakinta), todėl galima imti standartinį dvimatį normalųjį skirstinį, kurio tankis:

$$\phi(u,\ v;\ \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[u^2 - 2\rho uv + v^2]\}$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(v-\rho u)^2}{2(1-\rho^2)}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}.$$

Gauname

$$\mu_1(\rho) = 2\int_0^\infty \left[1 - 2\Phi\left(\frac{-\rho u}{\sqrt{1-\rho^2}}\right)\right] \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du, \quad \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

Diferencijuodami pagal $\rho$ turime

$$\mu_1'(\rho) = \frac{2}{\pi^2(1-\rho^2)^{3/2}}\int_0^\infty ue^{-\frac{u^2}{2(1-\rho^2)}}du = \frac{2}{\pi\sqrt{1-\rho^2}}, \quad \mu_1'(0) = \frac{2}{\pi}.$$

Kita vertus, empirinis koreliacijos koeficientas $r$ asimptotiškai turi normalųjį skirstinį

$$\sqrt{n}(r-\rho) \xrightarrow{d} Z \sim N(0,(1-\rho^2)^2), \tag{5.3.22}$$

taigi

$$\mu_2(\rho) = \rho, \quad \mu_2'(0) = 1.$$

Jeigu $\rho = 0$, tai naudodami statistikos $r_K$ dispersijos formulę (5.3.9) ir atsižvelgę į konvergavimą (5.3.22), gauname

$$\sigma_1^2(0) = \lim_{n\to\infty} nV_0(r_K) = \frac{4}{9}, \quad \sigma_2^2(0) = 1.$$

Pagal (1.6.1) kriterijaus, grindžiamo Kendalo koreliacijos koeficientu ASE atžvilgiu kriterijaus, grindžiamo Pirsono empiriniu koreliacijos koeficientu normaliojo skirstinio atveju yra

$$e(r_K, r) = \frac{(\frac{2}{\pi})^2 : \frac{4}{9}}{1:1} = \frac{9}{\pi^2}.$$

▲

## 5.3.4. Normaliųjų žymių kriterijus

Įrodėme, kad kai a. v. $(X,Y)^T$ yra normalusis, tai kriterijaus, grindžiamo Kendalo koreliacijos koeficientu (arba jam ekvivalenčiu kriterijumi, grindžiamu Spirmeno ranginiu koreliacijos koeficientu) ASE atžvilgiu kriterijaus, grindžiamo Pirsono empiriniu koreliacijos koeficientu, yra artimas 1. Kyla klausimas, ar galima rasti ranginį kriterijų, kuriam minėtas ASE būtų lygus 1?

Atsakymas yra teigiamas. Sudarant tokį kriterijų stebiniai keičiami ne rangais, o specialiai parinktomis rangų funkcijomis.

Stebėjimų poras $(X_i, Y_i)^T$ surikiuokime taip, kad $Y_1, ..., Y_n$ būtų išdėstyti didėjančia tvarka. Paskui eilutes $Y_1, ..., Y_n$ ir $X_1, ..., X_n$ pakeiskime jų rangų eilutėmis $(1, \dots, n)$ ir $(R_1, \dots, R_n)$. Pažymėkime

$$E_i = \mathbf{E}(U_{(i)}), \quad i = 1, 2, ..., n, \tag{5.3.23}$$

standartinio normaliojo skirstinio $U \sim N(0,\ 1)$ pozicinių statistikų $U_{(i)}$ paprastoje didumo $n$ imtyje $(U_1, ..., U_n)^T$ vidurkius, kurie vadinami *normaliosiomis žymėmis*. Jų reikšmės yrs tabuliuotos.

Vietoje Spirmeno koreliacijos koeficiento, t. y. empirinio korreliacijos koeficiento tarp $(1, \dots, n)$ ir $(R_1, \dots, R_n)$, apibrėžkime empirinį koreliacijos koeficientą tarp $(1, \dots, n)$ ir $(E_{R_1}, \dots, E_{R_n})$ :

$$r_{ns} = \frac{\sum_{i=1}^n (E_{R_i} - \bar{E})(i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^n (E_{R_i} - \bar{E})^2 \sum_{i=1}^n (i - \frac{n+1}{2})^2}} = \frac{\frac{1}{n}\sum_{i=1}^n iE_{R_i} - \frac{n+1}{2}\bar{E}}{\sqrt{\frac{n^2-1}{12}\frac{1}{n}\sum_{i=1}^n (E_{R_i} - \bar{E})^2}}. \tag{5.3.24}$$

**Normaliųjų žymių kriterijus:** nepriklausomumo hipotezė atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$r_{ns} \leq c_1, \quad \text{arba} \quad r_{ns} \geq c_2;$$

čia $c_1$ yra maksimali, o $c_2$ – minimali galima statistikos $r_{ns}$ reikšmės, atitinkamai tenkinančios nelygybes

$$\mathbf{P}\{r_{ns} \leq c_1\} \leq \alpha/2, \quad \mathbf{P}\{r_{ns} \geq c_1\} \leq \alpha/2.$$

Įrodoma, kad jei a. v. $(X, Y)^T$ turi dvimatį normalųjį skirstinį, tai nepriklausomumo kriterijaus, grindžiamo statistika $r_{ns}$, ASE kriterijaus, grindžiamo Pirsono empiriniu koreliacijos koeficientu, atžvilgiu lygus 1.

## 5.4. Ranginiai atsitiktinumo kriterijai

Tarkime, kad imties $\mathbf{X} = (X_1, ..., X_n)^T$ koordinatės yra nepriklausomi absoliučiai tolydūs a. d. Pažymėkime $F_i(x)$ atsitiktinio dydžio $X_i$ pasiskirstymo funkciją.

**Atsitiktinumo hipotezė:**

$$H_0 : F_1(x) \equiv F_2(x) \equiv \cdots \equiv F_n(x).$$

Ši hipotezė tvirtina, kad $\mathbf{X}$ yra paprastoji imtis a. d. $X$, t. y. kad a. d. $X_1, ..., X_n$ yra vienodai pasiskirstę.

Jei hipotezė $H_0$ teisinga, tai imties $(X_1, ..., X_n)^T$ rangų vektoriaus $(R_1, \ldots, R_n)^T$ ir jo koordinačių skirstiniai pateikti 5.3.1 skyrelyje:

$$\mathbf{P}\{(R_1, ..., R_n) = (j_1, ..., j_n)\} = \frac{1}{n!}, \quad \mathbf{P}\{R_i = j\} = \frac{1}{n},$$

su bet kokiu aibės $(1, ..., n)$ kėliniu $(j_1, ..., j_n)$ ir su bet kokiais $i, j = 1, \ldots n$.

Kai teisingos monotoninės alternatyvos

$$\bar{H}_1 : F_1(x) \leq F_2(x) \leq \cdots \leq F_n(x),$$

arba

$$\bar{H}_2 : F_1(x) \geq F_2(x) \geq \cdots \geq F_n(x),$$

kurios reiškia didėjantį arba mažėjantį trendą, rangai turės tendenciją išsidėstyti atitinkamai didėjančia arba mažėjančia tvarka. Taigi rango $R_i$ skirstinys priklauso nuo eksperimento numerio $i$.

### 5.4.1. Kendalo ir Spirmeno atsitiktinumo kriterijai

Apskaičiuokime Spirmeno arba Kendalo ranginius koreliacijos koeficientus $r_S$ arba $r_K$ naudodami vektorių $(1, 2, \ldots, n)^T$ ir rangų vektorių $(R_1, \ldots, R_n)^T$. Jeigu hipotezė $H_0$ teisinga, šių koeficientų skirstiniai yra tokie patys kaip ir skyrelyje 5.3.2 arba skyrelyje 5.3.3, t. y. jų reikšmės turės tendenciją įgyti artimas 0 reikšmes.

Jeigu teisingos alternatyvos $\bar{H}_1$ arba $\bar{H}_2$, tai rangų vektorius $(R_1, \ldots, R_n)^T$ turės tendenciją panašėti atitinkamai į vektorių $(1, \ldots, n)^T$ arba vektorių $(n, \ldots, 1)^T$. Taigi statistikos $r_S$ ir $r_K$ turės tendenciją įgyti reikšmes, artimas $+1$ arba $-1$.

**Spirmeno (Kendalo) atsitiktinumo kriterijus:** atsitiktinumo hipotezė $H_0$, kai alternatyva dvipusė $\bar{H}_1 \cup \bar{H}_2$, atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $|r_S|$ (atitinkamai $|r_K|$) viršija statistikos $r_S$ (atitinkamai $r_K$) $\alpha/2$ kritinę reikšmę.

Rasime Kendalo atsitiktinumo kriterijaus ASE optimalaus parametrinio kriterijaus atžvilgiu normaliojo skirstinio ir specialios trendo alternatyvos atveju.

Tarkime, kad esant teisingai alternatyvai imties elementas $X_i$ nusakytas tiesiniu regresijos modeliu:

$$X_i = \beta_0 + i\beta + e_i, \quad i = 1, 2, ..., n; \tag{5.4.1}$$

čia $e_i$ yra vienodai pasiskirstę n. a. d. ir $e_i \sim N(0,\ 1)$.

Šiame modelyje atsitiktinumo hipotezė ekvivalenti parametrinei hipotezei $H : \beta = 0$.

Iš regresinės analizės žinoma, kad TGN kriterijus šiai hipotezei tikrinti grindžiamas parametro $\beta$ įvertiniu

$$\hat{\beta} = \frac{\sum_i (X_i - \bar{X})(i - \bar{i})}{\sum_i (i - \bar{i})^2} \sim N\left(\beta,\ \frac{12}{n(n^2-1)}\right). \tag{5.4.2}$$

**5.4.1 teorema.** *Kendalo atsitiktinumo kriterijaus ASE atžvilgiu kriterijaus, grindžiamo įvertiniu $\hat{\beta}$, yra*

$$e(r_K, \hat{\beta}) = \left(\frac{3}{\pi}\right)^{1/3} \approx 0,985.$$

**Įrodymas.** Pagal (5.4.2) gauname:

$$n^{3/2}(\hat{\beta} - \beta)/\sqrt{12} \xrightarrow{d} Y \sim N(0,\ 1).$$

Todėl ASE išraiškoje (1.6.1) dydžiai, atitinkantys parametrinį kriterijų, grindžiamą įvertiniu $\hat{\beta}$, yra

$$\mu_2(\beta) = \beta, \quad \sigma_2(\beta) = \sqrt{12}, \quad \mu_2'(0) = 1, \quad \sigma_2(0) = \sqrt{12}, \quad \delta = 3/2.$$

Nagrinėkime kriterijų, grindžiamą Kendalo koreliacijos koeficientu

$$r_K = 1 - 4I_n/(n(n-1))$$

(čia $I_n$ inversijų skaičius).

Reikia rasti $I_n$ vidurkį, kai teisinga alternatyva (5.4.1). Gauname

$$\mathbf{E}_\beta(I_n) = \mathbf{E}_\beta(\sum_{i<j} h_{ij}) = \sum_{i<j} \mathbf{E}_\beta(h_{ij}).$$

Pagal (5.4.1) modelį skirtumas $X_i - X_j \sim N(\beta(i-j),\ 2)$. Taigi

$$\mathbf{E}_\beta(h_{ij}) = \mathbf{P}_\beta\{X_i - X_j > 0\} =$$

$$= \frac{1}{2\sqrt{\pi}} \int_0^\infty \exp\{-\frac{1}{4}(t-\beta(i-j))^2\}dt = 1 - \Phi\left(\frac{-\beta(i-j)}{\sqrt{2}}\right).$$

Kadangi

$$\frac{\partial}{\partial\beta}(\mathbf{E}_\beta(h_{ij}))|_{\beta=0} = \varphi(0)\frac{i-j}{\sqrt{2}} = \frac{i-j}{2\sqrt{\pi}},$$

o

$$\frac{\partial}{\partial\beta}\mathbf{E}_\beta(I_n)|_{\beta=0} = \frac{1}{2\sqrt{\pi}}\sum_{i<j}(i-j) = -\frac{n(n^2-1)}{12\sqrt{\pi}},$$

Tai norėdami, kad $\mu_1'(0)$ nepriklausytų nuo $n$, imsime statistiką $r_K^* = r_K/(n+1)$ (kuri ekvivalenti statistikai $r_K$). Tada gauname

$$\mu_1'(0) = \frac{\partial}{\partial\beta}\mathbf{E}_\beta(r_K^*)|_{\beta=0} = \frac{1}{3\sqrt{\pi}}. \tag{5.4.3}$$

Teoremoje 5.3.3 gauta

$$\mathbf{V}_\beta(I_n)|_{\beta=0} = \frac{n(n-1)(2n+5)}{72},$$

taigi

$$\sigma_1^2(0) = \lim_{n\to\infty} \mathbf{Var}_\beta(n^{3/2}r_k^*)|_{\beta=0} = \lim_{n\to\infty} \frac{n^3}{(n+1)^2}\frac{16}{n^2(n-1)^2}\frac{n(n-1)(2n+5)}{72} = \frac{4}{9}.$$

Gavome

$$\mu_1'(0) = \frac{1}{3\sqrt{\pi}}, \quad \sigma_1(0) = \frac{2}{3}.$$

Kadangi nagrinėjamu atveju $\delta = 3/2$ (žr.(1.6.1)), tai kriterijaus, grindžiamo koreliacijos koeficientu $r_K$, ASE atžvilgiu kriterijaus, grindžiamo įvertiniu $\hat{\beta}$, yra

$$e(r_K, \hat{\beta}) = \left(\frac{\frac{1}{3\sqrt{\pi}}\frac{3}{2}}{\frac{1}{\sqrt{12}}}\right)^{2/3} = \left(\frac{3}{\pi}\right)^{1/3}.$$

**5.4.1 pavyzdys.** Lentelėje pateikiama tam tikroje vietovėje pamatuota 2008 metų lapkričio mėnesio paros temperatūros deviacija.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Diena | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Deviacija | 12 | 13 | 12 | 11 | 5 | 2 | −1 | 2 | −1 | 3 |
| Diena | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Deviacija | 2 | −6 | −7 | −7 | −12 | −9 | 6 | 7 | 10 | 6 |
| Diena | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Deviacija | 1 | 1 | 3 | 7 | −2 | −6 | −6 | −5 | −2 | −1 |

Ar neprieštarauja šie duomenys prielaidai, kad buvo stebimi vienodai pasiskirstę n. a. d.?

Surastas rangų reikšmes pateikiame lentelėje.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Diena | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Rangai | 28,5 | 30 | 28,5 | 27 | 21 | 17 | 12 | 17 | 12 | 19,5 |
| Diena | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Rangai | 17 | 6 | 3,5 | 3,5 | 1 | 2 | 22,5 | 24,5 | 26 | 22,5 |
| Diena | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Rangai | 14,5 | 14,5 | 19,5 | 24,5 | 9,5 | 6 | 6 | 8 | 9,5 | 12 |

Naudodami SPSS paketą gauname: $r_S = -0,425$, $r_K = -0,321$. Atitinkamos $P$ reikšmės yra $pv = 0,01424$ ir $pv = 0,01932$. Atsitiktinumo hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija $0,01932$.

### 5.4.2. Bartelio ir Neimano atsitiktinumo kriterijus

Kitas atsitiktinumo kriterijus grindžiamas gretimų rangų skirtumais. Jeigu trendas egzistuoja, tai gretimų stebėjimų rangų skirtumai turi tendenciją įgyti mažesnes reikšmes. *Bartelio ir Neimano ranginio atsitiktinumo kriterijaus statistika* turi tokį pavidalą:

$$T_{BN} = \sum_{i=1}^{n-1} (R_{i+1} - R_i)^2. \tag{5.4.4}$$

Ši statistika įgyja reikšmes nuo $n-1$ iki $(n-1)(n^2+n-3)/3$, kai $n$ yra lyginis, ir nuo $n-1$ iki $[(n-1)(n^2+n-3)/3]-1$, kai $n$ nelyginis.

**Bartelio ir Neimano ranginis kriterijus.** Atsitiktinumo hipotezė atmetama $\alpha$ lygmens kriterijumi, kai $T_{BN} \leq c$; čia $c$ yra maksimalus skaičius, tenkinantis sąlygą $\mathbf{P}\{T_{BN} \leq c\} \leq \alpha$.

Kai $4 \leq n \leq 10$, tai $P$ reikšmės $pv = \mathbf{P}\{T_{BN} \leq c\}$ su įvairiomis statistikos $T_{BN}$ galimomis realizacijomis $c$ yra tabuliuotos (žr., pvz., [13]).

Didesniems $n$ paprastai naudojama normuota statistika

$$\bar{T}_{BN} = \frac{T_{BN}}{\sum_{i=1}^{n} (R_i - (n+1)/2)^2}.$$

Jeigu visi rangai skirstingi, tai šios statistikos vardiklis yra lygus $n(n^2-1)/12$.

Kai $10 \leq n \leq 100$, kritinės reikšmės gaunamos aproksimuojant $\bar{T}_{BN}$ skirstinį beta skirstiniu. Šios aproksimacijos pagrindu sudarytas lenteles taip pat galima rasti knygoje [13].

Kai $n > 100$, statistikos $\bar{T}_{BN}$ skirstinį rekomenduojama aproksimuoti normaliuoju skirstiniu $N(2, \sigma_n^2)$; čia

$$\sigma_n^2 = \frac{4(n-2)(5n^2-2n-9)}{5n(n+1)(n-1)^2}).$$

**Asimptotinis Bartelio ir Neimano kriterijus.** Atsitiktinumo hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$Z_n = \frac{\bar{T}_{BN} - 2}{\sigma_n} \leq -z_\alpha$$

.

**5.4.2 pavyzdys.** (**5.4.1** pavyzdžio tęsinys). Patikrinsime atsitiktinumo hipotezę pagal **5.4.1** pavyzdžio duomenis naudodami Bartelio ir Neimano kriterijų.

Rangai ir jų skirtumai pateikiai lentelėje.

| Diena | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Rangai | 28, 5 | 30 | 28, 5 | 27 | 21 | 17 | 12 | 17 | 12 | 19, 5 |
| Skirtumai | | 1, 5 | −1, 5 | −1, 5 | −6 | −4 | −5 | 5 | −5 | 7, 5 |
| Diena | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Rangai | 17 | 6 | 3, 5 | 3, 5 | 1 | 2 | 22, 5 | 24, 5 | 26 | 22, 5 |
| Skirtumai | −2, 5 | −11 | −2, 5 | 0 | −2, 5 | 1 | 20, 5 | 2 | 1, 5 | −3, 5 |
| Diena | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Rangai | 14, 5 | 14, 5 | 19, 5 | 24, 5 | 9, 5 | 6 | 6 | 8 | 9, 5 | 12 |
| Skirtumai | −8 | 0 | 5 | 5 | −15 | −3, 5 | 0 | 2 | 1, 5 | 2, 5 |

Gauname:

$$T_{BN} = \sum_{i=1}^{29}(R_{i+1} - R_i)^2 = 1133,25, \quad \sum_{i=1}^{30}(R_i - 15,5)^2 = 2238,$$

$$\bar{T}_{BN} = 1133,25/2238 = 0,506367.$$

Knygos [13] lentelėse randame, kad reikšmingumo lygmens $0,005$; $0,01$; $0,05$; $0,1$ kritinės reikšmės yra atitinkamai $1,11$, $1,19$, $1,41$, $1,54$. Kadangi gautoji statistikos $\bar{T}_{BN}$ yra gerokai mažesnė, hipotezė atmetama. Rasime asimptotinę $P$ reikšmę taikydami aproksimaciją normaliuoju skirstiniu. Statistika $Z_n$ įgijo reikšmę $-4,1928$ ir $pv_a = \Phi(-4,1928) = 0,0000138$. Hipotezė atmetama.

Matome, kad šiame pavyzdyje Bartelio ir Neimano kriterijus pasirodė kur kas galingesnis už Spirmeno ar Kendalo kriterijų. Tai, matyt, gali būti aiškinama tuo, kad Bartelio ir Neimano atsitiktinumo kriterijus yra tinkamesnis, kai alternatyvos nėra monotoninės. Pavyzdžiui, kažkuriuo momentu trendas pakinta iš teigiamo į neigiamą arba atvirkščiai.

# 5.5. Ranginiai homogeniškumo kriterijai

Tegu $\mathbf{X} = (X_1, ..., X_m)^T$ ir $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ yra dvi nepriklausomos paprastosios imtys, gautos stebint absoliučiai tolydžius a. d. $X \sim F(x)$ ir $Y \sim G(x)$. Reikia patikrinti hipotezę, kad pasiskirstymo funkcijos sutampa:

$$H_0 : F(x) = \mathbf{P}\{X \leq x\} \equiv \mathbf{P}\{Y \leq x\} = G(x). \tag{5.5.1}$$

## 5.5.1. Vilkoksono (Mano,Vitnio ir Vilkoksono) kriterijus

Tarkime, kad alternatyva yra *poslinkio*: egzistuoja toks $\theta \neq 0$, kad su visais $x \in \mathbf{R}$ teisinga hipotezė

$$H_1 : G(x) = F(x - \theta). \tag{5.5.2}$$

**Vilkoksono ranginio kriterijaus statistika.** Pažymėkime $R_1, R_2, ..., R_m$ stebėjimų $X_1, ..., X_m$ rangus jungtinėje didumo $m+n$ imtyje $(X_1, ..., X_m, Y_1, ..., Y_n)^T$. Tada Vilkoksono kriterijaus statistika $W$ yra lygi šių rangų sumai:

$$W = \sum_{i=1}^{m} R_i.$$

**Vilkoksono kriterijaus statistikos skirstinys.** Jeigu hipotezė $H_0$ teisinga, tai statistikos $W$ skirstinys nepriklauso nuo nežinomų parametrų, o priklauso tik nuo imčių didumų $m$ ir $n$, nes remiantis (5.2.2)

$$\mathbf{P}\{(R_1, ..., R_m) = (j_1, ..., j_m)\} = \frac{n!}{(m+n)!} \tag{5.5.3}$$

su kiekvienu vektoriumi $(j_1, ..., j_m)^T$, susidedančiu iš $m$ skirtingų aibės $\{1, 2, ..., m+n\}$ elementų. Statistikos $W$ minimali reikšmė

$$\omega_1 = 1 + ... + m = m(m+1)/2,$$

o maksimali

$$\omega_2 = (n+1) + ... + (n+m) = m(2n+m+1)/2.$$

Taigi su kiekvienu $k = \omega_1, ..., \omega_2$

$$\mathbf{P}\{W = k\} = N_k \frac{n!}{(m+n)!},$$

čia $N_k$ yra skaičius pirmiau aprašytų vektorių $(j_1, ..., j_m)^T$, tenkinančių sąlygą $j_1 + ... + j_m = k$.

**Mano ir Vitnio kriterijaus statistika.** Pažymėkime $U$ skaičių tokių atvejų, kai pirmosios imties elementai viršija antrosios imties elementus:

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} h_{ij}, \quad h_{ij} = \begin{cases} 1, & \text{kai } X_i > Y_j, \\ 0, & \text{kai } X_i < Y_j. \end{cases} \tag{5.5.4}$$

Statistikos $W$ ir $U$ yra glaudžiai susijusios. Iš tikrųjų, tegu $i_1, \ldots, i_m$ yra pirmosios imties $(X_1, \ldots, X_m)^T$ elementų indeksai, tenkinantys nelygybes $R_{i_1} < \cdots < R_{i_m}$. Tada prieš elementą $X_{i_l}$ sujungtoje surikiuotoje imtyje yra $R_{i_l} - 1$ elementas, iš jų $l-1$ pirmosios ir $R_{i_l} - l$ antrosios imties elementų. Taigi

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} h_{ij} = \sum_{l=1}^{m} (R_{i_l} - l) = W - m(m+1)/2. \tag{5.5.5}$$

Vilkoksono kriterijus grindžiamas statistika $W$, o Manio ir Vitnio – statistika $U$. Kadangi statistikos skiriasi tik konstanta, abu kriterijai yra ekvivalentūs.

Kai teisinga alternatyva $\theta > 0$, tai antrosios imties elementai turės tendenciją įgyti didesnes reikšmes negu pirmosios imties elementai, taigi rangų suma $W$ turės tendenciją įgyti mažesnes reikšmes. Atvirkščiai, kai $\theta < 0$, statistika $W$ turės tendenciją įgyti didesnes reikšmes.

**Vilkoksono kriterijus:** kai alternatyva dvipusė, atveju hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$W \leq c_1 \quad \text{arba} \quad W \geq c_2;$$

čia $c_1$ yra maksimalus, o $c_2$ minimalus skaičiai, tenkinantys nelygybes

$$\sum_{k=w_1}^{c_1} \mathbf{P}\{W = k|H_0\} \leq \alpha/2. \quad \sum_{i=c_2}^{w_2} \mathbf{P}\{W = k|H_0\} \leq \alpha/2.$$

Kai $m$ ir $n$ nedideli, statistikos $W$ kritinės reikšmės yra tabuliuotos (žr. [7]).

Kai alternatyvos vienpusės, ( $\theta > 0$ arba $\theta < 0$), kritinė sritis yra vienpusė, t. y. turi atitinkamai pavidalą $W \geq d$ arba $W \leq c$; kritinės reikšmės $c$ ir $d$ randamos analogiškai kaip $c_1$ ir $c_2$ pakeičiant $\alpha/2$ į $\alpha$.

**Didelių imčių atvejis.** Jeigu $m$ ir $n$ yra dideli, tai statistikos $W$ skirstinys aproksimuojamas normaliuoju. Tegu $N = m + n$. Remiantis (5.2.3) rangų sumos $W$ vidurkis ir dispersija yra

$$\mathbf{E}(W) = \frac{m(N+1)}{2}, \quad \mathbf{V}(W) = \sum_{j=1}^{m} \mathbf{V}(R_j) + \sum\sum\nolimits_{i \neq j} \mathbf{cov}(R_i, R_j) =$$

$$m\frac{N^2-1}{12} - m(m-1)\frac{N+1}{12} = \frac{mn(N+1)}{12}. \tag{5.5.6}$$

Pažymėkime

$$Z_{m,n} = \frac{W - \mathbf{E}(W)}{\sqrt{\boldsymbol{V}(W)}} = \frac{U - \mathbf{E}(U)}{\sqrt{\boldsymbol{V}(U)}}.$$

**5.5.1 teorema.** *Jeigu stebimų a. d. $X$ ir $Y$ skirstiniai absoliučiai tolydūs, $N \to \infty$, $m/N \to p \in (0, 1)$, tai esant teisingai hipotezei $H_0$*

$$Z_{m,n} \xrightarrow{d} Z \sim N(0, 1).$$

**Įrodymas**. Tegu $S_N$ yra inversijų skaičius jungtinėje didumo $N = m + n$ imtyje. Jis gaunamas atliekant $N(N+1)/2$ porinių visų stebėjimų palyginimų. Jeigu iš $S_N$ atimsime inversijų skaičius $S'_m$ ir $S''_n$, kurie gaunami lyginant atitinkamai pirmosios ir antrosios imties elementus, tai liks tik inversijų skaičius $U$, gautas lyginant pirmosios imties elementus su antrosios imties elementais:

$$S_N = S'_m + S''_n + U, \quad W = U + \frac{m(m+1)}{2} = S_N - S'_m - S''_n + \frac{m(m+1)}{2}. \tag{5.5.7}$$

Kai hipotezė teisinga, a. d. $S'_m, S''_n$ ir $U$ yra nepriklausomi. Remiantis teorema 5.2.2 a, d. $S_N, S'_m, S''_n$ asimptotiškai normalieji. Gauname

$$\frac{S_N - \mathbf{E}S_N}{\sqrt{\boldsymbol{V}S_N}} = \frac{S'_m + S''_n - \mathbf{E}(S'_m + S''_n)}{\sqrt{\boldsymbol{V}(S'_m + S''_n)}}\sqrt{\frac{\boldsymbol{V}(S'_m + S''_n)}{\boldsymbol{V}S_N}} + \frac{U - \mathbf{E}U}{\sqrt{\boldsymbol{V}U}}\sqrt{\frac{\boldsymbol{V}U}{\boldsymbol{V}S_N}}. \tag{5.5.8}$$

Kadangi

$$\frac{\boldsymbol{V}(S'_m + S''_n)}{\boldsymbol{V}S_N} \to 1 - 3pq, \quad \frac{\boldsymbol{V}U}{\boldsymbol{V}S_N} \to 3pq, \quad q = 1 - p,$$

tai pirmasis (5.5.8) dešinės lygybės pusės narys artėja į a. d. $V_1 \sim N(0,\, 1-3pq)$, kairioji pusė – į a. d. $V \sim N(0,\, 1)$, todėl antrasis dešinės lygybės pusės narys – į a. d. $V_2 \sim N(0, 3pq)$. Taigi

$$Z_{m,n} = \frac{U - \mathbf{E}U}{\sqrt{\mathbf{V}U}} \xrightarrow{d} Z \sim N(0,\, 1).$$

▲

Nustatyta, kad konvergavimas į normalųjį skirstinį gana greitas.

**Asimptotinis Vilkoksono kriterijus:** jeigu $m$ ir $n$ nėra maži, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|Z_{m,n}| > z_{\alpha/2}.$$

**Sutampančios reikšmės.** Jei yra sutampančių reikšmių, tai pagal 5.2.2 teoremą

$$\mathbf{V}(W) = \sum_{j=1}^{m} \mathbf{V}(R_j) + \sum\sum\nolimits_{i \neq j} \mathbf{cov}(R_i, R_j) = m(\frac{N^2-1}{12} - \frac{\mathbf{E}T}{12N}) +$$

$$m(m-1)(-\frac{N+1}{12} + \frac{\mathbf{E}T}{12N(N-1)}) = \frac{mn(N+1)}{12}(1 - \frac{\mathbf{E}T}{N^3 - N});$$

čia

$$T = \sum_{i=1}^{k} T_i, \quad T_i = (t_i^3 - t_i),$$

$k$ yra skaičius sutampančių elementų grupių, o $t_i$ yra $i$-osios grupės didumas.

Taigi, kai $m$ ir $n$ nėra maži ir yra sutampančių reikšmių, statistika $Z_{m,n}$ modifikuojama:

$$Z^*_{m,n} = \frac{Z_{m,n}}{\sqrt{1 - T/(N^3 - N)}}.$$

**Modifikuotas asimptotinis Vilkoksono kriterijus:** hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|Z^*_{m,n}| > z_{\alpha/2}.$$

**5.5.1 pavyzdys** ( **4.5.1** pavyzdžio tęsinys). Naudodami **4.5.1** pavyzdžio duomenis Vilkoksono, Mano ir Vitnio kriterijais tikrinsime hipotezę, kad fungicidų naudojimas neturi įtakos kavos medelių sergamumui.

Imčių didumai $m = n = 7$, jungtinės imties didumas $N = m + n = 14$. Pateikiame didėjimo tvarka surikiuotus duomenis (imties numeris nurodytas skliausteliuose).

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 0,75(1) | 1,76(1) | 2,46(1) | 4,88(1) | 5,10(1) | 5,68(2) | 5,68(2) |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 6.01(1) | 7,13(1) | 11,63(2) | 16,30(2) | 21,46(2) | 33,30(2) | 44,20(2) |

Rangai:

| | | | | | | |
|---|---|---|---|---|---|---|
| 1(1) | 2(1) | 3(1) | 4(1) | 5(1) | 6,5(2) | 6,5(2) |
| 8(1) | 9(1) | 10(2) | 11(2) | 12(2) | 13(2) | 14(2) |

Pirmosios imties rangų suma (Vilkoksono statistika) yra

$$W = 1 + 2 + 3 + 4 + 5 + 8 + 9 = 32.$$

Antrosios imties rangų suma $N(N+1)/2 - W = 105 - 32 = 73$ yra gerokai didesnė, todėl tikėtina, kad fungicidų naudojimas sumažina medelių sergamumą.

Mano ir Vitnio statistika

$$U = W - \frac{m(m+1)}{2} = 32 - \frac{7 \cdot 8}{2} = 4.$$

Randame

$$\mathbf{E}W = m(N+1)/2 = (7 \cdot 15)/2 = 52,5, \quad \boldsymbol{V}W = mn(N+1)/12 = (7 \cdot 7 \cdot 15)/12 = 61,25,$$

ir

$$Z_{m,n} = \frac{W - m(N+1)/2}{\sqrt{mn(N+1)/12}} = \frac{32 - 52,5}{\sqrt{61,25}} \approx -2,619394.$$

Yra viena pora sutampančių rangų: $k = 1$, $t_1 = 2$, $T_1 = (2^3 - 2) = 6$ ir

$$1 - \frac{\sum_{i=1}^{k} T_i}{n^3 - n} = 1 - \frac{6}{14^3 - 14} = 0,99782.$$

Modifikuotosios statistikos reikšmė yra

$$Z^*_{m,n} = Z_{m,n}/\sqrt{0,99782} \approx -2,6223.$$

Atlikdami skaičiavimus SPSS programų paketu gauname $P$ reikšmę $pv = 0,006410$. Asimptotinė $P$ reikšmė yra

$$pv_a = 2(1 - \Phi(2,6223)) \approx 0,008734.$$

Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,00641.

Šiame pavyzdyje Vilkoksono kriterijus labiau atskiria turimas imtis negu dviejų imčių Kolmogorovo ir Smirnovo kriterijus ir yra palyginamas su dviejų imčių Kramero ir Mizeso kriterijumi.

Šiame pratime natūralu tikrinti nulinę hipotezę su alternatyva, kad pesticidų naudojimas sumažina medelių sergamumą, t. y. su vienpuse alternatyva:

$$H_1: \quad \exists\, \theta < 0: \quad G(x) = F(x - \theta) \quad \text{su visais} \quad x \in \mathbf{R}.$$

Tokiu atveju hipotezė atmetama, kai $W \le c$. Hipotezė atmetama asimptotiniu kriterijumi, kai $Z^*_{m,n} < -z_\alpha$. Skaičiuodami SPSS paketu gauname $P$ reikšmę $pv = 0,003205$. Asimptotinė $P$ reikšmė yra

$$pv_a = \Phi(-2,6223)) \approx 0,004367.$$

Hipotezė atmetama.

Matome, kad ir su mažais imčių didumais, $P$ reikšmių aproksimavimo paklaida, palyginus, nedidelė.

## 5.5.2. Vilkoksono kriterijaus galia

Rasime Vilkoksono kriterijaus galią, kai imtys didelės. Kaip ir anksčiau, nagrinėsime poslinkio alternatyvą $\bar{H}$ (žr. (5.5.2)). Homogeniškumo hipotezė ekvivalenti parametrinei hipotezei $H_0: \theta = 0$ dėl poslinkio parametro reikšmės.

Pažymėkime $f(x)$ ir $f(x - \theta)$ a. d. $X$ ir $Y$ tankio funkcijas.

Jei teisinga alternatyva, tai (5.5.4) apibrėžtų a. d. $h_{ij}$ skirstiniai randami pagal pilnosios tikimybės formulę:

$$p_1(\theta) = \mathbf{P}\{h_{11} = 1\} = \mathbf{P}\{X_1 > Y_1\} = \int_{-\infty}^{\infty} F(x - \theta) f(x)\,dx,$$

$$p_2(\theta) = \mathbf{P}\{h_{11} = 1,\ h_{12} = 1\} = \mathbf{P}\{X_1 > Y_1, X_1 > Y_2\} = \int_{-\infty}^{\infty} F^2(x - \theta) f(x)\,dx,$$

$$p_3(\theta) = \mathbf{P}\{h_{11}=1, h_{21}=1\} = \mathbf{P}\{X_1 > Y_1, X_2 > Y_1\} = \int_{-\infty}^{\infty} [1-F(x)]^2 f(x-\theta)dx. \tag{5.5.9}$$

Pasinaudoję (5.5.5) išraiška, gauname

$$\mu(\theta) = \mathbf{E}(U) = mn\, p_1(\theta),$$

$$\sigma^2(\theta) = \boldsymbol{V}(U) = mn\boldsymbol{V}(h_{11}) + mn(n-1)\mathbf{cov}(h_{11}, h_{12}) + nm(m-1)\mathbf{cov}(h_{11}, h_{21})$$

$$= mn\,[p_1(\theta) - p_1^2(\theta) + (n-1)(p_2(\theta) - p_1^2(\theta)) + (m-1)(p_3(\theta) - p_1^2(\theta))].$$

Jeigu $m$ ir $n$ yra dideli, tai pagal CRT a. d. $(U-\mu(\theta))/\sigma(\theta)$ skirstinys aproksimuojamas normaliuoju. Taigi gauname kriterijaus galios aproksimaciją:

$$\beta(\theta) = \mathbf{P}_\theta\{|\frac{U-\mu(0)}{\sigma(0)}| > z_{\alpha/2}\} = \mathbf{P}_\theta\{\frac{U-\mu(\theta)}{\sigma(\theta)} > \frac{\mu(0)-\mu(\theta)+\sigma(0)z_{\alpha/2}}{\sigma(\theta)}\}$$

$$+\mathbf{P}_\theta\{\frac{U-\mu(\theta)}{\sigma(\theta)} < \frac{\mu(0)-\mu(\theta)-\sigma(0)z_{\alpha/2}}{\sigma(\theta)}\}$$

$$\approx 1 - \Phi\left(\frac{\mu(0)-\mu(\theta)+\sigma(0)z_{\alpha/2}}{\sigma(\theta)}\right) + \Phi\left(\frac{\mu(0)-\mu(\theta)-\sigma(0)z_{\alpha/2}}{\sigma(\theta)}\right).$$

Reikia pažymėti, kad funkcijos $p_1, p_2$ ir $p_3$ bei galia priklauso ne tik nuo parametro $\boldsymbol{\theta}$, bet ir nuo pasiskirstymo funkcijos $F$.

Specialioms funkcijų $F$ klasėms galima gauti aproksimacines galios išraiškas, kurios priklauso tik nuo vienmačio parametro.

Tarkime, kad funkcija $F(x)$ priklauso tik nuo mastelio ir poslinkio parametrų, t. y. priklauso pasiskirstymo funkcijų šeimai

$$\{F_0((x-\mu)/\sigma), \mu \in \mathbf{R}, \sigma > 0\};$$

čia $F_0(y)$ yra žinoma funkcija. Tegu $f_0$ yra tankio funkcija esant teisingai hipotezei $H : \eta = 0$; $\eta = \mu/\sigma$. Šiuo atveju funkcijos $p_1, p_2$ ir $p_3$ priklauso tik nuo parametro $\eta$:

$$p_1(\eta) = \int_{-\infty}^{\infty} F_0(y-\eta) f_0(y)dy, \quad p_2(\eta) = \int_{-\infty}^{\infty} F_0^2(y-\eta) f_0(y)dy,$$

$$p_3(\eta) = \int_{-\infty}^{\infty} [1-F_0(y)]^2 f_0(y-\eta)dy.$$

Taigi

$$\beta(\eta) \approx 1 - \Phi\left(\frac{\mu(0)-\mu(\eta)+\sigma(0)z_{\alpha/2}}{\sigma(\eta)}\right) + \Phi\left(\frac{\mu(0)-\mu(\eta)-\sigma(0)z_{\alpha/2}}{\sigma(\eta)}\right).$$

### 5.5.3. Vilkoksono kriterijaus ASE Stjudento kriterijaus atžvilgiu

Esant poslinkio alternatyvai teisinga lygybė $\mathbf{E}Y_j = \mathbf{E}X_i + \theta$, o homogeniškumo hipotezė, būdama ekvivalenti hipotezei $H_0 : \theta = 0$, yra ekvivalenti dviejų imčių vidurkių lygybės hipotezei $H_0 : \mu_1 = \mu_2$; čia $\mu_1 = \mathbf{E}X_i$, $\mu_2 = \mathbf{E}Y_i$. Hipotezei tikrinti gali būti naudojamas asimptotinis Stjudento kriterijus.

**Asimptotinio Stjudento kriterijaus statistika.**

$$\text{Kai } H_0, \text{ tai } X_i \sim F, Y_j \sim F \text{ ir } \mathbf{E}X_i = \mathbf{E}Y_j, \quad \boldsymbol{V}X_i = \boldsymbol{V}Y_j := \tau^2,$$

$$\mathbf{E}(\bar{X} - \bar{Y}) = 0, \quad \boldsymbol{V}(\bar{X} - \bar{Y}) = \tau^2(\frac{1}{m} + \frac{1}{n}).$$

Pažymėkime

$$S^2 = ((m-1)s_1^2 + (n-1)s_2^2)/(m+n-2);$$

čia $s_1^2$ ir $s_2^2$ – nepaslinktieji dispersijos įvertiniai, atitinkamai surasti pagal pirmąją ir antrąją imtį.

**5.5.2 teorema.** *Jeigu* $N = m + n \to \infty$, $m/N \to p \in (0, 1)$, *tai*

$$t = \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \xrightarrow{d} Z \sim N(0, 1). \tag{5.5.10}$$

**Asimptotinis Stjudento kriterijus:** jeigu $m$ ir $n$ yra dideli, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $|t| > z_{\alpha/2}$.

Kai skirstinys normalusis, t. y. kai $F_0 = \Phi$, statistika $t$ naudojama ir mažoms imtims. Kai hipotezė teisinga, ji turi Stjudento skirstinį su $m + n - 2$ laisvės laipsnių. Hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $|t| > t_{\alpha/2}(m + n - 2)$.

Rasime Vilkoksono kriterijaus ASE Stjudento kriterijaus atžvilgiu.

**5.5.3 teorema.** *Jeigu* $N \to \infty$, $m/N \to p \in (0, 1)$, *tai Vilkoksono kriterijaus ASE Stjudento kriterijaus atžvilgiu yra*

$$e(W, t) = 12\tau^2 \left[\int_{-\infty}^{\infty} f^2(x)dx\right]^2, \quad \tau^2 = \boldsymbol{V}(X_i). \tag{5.5.11}$$

**Įrodymas.** Stjudento kriterijaus statistika $t$ asimptotiškai ($N \to \infty$, $m/N \to p \in (0, 1)$) ekvivalenti normuotai statistikai $\bar{X} - \bar{Y}$, nes $S \xrightarrow{P} \tau$. Kai teisinga poslinkio alternatyva $H_1$, tai

$$\frac{\sqrt{N}(\bar{X} - \bar{Y} + \theta)}{\tau\sqrt{1/p + 1/q}} \xrightarrow{d} Z \sim N(0\,1), \quad q = 1 - p.$$

Taigi funkcijos, atitinkančios Stjudento kriterijų, ASE formulėje (1.6.1) turi tokį pavidalą:

$$\mu_2(\theta) = -\theta, \quad \sigma_2(\theta) = \tau\sqrt{1/p + 1/q} = \frac{\tau}{\sqrt{pq}},$$

$$\mu_2'(0) = -1, \quad \sigma_2(0) = \frac{\tau}{\sqrt{pq}}.$$

Nagrinėjant Vilkoksono kriterijaus galią buvo gauta, kad

$$(U - \mu(\theta))/\sigma(\theta) \xrightarrow{d} Z \sim N(0, 1).$$

Statistika $U^* = U/(mn)$ ekvivalenti statistikai $U$, todėl:

$$\frac{\sqrt{N}(U^* - p_1(\theta))}{\sqrt{(p_2(\theta) - p_1^2(\theta))/q + (p_3(\theta) - p_1^2(\theta))/p}} \xrightarrow{d} Z \sim N(0, 1).$$

Taigi funkcijos, atitinkančios Vilkoksono kriterijų, ASE formulėje (1.6.1) turi tokį pavidalą:

$$\mu_1(\theta) = p_1(\theta), \quad \sigma_1^2(\theta) = (p_2(\theta) - p_1^2(\theta))/q + (p_3(\theta) - p_1^2(\theta))/p.$$

Diferencijuodami pagal $\theta$ gauname

$$\mu_1'(\theta) = -\int_{-\infty}^{\infty} f(x-\theta)f(x)dx, \quad \mu_1'(0) = -\int_{-\infty}^{\infty} f^2(x)dx.$$

Kadangi $p_1(0) = 1/3$, $p_2(0) = p_3(0) = 1/3$, tai $\sigma_1^2(0) = \frac{1}{12pq}$.

Remiantis (1.6.1) Vilkoksono kriterijaus ASE Stjudento kriterijaus atžvilgiu yra

$$e(W, t) = \left( \frac{-\int_{-\infty}^{\infty} f^2(x)dx\, 2\sqrt{3pq}}{(-1)\frac{\sqrt{pq}}{\tau}} \right)^2 = 12\tau^2 \left[ \int_{-\infty}^{\infty} f^2(x)dx \right]^2$$

▲

**Skirstiniai, priklausantys nuo poslinkio ir mastelio parametrų.** Jeigu funkcija $F(x)$ priklauso šeimai

$$\{F_0((x-\mu)/\sigma), \mu \in \mathbf{R}, \sigma > 0\};$$

čia $F_0(y)$ yra žinoma funkcija, tai ASE nepriklauso nuo nežinomų parametrų:

$$e(W, t) = 12\tau_0^2 \left[ \int_{-\infty}^{\infty} f_0^2(y)dy \right]^2;$$

čia

$$\tau_0^2 = \boldsymbol{V}((X_i - \mu)/\sigma) = \int_{-\infty}^{\infty} y^2 dF_0(y) - (\int_{-\infty}^{\infty} y dF_0(y))^2,$$

nes

$$\int_{-\infty}^{\infty} f^2(x)dx = \sigma^{-1}\int_{-\infty}^{\infty} f_0^2(x)dx, \quad \tau^2 = \sigma^2\tau_0^2.$$

Pateiksime keletą pavyzdžių.

1) Normalusis skirstinys: $F_0 = \Phi$, $f_0 = \varphi$, $\tau_0^2 = 1$,

$$\int_{-\infty}^{\infty} \varphi^2(y)dy = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\{-x^2\}dx = \frac{1}{2\sqrt{\pi}},$$

taigi

$$e(W,t) = \frac{3}{\pi} \approx 0,95.$$

2) Tolygusis skirstinys:

$$F_0(x) = \begin{cases} 0, & \text{kai } x \le -1, \\ (x+1)/2, & \text{kai } x \in (-1.1), \\ 1, & \text{kai } x \ge 1, \end{cases}$$

$$f_0(x) = \frac{1}{2}\mathbf{1}_{(-1,\,1)}(x), \quad \tau_0^2 = 1/3, \quad \int_{-\infty}^{\infty} f_0^2(x)dx = \frac{1}{2},$$

taigi

$$e(W,t) = 1.$$

3) Logistinis skirstinys:

$$F_0(x) = \frac{1}{1+e^{-x}}, \quad f_0(x) = \frac{e^{-x}}{(1+e^{-x})^2}, \quad \tau_0^2 = \frac{\pi^2}{3},$$

$$\int_{-\infty}^{\infty} f_0^2(x)dx = \int_{-\infty}^{\infty} \frac{e^{-2x}}{(1+e^{-x})^4}dx = \int_0^{\infty} \frac{y}{(1+y)^4}dx = \frac{1}{6},$$

taigi

$$e(W,t) = \frac{\pi^2}{9} \approx 1,097.$$

4) Ekstremalių reikšmių skirstinys:

$$F_0(x) = 1 - e^{-e^x}, \quad f_0(x) = e^x e^{-e^x}, \quad \tau_0^2 = \frac{\pi^2}{6},$$

$$\int_{-\infty}^{\infty} f_0^2(x)dx = \frac{1}{4}\int_0^{\infty} ye^{-y}dy = \frac{1}{4},$$

taigi

$$e(W,t) = \frac{\pi^2}{8} \approx 1,23.$$

5) Dvipusis eksponentinis (Laplaso) skirstinys:

$$F_0(x) = \begin{cases} \frac{1}{2}e^x, & \text{kai } x \le 0, \\ 1 - \frac{1}{2}e^{-x}, & \text{kai } x > 0, \end{cases}$$

$$f_0(x) = \frac{1}{2}e^{-|x|}, \quad \tau_0^2 = 2,$$

$$\int_{-\infty}^{\infty} f_0^2(x)dx = \frac{1}{2}\int_0^{\infty} e^{-2x}dx = \frac{1}{4},$$

taigi

$$e(W,t) = 2.$$

**5.5.1 pastaba.** Pateikti pavyzdžiai rodo, kad kai kurioms šeimoms Vilkoksono kriterijus yra efektyvesnis už Stjudento kriterijų. Maža to, integralas formulėje (5.5.11) gali įgyti ir begalines reikšmes (žr. 5.8 pratimą). Taigi reikšmė $e(W,t) = \infty$ taip pat yra galima.

Rasime skirstinį, su kuriuo Vilkoksono kriterijaus ASE atžvilgiu Stjudento kriterijaus yra minimali, ir tą minimumą.

**5.5.4 teorema.** *Vilkoksono kriterijaus ASE atžvilgiu Stjudento kriterijaus minimumas yra* $\inf_f e(W,t) = 0,864$.

**Įrodymas.** Pagal (5.5.11) pakanka minimizuoti

$$\mathbf{E}(f(X)) = \int_{-\infty}^{\infty} f^2(x)dx;$$

čia $X$ – absoliučiai tolydusis a. d., kurio tikimybių tankis $f(x)$ ir dispersija $\tau^2 = 1$. Kadangi statistikos $W$ ir $t$ nepakinta, jei a. d. $X_i$ ir $Y_j$ pakeičiame a. d. $X_i + \mu$ ir $Y_j + \mu$, čia $\mu = \mathbf{E}X_i$, tai nemažindami bendrumo galime imti $\mu = 0$. Remiantis (5.5.11) reikia minimizuoti integralą

$$\mathbf{E}(f(X)) = \int_{-\infty}^{\infty} f^2(x)dx$$

su sąlygomis

$$\int_{-\infty}^{\infty} f(x)dx = 1, \int_{-\infty}^{\infty} x^2 f(x)dx = 1. \tag{5.5.12}$$

Naudojant Lagranžo neapibrėžtinių daugiklių metodą reikia minimizuoti integralą

$$\int_{-\infty}^{\infty} [f(x) - \lambda_1 - \lambda_2 x^2] f(x)dx.$$

Kadangi $f(x)$ yra neneigiama, tai šis integralas įgyja minimalią reikšmę, kai

$$f(x) = \begin{cases} \lambda_1 + \lambda_2 x^2, & \text{kai } \lambda_1 + \lambda_2 x^2 \geq 0, \\ 0, & \text{kai } \lambda_1 + \lambda_2 x^2 < 0. \end{cases}$$

Įstatę šią funkcijos išraišką į (5.5.12), gauname dviejų lygčių sistemą neapibrėžtiniams daugikliams $\lambda_1$ ir $\lambda_2$ rasti. Gauname

$$\lambda_1 = \frac{3}{4\sqrt{5}}, \quad \lambda_2 = -\frac{3}{20\sqrt{5}}, \quad f(x) = \begin{cases} \frac{3}{4\sqrt{5}} - \frac{3}{20\sqrt{5}}x^2, & \text{kai } |x| \leq \sqrt{5}, \\ 0, & \text{kai } |x| > \sqrt{5}, \end{cases}$$

$$[\int_{-\sqrt{5}}^{\sqrt{5}} \left( \frac{3}{4\sqrt{5}} - \frac{3}{20\sqrt{5}} x^2 \right)^2 dx]^2 = \frac{9}{125}.$$

Pagal (5.5.11)

$$\inf_f e(W, t) = 12 \frac{9}{125} = 0,864. \qquad \blacktriangle$$

**5.5.2 pastaba.** Atlikta analizė rodo, kad normaliojo skirstinio atveju, kai imtys pakankamai didelės, pakeičiant Stjudento kriterijų Vilkoksono ranginiu kriterijumi prarandama apie 5 % stebėjimų. Pačiu nepalankiausiu atveju prarandamų stebėjimų procentas neviršija 14 %.

Kita vertus, kai kuriems skirstiniams Vilkoksono kriterijus yra efektyvesnis (žr. pavyzdžius 3),4) ir pratimą 5.8). Taigi, jeigu nesame įsitikinę, kad stebimi a. d. yra normalieji, tai pirmenybę, matyt, reikėtų teikti Vilkoksono kriterijui.

### 5.5.4. Van der Vardeno kriterijus

Matėme, kad kai skirstinys normalusis, Vilkoksono ranginio homogeniškumo kriterijaus ASE Stjudento kriterijaus atžvilgiu apytiksliai lygus $0,95$. Kyla klausimas, ar galima rasti ranginį kriterijų, kurio ASE, palyginti su Stjudento kriterijumi, būtų lygus 1? Atsakymas yra teigiamas (žr. [29]), jeigu vietoje rangų sumų $W = \sum_{i=1}^m R_i$ imsime specialiai parinktų rangų funkcijų sumas:

$$V = v(R_1) + \cdots + v(R_m), \quad v(r) = \Phi^{-1}\left( \frac{r}{N+1} \right). \tag{5.5.13}$$

Atsitiktinio dydžio $V$ skirstinys yra simetriškas 0 atžvilgiu. Iš tikrųjų, remiantis $\Phi^{-1}(z) = -\Phi^{-1}(1-z)$, gaunama

$$-V = -\sum_{r=1}^m \Phi^{-1}\left( \frac{r}{N+1} \right) = \sum_{r=1}^m \Phi^{-1}\left( \frac{N+1-r}{N+1} \right).$$

Kai hipotezė teisinga, a. v. $(R_1, \ldots, R_m)$ skirstinys sutampa su skirstiniu a. v. $(N+1-R_1, \ldots, N+1-R_m)$, taigi ir a. d. $V$ ir $-V$ skirstiniai yra vienodi.

Kriterijus, grindžiamas statistika $V$, vadinamas Van der Vardeno kriterijumi.

**Van der Vardeno kriterijus:** kai alternatyva dvipusė, homogeniškumo hipotezė atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $|V| > c$; čia $c$ yra mažiausias skaičius, tenkinantis nelygybę $\mathbf{P}\{|V| > c | H_0\} \leq \alpha/2$.

Jeigu $m$ ir $n$ nedideli, tai statistikos $V$ kritinės reikšmės yra tabuliuotos (žr. [7]). Jų reikšmes taip pat galima rasti naudojant matematinės statistikos programų paketus (SAS, SPSS).

Kai $N \to \infty, m/N \to p \in (0, 1)$, tai statistikos $V$ skirstinys aproksimuojamas normaliuoju $N(0, \sigma_V^2)$; čia

$$\sigma_V^2 = \frac{mnQ}{N-1}, \quad Q = \frac{1}{N} \sum_{r=1}^N v^2(r). \tag{5.5.14}$$

Pažymėkime $Z_{m,n} = V/\sigma_V$.

**Asimptotinis Van der Vardeno kriterijus:** jeigu $m$ ir $n$ yra dideli, tai homogeniškumo hipotezė, kai alternatyva dvipusė, atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, jei $|Z_{m,n}| > z_{\alpha/2}$.

**5.5.2 pavyzdys.** (**4.5.1** ir **5.5.1** pavyzdžio tęsinys.) Pagal pratimo **4.5.1** duomenis patikrinsime homogeniškumo hipotezę naudodami Van der Vardeno kriterijų.

Atlikdami analizę SAS programų paketu gauname $V = -4,1701$ ir $pv = 0,0052, pv_a = 0,0097$. Hipotezė atmestina.

## 5.5.5. Ranginiai dviejų imčių homogeniškumo kriterijai, kai alternatyva yra mastelio

Tarkime, kad homogeniškumo hipotezės alternatyva yra mastelio:

$$H_1 : G(x) = F(\frac{x}{\sigma}), \quad \sigma > 0. \tag{5.5.15}$$

Jeigu $\sigma > 1$ $(0 < \sigma < 1)$, tai a. d. $Y$ sklaida yra didesnė (mažesnė) už a. d. $X$ sklaidą. Atsitiktinių dydžių $X$ ir $Y$ medianos sutampa.

**5.5.3 pastaba.** Jeigu a. d. $X$ ir $Y$ įgyjamų reikšmių sritis yra $(0, \infty)$, tai, atlikus transformacijas

$$X_i^* = \ln X_i, \quad Y_j^* = \ln Y_j,$$

a. d. $X_i^*$ ir $Y_j^*$ pasiskirstymo funkcijos įgyja tokį pavidalą

$$F^*(x) = F(e^x) \quad \text{ir} \quad G^*(x) = F(\frac{e^x}{\sigma}) = F(e^{x-\ln\sigma}) = F^*(x-\theta),$$

čia $\theta = \ln\sigma$. Taigi transformuotų dydžių alternatyva tampa poslinkio, todėl imtims, sudarytoms iš elementų $\ln X_i$ ir $\ln Y_j$, galima taikyti ankstesnio skyrelio kriterijus.

**5.5.4 pastaba.** Jeigu a. d. $X$ ir $Y$ įgyjamų reikšmių sritis yra $\mathbf{R}$, tai Vilkoksono ir Van der Vardeno kriterijai nėra efektyvūs, kai homogeniškumo hipotezės alternatyva yra mastelio.

Iš tikrųjų, tarkime, kad $F(x) = F_0((x-\mu)/\tau)$. Tada, kai teisinga alternatyva, $G(x) = F_0((x-\mu)/(\sigma\tau))$ ir a. d. $X - \mu$ ir $(Y-\mu)/\sigma$ skirstiniai sutampa. Jeigu $\sigma > 1 (0 < \sigma < 1)$, tai a. d. $X$ stebiniai turi tendenciją koncentruotis arčiau $\mu$ negu $Y$ stebiniai, t. y. pirmosios imties elementai bendroje variacinėje eilutėje turės tendenciją koncentruotis viduryje (atitinkamai abiejuose bendros variacinės eilutės galuose). Taigi statistikos $W$ ar kitos panašios statistikos įgyjamos reikšmės gali būti nei labai didelės, nei labai mažos, lygiai kaip ir esant teisingai hipotezei. Todėl šiomis statistikomis grindžiami kriterijai gali neskirti hipotezės nuo alternatyvos. Tokiu atveju parenkamos specialios rangų funkcijos.

**Kriterijų sudarymo idėja.** Bendroje variacinėje eilutėje $r$-jam elementui priskirkime reikšmę $s(r)$; čia $s$ yra aibėje $(1, 2, ..., N)$ apibrėžta funkcija. Apibrėžkime statistiką

$$S = s(R_1) + \cdots + s(R_m); \tag{5.5.16}$$

čia, kaip ir Vilkoksono kriterijaus atveju, $R_i$ yra pirmosios imties narių rangai bendroje variacinėje eilutėje.

Natūralu mažiausias $s$ reikšmes priskirti patiems mažiausiems ir patiems didžiausiems, o didžiausias – viduriniams variacinės eilutės nariams. Esant teisingai alternatyvai $\sigma > 1$ $(0 < \sigma < 1)$ pirmosios imties elementai koncentruosis variacinės eilutės viduryje (atitinkamai galuose), todėl suma $S$ įgyja dideles (atitinkamai mažas) reikšmes. Kai hipotezė teisinga, suma $S$ įgyja vidutines reikšmes. Taigi hipotezė atmetama vienpusės alternatyvos $\sigma > 1$ $(0 < \sigma < 1)$ naudai, kai $S > c_2$ $(S < c_1)$; čia $c_1$ ir $c_2$ yra statistikos $S$ kritinės reikšmės esant teisingai hipotezei. Panašiai, jei alternatyva yra dvipusė, hipotezė atmetama, kai $S < c_1^*$ arba $S > c_2^*$. Taip sudaromi *Zygelio ir Tjukio* bei *Ansario ir Bredlio* kriterijai.

Alternatyviai galima didžiausias $s$ reikšmes priskirti patiems mažiausiems ir patiems didžiausiems, o vidutines reikšmes – viduriniams variacinės eilutės nariams. Tada kritinėse srityse nelygybių ženklai pakeičiami priešingais. Taip sudaromi *Mūdo* ir *Klotso* kriterijai.

**Funkcijos $s$ parinkimas**:

1. Zygelio ir Tjukio kriterijus:

$$s(1) = 1,\ s(N) = 2,\ s(N-1) = 3,\ s(2) = 4,$$
$$s(3) = 5,\ s(N-2) = 6,\ s(N-3) = 7,\ s(4) = 8, \cdots \tag{5.5.17}$$

2. Ansario ir Bredlio kriterijus:

$$s(1) = 1,\ s(N) = 1,\ s(2) = 2,\ s(N-1) = 2. \cdots \tag{5.5.18}$$

3. Mūdo kriterijus:

$$s(r) = \left(r - \frac{N+1}{2}\right)^2. \tag{5.5.19}$$

4. Klotso kriterijus:

$$s(r) = \left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2. \tag{5.5.20}$$

Kai teisinga homogeniškumo hipotezė, pirmieji du Zygelio ir Tjukio, Ansario ir Bredlio, Mūdo, Klotso statistikų momentai yra:

$$\mathbf{E}S_{ZT} = m(N+1)/2, \quad \boldsymbol{V}S_{ZT} = mn(N+1)/12,$$

$$\mathbf{E}S_{AB} = m(N+1)/4, \quad \boldsymbol{V}S_{AB} = mn(N+1)^2/(48N),$$
$$\mathbf{E}S_M = m(N^2-1)/12, \quad \boldsymbol{V}S_M = mn(N+1)(N^2-4)/180,$$
$$\mathbf{E}S_K = \frac{m}{N}\sum_{i=1}^{N}\left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^2,$$
$$\boldsymbol{V}S_K = \frac{mn}{N(N-1)}\sum_{i=1}^{N}\left[\Phi^{-1}\left(\frac{i}{N+1}\right)\right]^4 - \frac{n}{m(N-1)}[\mathbf{E}S_K]^2.$$

Kai teisinga hipotezė, Zygelio ir Tjukio statistikos skirstinys sutampa su Vilkoksono statistikos skirstiniu. Be to, ši statistika asimptotiškai ekvivalenti Ansario ir Bredlio statistikai.

**Didelės imtys.** Kai hipotezė teisinga, visos statistikos asimptotiškai normaliosios:

$$Z_{m,n} = \frac{S - \mathbf{E}S}{\sqrt{\boldsymbol{V}S}} \xrightarrow{d} Z \sim N(0,\ 1), \quad \text{kai} \quad n \to \infty,\ m/n \to p \in (0,\ 1). \quad (5.5.21)$$

Kai $m$ ir $n$ yra dideli, kriterijai grindžiami statistika $Z_{m,n}$. Priklausomai nuo alternatyvų hipotezė atmetama, kai ši statistika viršija ar yra mažesnė už atitinkamas standartinio normaliojo skirstinio kritines reikšmes.

Normaliojo skirstinio atveju yra gauta šių kriterijų ASE atžvilgiu Fišerio dviejų dispersijų palyginimo kriterijaus: Ansario ir Bredlio bei Zygelio ir Tjukio kriterijų ASE yra $6/\pi^2 \approx 0,608$, Mūdo kriterijaus – $15/(2\pi^2) \approx 0,760$, Klotso kriterijaus – 1.

**5.5.3 pavyzdys.** Tam tikra televizorių elektrinių selektorių charakteristika buvo matuojama dviejų tipų prietaisais. Gautos atsitiktinių paklaidų reikšmės: $m = 10$ paklaidų $X_1, ..., X_{10}$, atliekant matavimus pirmo tipo prietaisu, ir $n = 20$ paklaidų $Y_1, ..., Y_{20}$, atliekant matavimus antro tipo prietaisu. Pateikiame gautus rezultatus (padaugintus iš 100).

a) Matuota pirmo tipo prietaisu: 2,2722; -1,1502; 0,9371; 3,5368; 2,4928; 1,5670; 0,9585; -0,6089; -1,3895;-0,5112.

b) Matuota antro tipo prietaisu: 0,6387; -1,8486; -0,1160; 0,6832; 0,0480; 1,2476; 0,3421; -1,5370; 0,6595; -0,7377; -0,0726; 0,6913; 0,4325; -0,2853; 1,8385; -0,6965; 0,0037; -0,3561; -1,9286; 0,4121.

Tardami, kad imtys gautos stebint absoliučiai tolydžius nepriklausomus a. d. $X \sim F(x)$ ir $Y \sim G(x)$ su vienodais vidurkiais $\mathbf{E}X = \mathbf{E}Y = 0$, tikrinsime hipotezę $H_0 : F(x) \equiv G(x)$ su vienpuse alternatyva $\bar{H} : G(x) \equiv F(x/\theta), \theta < 1$, kad pirmo tipo prietaisas yra mažiau tikslus.

Randame statistikų įgytas reikšmes:

$$S_{ZT} = 100, \quad S_{AB} = 52, \quad S_M = 1128,5, \quad S_K = 12,363.$$

Atlikdami analizę SAS programų paketu gauname tokias $P$ reikšmes: 0,0073; 0,0067; 0,0168; 0,0528. Remdamiesi aproksimacija (5.5.21) gauname asimptotines $P$ reikšmes: 0,0082; 0,0068; 0,0154; 0,0462. Homogeniškumo hipotezė atmestina.

Jeigu tartume, kad buvo stebėti nepriklausomi normalieji a. d. $X \sim N(0, \sigma_1^2)$ ir $Y \sim N(0, \sigma_2^2)$, tai hipotezė $H_0$ tampa parametrine dispersijų lygybės hipoteze $H_0 : \sigma_1^2 = \sigma_2^2$, kai vienpusė alternatyva yra $\bar{H} : \sigma_1^2 > \sigma_2^2$. Ši hipotezė tikrinama remiantis statistika $F = s_1^2/(s_2^2)$, čia $s_1^2$ ir $s_2^2$ yra dispersijų įvertiniai:

$$\hat{\sigma}_1^2 = s_1^2 = \frac{1}{m}\sum_{i=1}^{m} X_i^2, \quad \hat{\sigma}_2^2 = s_2^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Kai hipotezė $H_0$ yra teisinga, statistika $F$ turi Fišerio skirstinį su $m$ ir $n$ laisvės laipsnių. Hipotezė atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $F > F_\alpha(m,n)$.

Gauname: $s_1^2 = 3,2024$, $s_2^2 = 0,8978$, $F = 3,567$ ir $P$ reikšmė yra

$$pv = \mathbf{P}\{F_{m,n} > 3,567\} = 0,0075.$$

Šiame pavyzdyje parametrinio Fišerio kriterijaus $P$ reikšmė yra beveik tokia pat, kaip ir Zygelio ir Tjukio ar Ansario ir Bredlio kriterijų, o Klotso ir Mūdo kriterijai pasirodė mažiau galingi.

## 5.6. Vilkoksono ranginis ženklų kriterijus

Tarkime, $X_1, ..., X_n$ yra paprastoji imtis a. d. $X$, turinčio baigtinį antrąjį momentą, kurio pasiskirstymo funkcija $F$ priklauso absoliučiai tolydžių pasiskirstymo funkcijų aibei $\mathcal{F}$.

Pažymėkime $M$ a. d. $X$ medianą, dėl kurios reikšmių ir bus formuluojamos hipotezės.

**Hipotezė dėl medianos reikšmės:**

$$H_0 : F \in \mathcal{F},\ M = M_0;$$

čia $M_0$ – fiksuota medianos reikšmė.

**Vienpusės alternatyvos:** $H_1 : F \in \mathcal{F},\ M > M_0$ ir $H_2 : F \in \mathcal{F},\ M < M_0$.
**Dvipusė alternatyva:** $H_3 : F \in \mathcal{F},\ M \neq M_0$.

### 5.6.1. Vilkoksono ranginiai ženklų kriterijai

Vilkoksono ranginis ženklų kriterijus yra tinkamesnis, kai skirstinių šeima $\mathcal{F}$ susideda iš simetriškų skirstinių. Jeigu skirstinių šeimai priklauso ir nesimetriški skirstiniai, tai paprastas ženklų kriterijus (žr. 6.1.1 skyrelį) kartais gali būti galingesnis už Vilkoksono ranginį ženklų kriterijų.

Jeigu skirstinių šeima $\mathcal{F}$ susideda iš simetriškų skirstinių ir $n$ yra didelis, tai Vilkoksono ranginio ženklų kriterijaus konkurentas yra asimptotinis Stjudento kriterijus. Šis kriterijus naudojamas hipotezei $\tilde{H}_0 : F \in \mathcal{F},\ \mu = \mu_0$, čia $\mu = \mathbf{E}X_i$, tikrinti. Tačiau kai skirstiniai simetriški, vidurkis sutampa su mediana: $M = \mu$, todėl hipotezė $\tilde{H}_0$ yra ekvivalenti hipotezei $H_0$. Stjudento asimptotinis kriterijus grindžiamas statistika

$$t = \sqrt{n}\,\frac{\bar{X} - \mu_0}{s};$$

čia

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Kai hipotezė $\tilde{H}_0$ teisinga

$$\sqrt{n}(\bar{X} - \mu_0)/\sigma \xrightarrow{d} Z \sim N(0,\,1), \;\; s \xrightarrow{P} \sigma, \;\; t \xrightarrow{d} Z \sim N(0,\,1);\;\; \text{čia } \sigma^2 = \boldsymbol{V}X_i.$$

**Asimptotinis Stjudento kriterijus:** jei $n$ yra didelis, o alternatyva dvipusė, hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|t| > z_{\alpha/2}.$$

**Vilkoksono ranginio ženklų kriterijaus sudarymas.** Tegu $D_i = X_i - M_0$ ir $R_i$ yra elemento $|D_i|$ rangas sekoje $|D_1|, \ldots, |D_n|$, o $T^+$ ir $T^-$ – rangų, atitinkančių teigiamus ir neigiamus skirtumus $D_i$, sumos:

$$T^+ = \sum_{i:D_i>0} R_i, \quad T^- = \sum_{i:D_i<0} R_i. \tag{5.6.1}$$

Pavyzdžiui, jeigu $M_0 = 10$ ir imties realizacija yra 7, 16, 5, 8, 14, tai skirtumų $D_1, \ldots, D_5$ realizacija yra $-3$, 6, $-5$, $-2$, 4, o $|D_1|, \ldots, |D_5|$ – 3, 6, 5, 2, 4. Taigi rangai $R_1, \ldots, R_5$ įgijo reikšmes 2, 5, 4, 1, 3. Rangai, atitinkantys teigiamus ir neigiamus skirtumus $D_i$, yra 3, 5 ir 1, 2, 4 atitinkamai. Taigi $T^+ = 3 + 5 = 8$, $T^- = 1 + 2 + 4 = 7$.

Praktiškai $T^+$ ir $T^-$ yra patogu skaičiuoti taip: skirtumų $D_1, \ldots, D_5$ realizaciją $-3$, 6, $-5$,
$-2$, 4 išrikiuojame jų absoliutinių didumų didėjimo tvarka ir surašome palikdami jų ženklus: $-2$, $-3$, 4, $-5$, 6. Priskiriame rangus taip pat palikdami ženklus: $-1$, $-2$, 3, $-4$, 5. Tada $T^+ = 3 + 5 = 8$, $T^- = 1 + 2 + 4 = 7$.

Pakanka nagrinėti vieną iš statistikų $T^+$ arba $T^-$, nes jų suma $T^+ + T^- = R_1 + \cdots + R_n = n(n+1)/2$. Galimos statistikos $T^+$ reikšmės yra $0, 1, \ldots, n(n+1)/2$.

Kai hipotezė $H_0$ teisinga, skirtumai $D_i$ yra simetriškai pasiskirstę nulio atžvilgiu, todėl a. d. $T^+$ ir $T^-$ skirstiniai sutampa ir $\mathbf{E}T^+ = \mathbf{E}T^-$.

Jeigu $M > M_0$, tai a. d. $D_i$ skirstinys simetriškas taško $\theta = M - M_0 > 0$ atžvilgiu, nes su bet kokiu $c > 0$ a. d. $D_i$ įgyja reikšmes iš intervalų $(\theta - c, \theta)$ ir $(\theta, \theta + c)$ su vienodomis tikimybėmis. Taigi a. d. $D_i$ turi tendenciją dažniau įgyti teigiamas reikšmes negu neigiamas. Todėl $T^+$ turi tendenciją įgyti didesnes reikšmes už $T^-$ ir $\mathbf{E}T^+ > \mathbf{E}T^-$.

Analogiškai, jeigu $M < M_0$, tai $T^+$ turi tendenciją įgyti mažesnes reikšmes už $T^-$ ir $\mathbf{E}T^+ < \mathbf{E}T^-$.

Kadangi suma $T^+ + T^- = n(n+1)/2$ yra pastovi, tai sąlygos

$$\mathbf{E}T^+ = \mathbf{E}T^-, \quad \mathbf{E}T^+ > \mathbf{E}T^-, \quad \mathbf{E}T^+ < \mathbf{E}T^-$$

yra ekvivalenčios sąlygoms

$$\mathbf{E}T^+ = n(n+1)/4 =: N, \quad \mathbf{E}T^+ > N, \quad \mathbf{E}T^+ < N.$$

Taigi jei $M = M_0$, tai statistikos $T^+$ reikšmės koncentruojasi taško $N$ aplinkoje; jei $M > M_0$, – taško, didesnio už $N$, aplinkoje; jei $M < M_0$, – taško, mažesnio už $N$, aplinkoje. Tuo ir grindžiamas nagrinėjamas kriterijus.

**Vilkoksono ranginis ženklų kriterijus:** kai alternatyva dvipusė $H_3$, hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$T^+ \geq T^+_{\alpha/2}(n) \quad \text{arba} \quad T^+ \leq T^+_{1-\alpha/2}; \tag{5.6.2}$$

čia $T^+_{\alpha/2}(n)$ yra mažiausias skaičius, tenkinantis nelygybę $\mathbf{P}\{T^+ \geq T^+_{\alpha/2}(n)\} \leq \alpha/2$, o $T^+_{1-\alpha/2}$ yra didžiausias skaičius, tenkinantis nelygybę $\mathbf{P}\{T^+ \leq T^+_{1-\alpha/2}\} \leq \alpha/2$. Vienpusių alternatyvų $H_1$ arba $H_2$ atveju hipotezė $H_0$ atmetama, kai

$$T^+ \geq T^+_{\alpha}(n) \quad \text{arba} \quad T^+ \leq T^+_{1-\alpha}. \tag{5.6.3}$$

**5.6.1 pastaba.** Reikia pažymėti, kad Vilkoksono ranginis ženklų kriterijus gali būti neefektyvus, kai skirstinių šeimai $\mathcal{F}$ priklauso ir nesimetriški skirstiniai.

Iš tikrųjų, tegu teisinga alternatyva $H_1 : F \in \mathcal{F},\ M > M_0$, o skirstiniai turi ilgą kairiąją „uodegą“. Nors skaičius skirtumų $D_i$, įgyjančių teigiamas reikšmes, bus didesnis negu skaičius skirtumų, įgyjančių neigiamas reikšmes, tačiau pastarųjų skirtumų absoliutiniai didumai paprastai bus gerokai didesni. Todėl rangų suma $T^+$ gali būti palyginama su rangų suma $T^-$, kaip ir esant teisingai hipotezei. Taigi ranginis Vilkoksono kriterijus gali neskirti hipotezės nuo alternatyvos.

Kai $n$ nėra didelis (pvz., $n \leq 30$), kritines reikšmes $T^+_{\alpha}(n)$ (ir $P$ reikšmes) galime rasti toliau pateikiamu statistikos $T^+$ tikimybiniu skirstiniu.

**5.6.1 teorema.** *Jeigu hipotezė $H_0$ teisinga, tai*

$$\mathbf{E}(T^+) = \frac{n(n+1)}{4}, \quad \boldsymbol{V}(T^+) = \frac{n(n+1)(2n+1)}{24},$$
$$\mathbf{P}\{T^+ = k\} = \frac{c_{kn}}{2^n}, \quad k = 0, 1, \ldots, n(n+1)/2; \tag{5.6.4}$$

*čia $c_{kn}$ yra koeficientai prie $t^k$ sandaugoje $\prod_{k=1}^{n}(1+t^k)$.*

**Įrodymas.** Statistiką $T^+$ galima užrašyti kitaip. Nagrinėkime variacinę eilutę $|D|_{(1)}, \ldots, |D|_{(n)}$ gautą iš $|D_1|, \ldots, |D_n|$. Apibrėžkime

$$W_i = \begin{cases} 1, \text{ kai } \exists D_j > 0 : |D|_{(i)} = D_j, \\ 0, \qquad \text{kitais atvejais.} \end{cases}$$

Taigi $W_i = 1$, jei egzistuoja $D_j > 0$: $R_j = i$, todėl,

$$T^+ = \sum_{i=1}^{n} iW_i. \tag{5.6.5}$$

Kadangi a. d. $D_1, \ldots, D_n$ yra nepriklausomi ir įgyja teigiamas ir neigiamas reikšmes su vienodomis tikimybėmis $0,5$, tai su visais $k_1, \ldots, k_n \in \{0, 1\}$

$$\mathbf{P}\{W_1 = k_1, \ldots, W_n = k_n\} = \left(\frac{1}{2}\right)^n, \quad \mathbf{P}\{W_i = k_i\} = \frac{1}{2}.$$

taigi $W_1, \ldots, W_n$ yra nepriklausomi vienodai pasiskirstę Bernulio a. d.: $W_i \sim B(1, 1/2)$. Todėl

$$\mathbf{E}(T^+) = \sum_{i=1}^{n} i\frac{1}{2} = \frac{n(n+1)}{4}, \quad \boldsymbol{V}(T_+) = \sum_{i=1}^{n} i^2\frac{1}{4} = \frac{n(n+1)(2n+1)}{24}.$$

Atsitiktinio dydžio $T^+$ generuojančioji funkcija

$$\psi(t) = \mathbf{E}(t^{T^+}) = \sum_{k=0}^{M} t^k\mathbf{P}\{T^+ = k\} \tag{5.6.6}$$

turi tokį pavidalą

$$\psi(t) = \prod_{i=1}^{n} \mathbf{E}t^{iW_i} = \frac{1}{2^n}\prod_{i=1}^{n}(1+t^i) = \sum_{k=0}^{M} c_{kn}t^k. \tag{5.6.7}$$

Iš (5.6.6) ir (5.6.7) gauname (5.6.4). ▲

**4.6.1 pavyzdys.** Vertybinių popierių biržoje grąža (eurais už akciją) yra:

$$3,45; 4,21; 2,56; 6,54; 3,25; 7,11.$$

Tikrinsime hipotezes: a) mediana ne mažesnė už 4 eurus; b) mediana ne didesnė už 4 eurus; c) mediana ne didesnė už 3 eurus; d) mediana ne didesnė už 7 eurus; e) mediana lygi 3 eurams. Kriterijaus reikšmingumo lygmuo $\alpha = 0,1$.

a); b) Imdami $M_0$ =4, gauname skirtumus $D_i = X_i - 4$ : $-0,55; 0,21; -1,44; 2,54; -0,75; 3,11$. Išrikiuojame didėjančia tvarka palikdami ženklus: $0,21; -0,55; -0,75; -1,44; 2,54; 3,11$. Rangai (paliekant ženklus) yra: $1; -2; -3; -4; 5; 6$. Yra trys teigiami ir trys neigiami skirtumai. Rangų, atitinkančių teigiamus skirtumus, suma $T^+$ įgijo reikšmę $t^+ = 1 + 5 + 6 = 12$.

Atveju a) $P$ reikšmė yra

$$pv = \mathbf{P}\{T^+ \geq 12|H_0\} = 0,422.$$

Atveju b) $P$ reikšmė yra

$$pv = \mathbf{P}\{T^+ \leq 12|H\} = 0,656.$$

Duomenys neprieštarauja iškeltoms hipotezėms.

c) Randame skirtumus $D_i = X_i - 3$ : $0,45; 1,21; -0,44; 3,54; 0,25; 4,11$. Išrikiuojame didėjančia tvarka : $0,25; -0,44; 0,45; 1,21; 3,54; 4,11$. Rangai (paliekant ženklus): $1; -2; 3; 4; 5; 6$; $t^+ = 19$ ir

$$pv = \mathbf{P}\{T^+ \geq 19|H_0\} = 0,047.$$

Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,047.

d) Randame skirtumus $D_i = X_i - 7$ : $-3,55; -2,79; -4,44; -0,46; -3,75; 0,11$. Yra tik vienas teigiamas skirtumas, kurį atitinka rangas 1, taigi $T^+ = 1$ ir $P$ reikšmė

$$pv = \mathbf{P}\{T^+ \leq 1|H\} = 0,031.$$

Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,031.

e) Atveju c) gavome $T^+ = 19$. Tada

$$pv = 2\min(\mathbf{P}\{T^+ \leq 19|H\}, \mathbf{P}\{T^+ \geq 19|H\}) = 2\min(0,953; 0,047) = 0,094.$$

Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,094.

**5.6.2 pastaba** Jeigu dėl apvalinimo paklaidų kai kurie skirtumai $D_i$ lygūs nuliui, juos atmetame ir imties dydį sumažiname atmestųjų skirtumų skaičiumi.

**Didelės imtys.** Kai imtys didelės, imčių atveju asimptotinį kriterijų sudarome naudodami ribinį statistikos $T^+$ skirstinį.

**5.6.2 teorema.** *Jei hipotezė $H_0$ teisinga, tai*

$$Z_n = \frac{T^+ - \mathbf{E}(T^+)}{\sqrt{\boldsymbol{V}(T_+)}} \xrightarrow{d} Z \sim N(0,\, 1).$$

**Įrodymas.** Kadangi

$$\frac{T^+ - \mathbf{E}(T^+)}{\sqrt{\boldsymbol{V}(T_+)}} = \sum_{i=1}^{n} Y_i, \quad Y_i = \frac{iW_i - i/2}{\sqrt{\boldsymbol{V}(T_+)}},$$

$$\mathbf{E}Y_i = 0, \quad \boldsymbol{V}(\sum_{i=1}^{n} Y_i) = 1,$$

$$\mathbf{E}|Y_i|^3 = \mathbf{E}\left|\frac{|iW_i - i/2|}{\boldsymbol{V}(T_+)}\right|^3 = \frac{(i/2)^3}{[n(n+1)(2n+1)/24]^{3/2}} \le \frac{n^3}{8[2n^3/24]^{3/2}},$$

tai teoremos rezultatas išplaukia iš Liapunovo teoremos. ▲

**Asimptotinis Vilkoksono ranginis ženklų kriterijus.** Jeigu $n$ yra didelis, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|Z_n| \ge z_{\alpha/2}.$$

**5.6.3 pastaba.** Jeigu yra sutampančių reikšmių, tai statistika modifikuojama:

$$Z_n^* = \frac{Z_n}{\sqrt{1 - T/(2n(n+1)(2n+1))}};$$

čia $T = \sum_{l=1}^{k}(t_l^3 - t_l)$; $k$ yra sutampančių grupių skaičius, o $t_l$ – yra $l$-osios grupės didumas. Hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|Z_n^*| \ge z_{\alpha/2}.$$

**5.6.2 pavyzdys.** (**2.3.2** pavyzdžio tęsinys) Remdamiesi **2.3.2** pavyzdžio duomenimis patikrinsime hipotezę, kad a. d.$V$ mediana yra a) didesnė už 15; b) lygi 15.

Teigiamų skirtumų $D_i = X_i - 15$ skaičius yra 15 ir $T^+ = 347,5$. Statistika

$$Z_n = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{347,5 - \frac{49(49+1)}{4}}{\sqrt{\frac{49(49+1)(249+1)}{24}}} = -2,63603.$$

Kadangi yra trys grupės sutampančių rangų: $(12,5; 12,5)$, $(14,5; 14,5)$, $(33,5; 33,5)$ po du elementus ir viena grupė : $(18,5; 18,5; 18,5; 18,5)$ su 4 elementais, tai $k = 4$, $t_1 = t_2 = t_3 = 2$, $t_4 = 4$ ir $T = 3(2^3 - 2) + (4^3 - 4) = 78$. Korekcija yra nedidelė

$$\sqrt{1 - T/(2n(n+1)(2n+1))} = 0,999920 \quad \text{ir} \quad Z_n^* = \frac{Z_n}{0,999920} = -2,63677.$$

a) Hipotezė yra atmetama, kai statistika $Z_n^*$ įgyja mažas reikšmes. Asimptotinė $P$ reikšmė

$$pv_a = \Phi(-2,63677) = 0,00419.$$

Hipotezė atmetama. Tiksli $P$ reikšmė, rasta naudojant SPSS paketą, yra $pv = 0,003815$.

b) Hipotezė atmetama, kai statistika $|Z_n^*|$ įgyja dideles reikšmes. Asimptotinė $P$ reikšmė yra

$$pv_a = 2(1 - \Phi(2,63677)) = 0,00838.$$

Tiksli $P$ reikšmė, rasta naudojant SPSS paketą, yra $pv = 0,00763$.

**5.6.3 pastaba.** Kai kuriuose matematinės statistikos paketuose naudojama tokia kriterijaus modifikacija: vietoje statistikos $T^+$ naudojama statistika

$$T = \min(T^+, T^-) - \frac{n(n+1)}{4} = \min(T^+, \frac{n(n+1)}{2} - T^+) - \frac{n(n+1)}{4}. \quad (5.6.8)$$

Ši statistika įgyja neteigiamas reikšmes ir esant teisingai hipotezei $H_0$ dauguma jos reikšmių artimos 0. Šios statistikos skirstinys randamas naudojantis statistikos $T^+$ skirstiniu. Hipotezė $H_0$ atmetama, kai $T < T_{1-\alpha}$; čia $T_{1-\alpha}$ yra statistikos $T$ lygmens $\alpha$ kritinė reikšmė. Remdamiesi 4.6.2 teorema gauname, kad su visais $x < 0$

$$\mathbf{P}\{\frac{T}{\sqrt{\boldsymbol{V}(T_+)}} \leq x\} \to 2\Phi(x).$$

Todėl kai $n$ didelis, hipotezė $H_0$ atmetama, kai

$$T/\sqrt{\boldsymbol{V}(T_+)} < -z_\alpha;$$

čia $\Phi(x)$ ir $z_\alpha$ yra standartinio normaliojo skirstinio pasiskirstymo funkcija ir lygmens $\alpha$ kritinė reikšmė.

## 5.6.2. Vilkoksono ranginio ženklų kriterijaus ASE Stjudento kriterijaus atžvilgiu

Rasime Vilkoksono ranginio ženklų kriterijaus ASE Stjudento kriterijaus atžvilgiu, kai skirstiniai simetriški (tada hipotezės dėl medianos ir dėl vidurkio reikšmės sutampa).

Pažymėkime $f_D(x)$ ir $F_D(x)$ a. d. $D$ tankio funkciją ir pasiskirstymo funkciją, kai hipotezė $H_0$ yra teisinga. Tankio funkcija $f_D(z)$ simetriška nulio atžvilgiu: $f_D(-x) = f_D(x)$.

Kai teisinga alternatyva $H_3$, tai a. d. $D$ tankio funkcija yra $f_D(x|\theta) = f_D(x-\theta)$, $\theta = M - M_0 \neq 0$, o pasiskirstymo funkcija $F_D(x|\theta) = F_D(x - \theta)$.

**5.6.3 teorema.** *Vilkoksono ranginio ženklų kriterijaus ASE Stjudento kriterijaus atžvilgiu, kai skirstiniai simetriški yra*

$$e(T^+, t) = 12\tau^2 \left[\int_{-\infty}^{\infty} f_D^2(x)\, dx\right]^2, \quad \tau^2 = \boldsymbol{V}(D_i). \quad (5.6.9)$$

**Įrodymas.** Statistiką $T^+$ galima užrašyti tokiu pavidalu

$$T^+ = \sum_{1 \leq i \leq j \leq n} T_{ij}, \quad T_{ij} = \begin{cases} 1, \text{ kai } D_i + D_j > 0, \\ 0, \text{ priešingu atveju.} \end{cases} \quad (5.6.10)$$

Nagrinėkime variacinę eilutę $D_{(1)} \leq ... \leq D_{(n)}$. Tarkime, kad $D_{(k)} < 0$, $D_{(k+1)} > 0$. Tada gauname:

$$\sum_{1\leq i\leq j\leq n} \mathbf{1}_{\{D_i+D_j>0\}} = \sum_{1\leq i\leq j\leq n} \mathbf{1}_{\{D_{(i)}+D_{(j)}>0\}} = \sum_{j=k+1}^{n} \sum_{i=1}^{j} \mathbf{1}_{\{D_{(i)}+D_{(j)}>0\}} =$$

$$\sum_{j=k+1}^{n} (\sum_{i=1}^{k} \mathbf{1}_{\{D_{(i)}+D_{(j)}>0\}} + \sum_{i=k+1}^{j} 1) = \sum_{j=k+1}^{n} (\sum_{i=1}^{k} \mathbf{1}_{\{|D_{(j)}|>|D_{(i)}|\}} + j - k) = T^+.$$

Kai teisinga alternatyva

$$\mathbf{E}_\theta T^+ = n\mathbf{E}_\theta T_{11} + \frac{n(n-1)}{2}\mathbf{E}_\theta T_{12},$$

$$\mathbf{E}_\theta T_{11} = \mathbf{P}_\theta\{D_1 > 0\} = 1 - F_D(-\theta), \quad \mathbf{E}_\theta T_{12} = \mathbf{P}_\theta\{D_1 + D_2 > 0\} =$$

$$\int_{-\infty}^{\infty} \mathbf{P}_\theta\{D_1 > -x\} dF_{D_2}(x|\theta) = \int_{-\infty}^{\infty} [1 - F_D(-y-2\theta)] dF_D(y).$$

Kadangi kriterijus, grindžiamas statistika $T^+$, yra ekvivalentus kriterijui, grindžiamam statistika $V^+ = T^+/C_n^2$, galima nagrinėti statistiką $V^+$. Gauname:

$$\mu_{1n}(\theta) = \mathbf{E}_\theta V^+ = \frac{2}{n-1}[1 - F_D(-\theta)] + \int_{-\infty}^{\infty} [1 - F_D(-y-2\theta)] dF_D(y) \to$$

$$\int_{-\infty}^{\infty} [1 - F_D(-y-2\theta)] dF_D(y) = \mu_1(\theta), \quad \dot{\mu}_1(\theta) = 2\int_{-\infty}^{\infty} f_D(-y-2\theta) dF_D(y).$$

Esant simetriškam skirstiniui

$$\dot{\mu}_1(0) = 2\int_{-\infty}^{\infty} f_D^2(y) dy.$$

Kai hipotezė teisinga, a. d. $T^+$ dispersija pateikta 5.6.1 teoremoje:

$$\boldsymbol{V}_0(T_+) = n(n+1)(2n+1)/24,$$

taigi

$$\sigma_{1n}^2(0) = \boldsymbol{V}_0 V^+ = (n+1)(2n+1)/(6n(n^2-1)),$$

$$n\sigma_{1n}^2(0) \to \sigma_1^2(0) = \frac{1}{3}.$$

Stjudento kriterijus grindžiamas statistika $t = \sqrt{n}\bar{D}/s$, kuri asimptotiškai ekvivalenti statistikai $\bar{D}$, nes $s$ konverguoja pagal tikimybę į konstantą.

Turime

$$\mu_2(\theta) = \mathbf{E}_\theta \bar{D} = \theta, \quad \dot{\mu}_2(\theta) = 1,$$

$$\sigma_{2,n}^2(0) = \boldsymbol{V}_0 \bar{Z} = \tau^2/n, \quad \tau^2 = \boldsymbol{V}_0 D_i, \quad n\sigma_{2n}^2(0) \to \tau^2 = \sigma_2^2(0),$$

Remiantis 1.6.1 Vilkoksono ranginio ženklų kriterijaus ASE Stjudento kriterijaus atžvilgiu, kai skirstiniai simetriški, yra

$$e(T^+, t) = \left(\frac{\dot{\mu}_1(0)\sigma_2(0)}{\sigma_1(0)\dot{\mu}_2(0)}\right)^2 = \left(\frac{2\int_{-\infty}^{\infty} f_D^2(x)\,dx\,\tau}{\frac{1}{\sqrt{3}}\cdot 1}\right)^2 = 12\tau^2\left[\int_{-\infty}^{\infty} f_D^2(x)\,dx\right]^2.$$

▲

**5.6.4 pastaba.** Gavome lygiai tokią pat ASE išraišką, kaip ir lygindami dviejų nepriklausomų imčių ranginį Vilkoksono, Mano ir Vitnio kriterijų su Stjudento kriterijumi. Taigi Vilkoksono ranginis ženklų kriterijus yra geras „konkurentas“ Stjudento kriterijui. Jo ASE yra artimas 1 , kai $D_i$ skirstinys yra normalusis (ASE=$3/\pi \approx 0,955$), lygus 1, kai skirstinys tolygusis ir kartais gali būti didesnis už 1. Pavyzdžiui, logistinio skirstinio ASE=$\pi^2/9 \approx 1,097$, ekstremalių reikšmių skirstinio ASE=$\pi^2/8 \approx 1,23$, dvigubo eksponentinio skirstinio ASE=2. Remdamiesi 5.5.3 teorema gauname, kad ASE negali būti mažesnis už 0,864.

## 5.7. Vilkoksono ranginis ženklų kriterijus dviejų priklausomų imčių homogeniškumo hipotezei tikrinti

Tarkime $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ yra paprastoji imtis, gauta stebint a. v. $(X, Y)^T$, turinčio du baigtinius momentus, kurio pasiskirstymo funkcja $F(x, y)$ priklauso absoliučiai tolydžių dvimačių pasiskirstymo funkcijų šeimai $\mathcal{F}$. Pažymėkime $F_1(x)$ ir $F_2(y)$ marginaliąsias pasiskirstymo funkcijas.

**Dviejų priklausomų imčių homogeniškumo hipotezė:**

$$H_0: \ F \in \mathcal{F}, \ F_1(x) \equiv F_2(x).$$

Dviejų imčių Vilkoksono ranginį ženklų kriterijų sudarysime tuo atveju, kai skirstinių šeimos $\mathcal{F}$ pasiskirstymo funkcijos turi tokį pavidalą

$$F(x, y) = F_0(x, y + \theta), \quad \theta \in \mathbf{R};$$

čia $F_0(x, y)$ yra simetriška pasiskirstymo funkcija, taigi a. v. $(X, Y - \theta)^T$ turi simetrišką skirstinį.

Jeigu alternatyvų klasė apima ir nesimetriškus skirstinius, tai hipotezei $H_0$ tikrinti kartais geriau taikyti paprastąjį ženklų kriterijų (žr. 6.1.1 skyrelį).

Tegu $D = X - Y$. Pažymėkime $M$ atsitiktinio dydžio $D$ medianą ir nagrinėkime hipotezes dėl medianos reikšmės.

**Hipotezė dėl a. d. $D$ medianos lygybės 0:**

$$H_0^*: F \in \mathcal{F}, \ M = 0.$$

Ši hipotezė yra platesnė už homogeniškumo hipotezę, nes esant teisingai $H_0$ pasiskirstymo funkcija $F(x,y)$ yra simetriška, todėl $\mathbf{P}\{D>0\}=\mathbf{P}\{D<0\}$ ir $M=0$.

Hipotezei $H_0^*$ tikrinti galime naudoti imtį $D_1,\ldots,D_n$ sudarytą iš skirtumų $D_i=X_i-Y_i$, ir jos pagrindu sukonstruodami kriterijų, analogišką Vilkoksono ranginiam ženklų kriterijui vienos imties atveju (didelėms imtims taip pat galima taikyti ir asimptotinį Stjudento kriterijų). Visi anketesnio skyrelio rezultatai galioja.

*Dviejų imčių Vilkoksono ranginio ženklų kriterijaus statistika* turi tas pačias savybes kaip ir Vilkoksono ranginio ženklų kriterijaus statistika vienos imties atveju pakeičiant skirtumus $X_i-M_0$ į skirtumus $D_i=X_i-Y_i$ ir vietoje $M_0$ imant 0.

Jei hipotezė $H_0^*$ dėl medianos lygybės 0 atmetama, tai natūralu atmesti ir homogeniškumo hipotezę $H_0$, nes ji yra siauresnė.

**5.7.1 pavyzdys. 5.1.1** pavyzdžio tęsinys. Pagal **5.1.1** pavyzdžio duomenis patikrinsime hipotezę, kad skirtumo $D=X-Y$ mediana lygi nuliui.

Randame skirtumus $D_i=X_i-Y_i$, $i=1,...,50$, ir apskaičiuojame statistikos $T^+$ reikšmę $T^+=718,5$. Esant teisingai hipotezei gauname (žr. 5.6.1 teoremą): $\mathbf{E}T^+=637,5$, $\boldsymbol{V}T^+=10731,25$. Apskaičiuojame $Z_n=0,7819$, o modifikuotos statistikos reikšmė (atsižvelgiant į sutampančias reikšmes) $Z_n^*=0,7828$. Asimptotinė P reikšmė $pv_a=2(1-\Phi(0,7828))=0,4337$. Duomenys neprieštarauja iškeltai hipotezei.

## 5.8. Kruskalo ir Voliso kriterijus

Tarkime, kad

$$\mathbf{X}_1=(X_{11},...,X_{1n_1})^T,\quad \ldots\quad,\quad \mathbf{X}_k=(X_{k1},...,X_{kn_k})^T$$

yra $k$ paprastųjų imčių, gautų stebint n. a. d. $X_1,...,X_k$ su absoliučiai tolydžiomis pasiskirstymo funkcijomis $F_1(x),...,F_k(x)$.

**Kelių nepriklausomų imčių homogeniškumo hipotezė:**

$$H_0: F_1(x)=F_2(x)=\cdots=F_k(x)=: F(x),\ \forall x\in\mathbf{R}. \tag{5.8.1}$$

Tarkime, kad hipotezės $H_0$ alternatyva yra poslinkio

$$H_1: F_j(x)=F(x-\theta_j)\ \text{ su visais } x\in\mathbf{R},\quad j=1,\ldots,k,\quad \sum_{j=1}^k\theta_j^2>0. \tag{5.8.2}$$

Jeigu visi skirstiniai $F_j$ yra normalieji su vienoda dispersija $\sigma^2$ ir galbūt skirtingais vidurkiais $\mu_j$, tai alternatyva $H_1$ yra

$$H_1: \mu_j=\mu+\theta_j,\ \text{ su visais } x\in\mathbf{R},\quad j=1,\ldots,k,\quad \sum_{j=1}^k\theta_j^2>0,$$

ir hipotezė $H_0$ tampa parametrine:

$$H_0 : \mu_1 = ... = \mu_k.$$

Kaip žinoma, kriterijus normaliųjų skirstinių vidurkių lygybės hipotezei tikrinti, kai dispersijos vienodos, nagrinėjamas vienfaktorėje dispersinėje analizėje ir grindžiamas statistika

$$F = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 / \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i(X_{ij} - \bar{X}_{i.})^2; \qquad (5.8.3)$$

čia $\bar{X}_{i.}$ yra $i$-osios imties aritmetinis vidurkis, $\bar{X}_{..}$ – jungtinės imties aritmetinis vidurkis, $n = \sum_{i=1}^{k} n_i$.

Jeigu hipotezė (5.8.1) teisinga, tai statistika (5.8.3) turi Fišerio skirstinį su $k-1$ ir $n-k$ laisvės laipsnių. Jeigu $n_i \to \infty$, $n_i/n \to p_i \in (0,\, 1)$, $k$ yra fiksuotas, tai (5.8.3) vardiklis pagal tikimybę konverguoja į $\sigma^2$, o statistika $F$ konverguoja į a. d. $\chi^2_{k-1}/(k-1)$.

Jei $n_i$ yra dideli, tai statistika $F$ apytiksliai proporcinga (5.8.3) skaitikliui.

**Kriterijaus sudarymo idėja.** Kai skirstiniai nežinomi, stebiniai $X_{i1}, ..., X_{in_i}$ keičiami jų rangais $R_{i1}, ..., R_{in_i}$ jungtinėje visų stebinių variacinėje eilutėje.

*Kruskalo ir Voliso* statistika apibrėžiama pagal analogiją su Fišerio statistika (5.8.3) $\bar{X}_{i.}$ keičiant į $\bar{R}_{i.}$ ir $\bar{X}_{..}$ keičiant į $\bar{R}_{..} = (n+1)/2$ ir gautą skaitiklio išraišką dauginant iš proporcingumo koeficiento:

$$F_{KW} = \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i\left(\bar{R}_{i\cdot} - \frac{n+1}{2}\right)^2 = \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i \bar{R}^2_{i\cdot} - 3(n+1)$$

$$= \frac{12}{n(n+1)}\sum_{i=1}^{k} \frac{R^2_{i\cdot}}{n_i} - 3(n+1); \qquad (5.8.4)$$

čia

$$R_{i\cdot} = \sum_{j=1}^{n_i} R_{ij}$$

yra $i$-osios imties elementų rangų bendroje variacinėje eilutėje suma, $i = 1, \ldots, k$.

Proporcingumo koeficientas parenkamas taip, kad esant teisingai hipotezei $H_0$, statistikos $F_{KW}$ skirstinys, kai imčių didumai auga, artėtų į chi kvadrato skirstinį.

Esant teisingai hipotezei $H_0$ :

$$\mathbf{E}(\bar{R}_{i\cdot}) = \mathbf{E}(R_{ij}) = (n+1)/2$$

su visais $i$, o kai teisinga alternatyva, kai kurie vidurkiai didesni arba mažesni už $(n+1)/2$, taigi statistika $F_{KW}$, kuri grindžiama dėmenų $\left(\bar{R}_{i\cdot} - (n+1)/2\right)^2$ suma, turi tendenciją įgyti didesnes reikšmes, kai teisinga alternatyva.

**Kruskalo ir Voliso kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $F_{KW} > F_{KW}(\alpha)$; čia $F_{KW}(\alpha)$ yra minimalus skaičius $c$, tenkinantis nelygybę $\mathbf{P}\{F_{KW} > c|H_0\} \leq \alpha$.

**Didelės imtys.** Jeigu visos imtys yra didelės, tai statistikos $F_{KW}$ skirstinys aproksimuojamas chi kvadrato skirstiniu su $k-1$ laisvės laipsniu.

**5.8.1 teorema.** *Jeigu hipotezė $H_0$ teisinga, $n_i/n \to p_{i0} \in (0,\,1)$, $i = 1, ..., k$, tai*

$$F_{KW} \xrightarrow{d} S \sim \chi^2(k-1). \tag{5.8.5}$$

**Įrodymas.** Pažymėkime $R_i = \sum_{j=1}^{n_i} R_{ij}$ $i$-osios imties rangų jungtinėje visų stebėjimų variacinėje eilutėje sumą.

Kadangi $R_i$ dispersija sutampa su Vilkoksono statistikos dispersija $i$-ąją grupę interpretuodami kaip pirmąją, o visas likusias – kaip antrąją, gauname

$$\mathbf{E}R_i = n_i \frac{n+1}{2}, \quad \mathbf{V}R_i = \frac{n_i(n+1)(n-n_i)}{12}.$$

Be to, su visais $i \neq j$

$$\mathbf{cov}(R_i, R_j) = \sum_{l=1}^{n_i} \sum_{s=1}^{n_j} \mathbf{cov}(R_{il}, R_{js}) = -n_i n_j \frac{n+1}{12}.$$

Taigi a. v. $\boldsymbol{R} = (R_1, ..., R_k)^T$ kovariacinė matrica yra $\boldsymbol{\Sigma}_n = [\sigma_{ij}]_{k\times k}$; čia

$$\sigma_{ij} = \begin{cases} n_i(n+1)(n-n_i)/12, & i = j, \\ -n_i n_j (n+1)/12, & i \neq j. \end{cases}$$

Gauname:

$$\frac{12}{(n+1)n^2}\boldsymbol{\Sigma}_n = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \cdots & \cdots & \cdots & \cdots \\ -p_kp_1 & -p_kp_2 & \cdots & p(1-p_k) \end{pmatrix} = \boldsymbol{D} - \boldsymbol{p}\boldsymbol{p}^T;$$

čia $p_i = n_i/n$, $\boldsymbol{p} = (p_1, ..., p_k)^T$, $\boldsymbol{D}$ yra diagonalinė matrica su diagonaliniais elementais $p_1, ..., p_k$.

Matricos $\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{p}\boldsymbol{p}^T$ apibendrintoji atvirkštinė matrica yra

$$\boldsymbol{A}^- = (\boldsymbol{D} - \boldsymbol{p}\boldsymbol{p}^T)^- = \boldsymbol{D}^{-1} + \frac{1}{p_k}\mathbf{1}\mathbf{1}^T, \quad \mathbf{1} = (1, ..., 1)^T.$$

Iš tikrųjų, remdamiesi lygybėmis

$$\mathbf{1}^T\boldsymbol{D} = \boldsymbol{p}^T, \quad \mathbf{1}^T\boldsymbol{p} = \boldsymbol{p}^T\mathbf{1} = 1, \quad \boldsymbol{D}\mathbf{1} = \boldsymbol{p}, \quad \boldsymbol{p}^T\boldsymbol{D}^{-1} = \mathbf{1}^T,$$

gauname $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$. Taigi

$$\boldsymbol{\Sigma}_n^- = \frac{12}{n^2(n+1)}[\sigma^{ij}]_{k\times k}, \quad \sigma^{ii} = \frac{1}{p_k} + \frac{1}{p_i}, \quad \sigma^{ij} = \frac{1}{p_k}, \ i \neq j.$$

Gauname

$$(\boldsymbol{R}-\mathbf{E}\boldsymbol{R})^T\boldsymbol{\Sigma}_n^-(\boldsymbol{R}-\mathbf{E}\boldsymbol{R}) = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left(R_i-\frac{n_i(n+1)}{2}\right)^2 = F_{KW}.$$

Reikia pažymėti, kad

$$n^{-3}\boldsymbol{\Sigma}_n \to \boldsymbol{\Sigma} = \frac{1}{12}(\boldsymbol{D}_0-\boldsymbol{p}_0\boldsymbol{p}_0^T),$$

čia $\boldsymbol{p}_0 = (p_{10},\dots,p_{k0})^T$ ,o $\boldsymbol{D}_0$ yra diagonalinė matrica su diagonaliniais elementais $p_{10},\dots,p_{k0}$. Matricos $\boldsymbol{\Sigma}$ rangas yra $k-1$ (žr. 2.4 pratimą).

Remiantis CRT atsitiktinių vektorių sekoms

$$n^{-3/2}(\mathbf{R}-E(\mathbf{R})) \xrightarrow{d} \boldsymbol{Z} \sim N_k(\mathbf{0},\boldsymbol{\Sigma});$$

čia

$$\boldsymbol{\Sigma} = \frac{1}{12}(\boldsymbol{D}_0-\boldsymbol{p}_0\boldsymbol{p}_0^T), \quad \boldsymbol{p}_0=(p_{10},\dots,p_{k0})^T,$$

ir $\boldsymbol{D}_0$ yra diagonalinė matrica su diagonaliniais elementais $p_{10},\dots,p_{k0}$.

Naudosimės teorema, kuri tvirtina, kad jei

$$X \sim N_k(\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad \text{tai} \quad (X-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^-(X-\boldsymbol{\mu}) \sim \chi^2(r);$$

čia $\boldsymbol{\Sigma}^-$ yra matricos $\boldsymbol{\Sigma}$ apibendrintoji atvirkštinė matrica, t. y. matrica, tenkinanti sąlygą $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^-\boldsymbol{\Sigma}=\boldsymbol{\Sigma}$; $r$ – matricos $\boldsymbol{\Sigma}$ rangas.

Pagal šią teoremą

$$n^{-3}(\mathbf{R}-E(\mathbf{R}))^T\boldsymbol{\Sigma}^-(\mathbf{R}-E(\mathbf{R})) \xrightarrow{d} S \sim \chi^2(k-1),$$

iš čia išplaukia teoremos tvirtinimas.

▲

**Asimptotinis Kruskalo ir Voliso kriterijus:** jeigu $n_i$ nėra maži, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$F_{KW} > \chi^2_\alpha(k-1).$$

**Kruskalo ir Voliso kriterijaus ASE Fišerio kriterijaus atžvilgiu.** Kai $n_i \to \infty$, $k$ fiksuotas, tai Kruskalo ir Voliso kriterijaus ASE Fišerio kriterijaus atžvilgiu yra

$$e(F_{KW},F) = 12\sigma^2\left[\int_{-\infty}^{\infty} f^2(x)dx\right]^2;$$

čia $f$ ir $\sigma^2$ yra a. d. $X_{ij}$ tankio funkcija ir dispersija, kai hipotezė teisinga.

Ši formulė identiška ASE išraiškai, gautai lyginant Vilkoksono kriterijų su Stjudento kriterijumi, kai yra dvi imtys. ASE yra arti 1, kai skirstinys normalusis (ASE $= 3/\pi \approx 0,955$), lygus 1, kai skirstinys tolygusis, kartais ASE gali viršyti 1. Pavyzdžiui, logistinio skirstinio ASE $= \pi^2/9 \approx 1,097$, ekstremalių reikšmių skirstinio ASE $= \pi^2/8 \approx 1,234$ dvigubo eksponentinio skirstinio ASE $= 2$. Pačiu nepalankiausiu atveju ASE negali būti mažesnis už 0,864.

**Sutampantys duomenys.** Jei duomenys apvalinami, tai galimos vienodos kai kurių stebinių reikšmės netgi ir kai skirstiniai absoliučiai tolydūs. Jei yra sutampančių reikšmių, tai statistika $F_{KW}$ modifikuojama analogiškai Vilkoksono statistikai:

$$F^*_{KW} = F_{KW}/(1 - T/(n^3 - n));$$

čia $T = \sum_{i=1}^{s} T_i, \quad T_i = (t_i^3 - t_i)$, $s$ yra sutampančių narių grupių skaičius jungtinėje imtyje; $t_i$ yra $i$-osios grupės didumas.

**Modifikuotasis asimptotinis Kruskalo ir Voliso kriterijus:** hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $F^*_{KW} > \chi^2_\alpha(k-1)$.

**5.8.1 pavyzdys.** Serotonino kiekis buvo matuotas po trijų skirtingų vaistų injekcijų. Lentelėje pateiktjs gautosios serotonino kiekio reikšmės trijose pacientų grupėse. Ar šių trijų medikamentų poveikis vienodas?

| 1 (placebo) | 2 (vaistas 1 ) | 3 (vaistas 2) |
|---|---|---|
| 340 | 294 | 263 |
| 340 | 325 | 309 |
| 356 | 325 | 340 |
| 386 | 340 | 356 |
| 386 | 356 | 371 |
| 402 | 371 | 371 |
| 402 | 385 | 402 |
| 417 | 402 | 417 |
| 433 | | |
| 495 | | |
| 557 | | |

Sudarome jungtinę visų stebinių variacinę eilutę (grupės numeris nurodytas skliausteliuose):

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 263(3) | 294(2) | 309(3) | 325(2) | 325(2) | 340(1) | 340(1) | 340(2) | 340(3) |
| 10 | 11 | 12 | 13 | 14 | 5 | 16 | 17 | 18 |
| 356(1) | 356(2) | 356(3) | 371(2) | 371(3) | 371(3) | 385(2) | 386(1) | 386(1) |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| 402(1) | 402(1) | 402(2) | 402(3) | 417(1) | 417(3) | 433(1) | 495(1) | 557(1) |

Rangų sumos:

$$R_{1.} = 7,5 + 7,5 + 11 + 17,5 + 17,5 + 20,5 + 20,5 + 23,5 + 25 + 26 + 27 = 203,5,$$

$$R_{2.} = 2 + 4,5 + 4,5 + 7,5 + 11 + 14 + 16 + 20,5 = 80,$$

$$R_{3.} = 1 + 3 + 7,5 + 11 + 14 + 14 + 20,5 + 23,5 = 94,5.$$

Imčių didumai: $n_1 = 11$, $n_2 = 8$, $n_3 = 8$, bendras stebinių skaičius $n = 27$.

Kruskalo ir Voliso statistikos reikšmė yra

$$F_{KW} = \frac{12}{27 \cdot 28}\left(\frac{203,5^2}{11} + \frac{80^2}{8} + \frac{94,5^2}{8}\right) - 3 \cdot 28 = 6,175.$$

Yra $s = 7$ sutampančių rangų grupės

$$t_1 = 2,\ t_2 = 3,\ t_3 = 3,\ t_4 = 3,\ t_5 = 2,\ t_6 = 4,\ t_7 = 2,$$

taigi

$$T_1 = 8 - 2 = 6,\ T_2 = 27 - 3 = 24,\ T_3 = 24,\ T_4 = 24,\ T_5 = 6,\ T_6 = 64 - 4 = 60,\ T_7 = 6.$$

Modifikuotosios Kruskalo ir Voliso statistikos reikšmė yra:

$$F^*_{KW} = \frac{6,175}{1 - \frac{3\cdot 6 + 3\cdot 24 + 60}{27(729-1)}} = \frac{6,175}{0,99134} = 6,234.$$

$P$ reikšmė rasta, skaičiuojant SPSS paketu, yra $pv = 0,03948$.

Asimptotinė $P$ reikšmė yra

$$pv_a = \mathbf{P}\{\chi^2_2 > 6,234\} \approx 0,0443.$$

Jeigu kriterijaus reikšmingumo lygmuo 0,05, tai hipotezė atmetama tiek tiksliu, tiek asimptotiniu kriterijumi.

**5.8.1 pastaba.** Jeigu tikrinant hipotezę $H_0$ alternatyvos yra mastelio:

$$H_1 : F_j(x) = F_j(x/\theta_j), \quad x \in \mathbf{R}, \quad \theta_j > 0, \quad j = 1, ..., k;$$

$$\exists\, \theta_i \neq \theta_j, \quad 1 \leq i \neq j \leq k, \tag{5.8.6}$$

tai, kaip ir Vilkoksono kriterijus, Kruskalo ir Voliso kriterijus gali būti neefektyvus.

Analogiškai 5.4 skyreliui vietoje rangų $R_{ij}$ imkime $s(R_{ij})$, čia $s(r)$ yra tam tikra aibėje $\{1, 2, ..., n\}$ apibrėžta funkcija, ir apibrėžkime statistiką

$$F = \frac{1}{\sigma_s^2}\sum_{i=1}^{k} n_i(\bar{s}_i - \bar{s})^2, \quad \bar{s}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} s(R_{ij}),$$

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} s(R_{ij}), \quad \sigma_s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(s(R_{ij}) - \bar{s})^2. \tag{5.8.7}$$

Imdami tas pačias funkcijas kaip ir skyrelyje 5.4, gausime Zygelio ir Tjukio, Ansari ir Bredlio, Mūdo , Klotso statistikų analogus $F_{ZT}, F_{AB}, F_M, F_K$ tuo atveju, kai imčių skaičius $k > 2$.

Kai hipotezė $H_0$ teisinga, statistikos $F_{ZT}$ skirstinys sutampa su statistikos $F_{KW}$ skirstiniu.

Mažiems imčių dydžiams $n_i$ statistikų $F_{ZT}, F_{AB}, F_M, F_K$ P reikšmės gali būti surastas naudojant kai kuriuos programų paketus (SAS, SPSS).

Analogiškai 5.8.1 teoremai galima įrodyti, kad ir kitų statistikų $F_{ZT}$, $F_{AB}$, $F_M$, $F_K$ skirstiniai asimptotiškai yra chi kvadrato skirstiniai su $k-1$ laisvės laipsniu.

Didelėms imtims hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai:

$$F > \chi^2_\alpha(k-1), \tag{5.8.8}$$

čia $F$ yra bet kuri iš statistikų $F_{ZT}, F_{AB}, F_M, F_K$.

**5.8.2 pavyzdys.** (**5.8.1** pavyzdžio tęsinys).

Papildykime **5.8.1** pavyzdžio duomenis matavimų rezultatais, gautais atliekant matavimus dar dviem prietaisų tipais:

3 tipo prietaisas: 2,0742; -1,0499; 0,8555; 3,2287; 2,2756; 1,4305; 0,8750; -0,5559; -1,2684; -0,4667; 1,0099; -2,9228; -0,1835; 1,0803; 0,0759;

4 tipo prietaisas: 1,7644; 0,4839; -2,1736; 0,9326; -1,0432; -0,1026; 0,9777; 0,6117; -0,4034; 2,6000; -09851; 0,0052; -0,5035; 2,7274; 0,5828.

Pagal šiuos ir **4.5.3** pratimo duomenis patikrinkime hipotezę, kad visų keturių prietaisų paklaidos turi vienodus skirstinius.

Jeigu tartume, kad matavimo paklaidos turi normalųjį skirstinį su nuliniu vidurkiu, tai hipotezė tampa parametrine dėl dispersijų lygybės $H: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$. Šią hipotezę galima tikrinti Bartleto kriterijumi, grindžiamu tikėtinumų santykio statistika

$$R_{TS} = n\ln(s^2) - \sum_{i=1}^{4} n_i \ln(s_i^2),$$

čia $n = n_1 + n_2 + n_3 + n_4$, $s_i^2$ yra dispersijos įvertinys pagal $i$-osios imties duomenis, o $s^2 = (n_1 s_1^2 + ... + n_4 s_4^2)/n$. Gauname $R_{TS} = 6,7055$ ir asimptotinė $P$ reikšmė yra $pv_a = \mathbf{P}\{\chi_3^2 > 6,7055\} = 0,0819$.

Pritaikysime šio skyrelio neparametrinius kriterijus. Gauname statistikų reikšmes: $S_{ZT} = 8,9309; S_{AB} = 8,8003; S_M = 4,9060; S_K = 6,6202$. Naudodami SAS paketą gauname tokias asimptotines

$P$ reikšmes 0,0302; 0,0321; 0,1788; 0,0850. Matome, kad šiame pavyzdyje Klotso kriterijus duoda praktiškai tą patį atsakymą kaip ir parametrinis Bartleto kriterijus; kriterijai, grindžiami statistikomis $S_{ZT}$ ir $S_{AB}$, pasirodė galingesni, o Mūdo kriterijus mažiau galingas.

## 5.9. Frydmano kriterijus

Tarkime, kad turime n nepriklausomų vektorių

$$(X_{11}, ..., X_{1k})^T, \dots, (X_{n1}, ..., X_{nk})^T, \quad k > 2,$$

Tegu a. v. $\mathbf{X}_i = (X_{i1}, ..., X_{ik})^T$ pasiskirstymo funkcija $F_i = F_i(x_1, ..., x_k)$ priklauso neparametrinei $k$-mačių absoliučiai tolydžių pasiskirstymo funkcijų šeimai $\mathcal{F}$.

Duomenis galima interpretuoti ir kaip $k$ priklausomų (ar nepriklausomų) paprastųjų imčių

$$(X_{11}, ..., X_{n1})^T, \dots, (X_{1k}, ..., X_{nk})^T;$$

čia $\boldsymbol{Y}_j = (X_{1j}, ..., X_{nj})^T$ yra vektorius su nepriklausomomis koordinatėmis.

Visus turimus stebėjimus galime surašyti į tokią matricą:

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} \tag{5.9.1}$$

Pažymėkime $F_{i1}, \dots, F_{ik}$ marginaliąsias a. v. $\mathbf{X}_i = (X_{i1}, ..., X_{ik})^T$ pasiskirstymo funkcijas.

$k$ **priklausomų imčių homogeniškumo hipotezė:**

$$H_0 : F_{i1}(x) = ... = F_{ik}(x), \quad \forall\, x \in \mathbf{R}, \quad i = 1, ..., n.$$

Aptarsime keletą tokios hipotezės formuluočių konkrečiomis situacijomis.

**5.9.1 pavyzdys**. Buvo matuojamas $n = 6$ grupių darbuotojų darbo efektyvumas kiekvieną savaitės dieną ($k = 7$). Atsitiktinis dydis $X_{ij}$ žymi $i$-osios grupės darbo efektyvumą $j$-ąją savaitės dieną. Atsitiktinio vektoriaus $\mathbf{X}_i = (X_{i1}, ..., X_{ik})^T$ koordinatės yra priklausomos (ta pati grupė stebima $k$ kartų), bet nebūtinai vienodai pasiskirsčiusios. Vektoriai $\mathbf{X}_1$,...,$\mathbf{X}_n$ yra nepriklausomi (skirtingos darbuotojų grupės). Taigi a. v. $\boldsymbol{Y}_j = (X_{1j}, ..., X_{nj})^T$ koordinatės yra nepriklausomos ir gali būti vienodai pasiskirsčiusios (jeigu grupės vienodos kvalifikacijos) arba skirtingai pasiskirsčiusios (jeigu grupės skirtingos kvalifikacijos). Atsitiktiniai vektoriai $\boldsymbol{Y}_1$,...,$\boldsymbol{Y}_k$ yra priklausomi (stebime tas pačias grupes skirtingomis dienomis).

Reikia patikrinti hipotezę, kad darbo efektyvumas nepriklauso nuo savaitės dienos. Šiame uždavinyje reikia palyginti ne darbuotojų grupes, bet savaitės dienas *eliminuojant grupės faktorių.*

**5.9.2 pavyzdys**.Yra užfiksuotas $n = 10$ pacientų reakcijos laikas veikiant juos trijų tipų ($k = 3$) skirtingais vaistais. Atsitiktinis dydis $X_{ij}$ žymi $i$-ojo paciento reakcijos laiką veikiant $j$-uoju vaistu. Reikia patikrinti hipotezę, kad visų vaistų poveikis reakcijos laikui yra vienodas.

Tarkime, kad absoliučiai tolydžios pasiskirstymo funkcijos $F_i$, priklausančios aibei $\mathcal{F}_i$, turi tokį pavidalą

$$F_i(x_1, x_2, \ldots, x_k) = G_i(x_1, x_2 + \theta_{i2}, \ldots, x_k + \theta_{ik}), \quad \theta_{ij} \in \mathbf{R};$$

$G_i$ yra simetriškos funkcijos, t. y. su visais $x_1, \ldots, x_k \in \mathbf{R}$ ir su visais kėliniais $(j_1, \ldots, j_k)$ of $(1, \ldots, k)$

$$G_i(x_{j_1}, \ldots, x_{j_k}) = G_i(x_1, \ldots, x_{_k}). \tag{5.9.2}$$

Homogeniškumo hipotezės $H_0$ alternatyva $H_1$ turi tokį pavidalą:

$$H_1: \ F_i \in \mathcal{F}, \ \ F_{ij}(x) = F_{i1}(x - \theta_{ij}), \ i = 1, \ldots, n; \ j = 2, \ldots, k; \ \ \sum_{i=1}^{n} \sum_{j=2}^{k} \theta_{ij}^2 > 0.$$

t. y. marginaliosios pasiskirstymo funkcijos skiriasi tik poslinkio parametrais.

**Frydmano kriterijaus statistikos sudarymas.** Randame a. v. $(X_{i1}, ..., X_{ik})^T$ rangų vektorių $(R_{i1}, ..., R_{ik})^T$. Tada vietoje pradinių duomenų matricos $\mathbf{X}$ gauname rangų matricą

$$\boldsymbol{R} = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1k} \\ R_{21} & R_{22} & \cdots & R_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ R_{n1} & R_{n2} & \cdots & R_{nk} \end{pmatrix}$$

Rangų suma kiekvienoje eilutėje ta pati:

$$R_{i\cdot} = R_{i1} + ... + R_{ik} = k(k+1)/2, \quad i = 1, 2...., n.$$

Pažymėkime $j$-ojo stulpelio rangų sumą

$$\bar{R}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} R_{ij}.$$

Visų rangų aritmetinis vidurkis yra

$$\bar{R}_{\cdot\cdot} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} R_{ij} = \frac{k+1}{2}.$$

Kai skirstiniai simetriški, tai esant teisingai hipotezei $H_0$ atsitiktiniai dydžiai $\bar{R}_{\cdot 1}, ..., \bar{R}_{\cdot k}$ yra vienodai pasiskirstę ir jų įgyjamos reikšmės grupuojasi apie vidurkį $\bar{R}_{\cdot\cdot}$.

Frydmano kriterijaus statistika grindžiama skirtumais $\bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot} = \bar{R}_{\cdot j} - (k+1)/2$:

$$S_F = \frac{12n}{k(k+1)}\sum_{j=1}^{k}(\bar{R}_{\cdot j} - \frac{k+1}{2})^2 = \frac{12n}{k(k+1)}\sum_{j=1}^{k}\bar{R}_{\cdot j}^2 - 3n(k+1)$$

$$= \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_{\cdot j}^2 - 3n(k+1); \tag{5.9.3}$$

čia $R_{\cdot j} = \sum_{i=1}^{n} R_{ij}$. Normuojantis daugiklis $12n/k(k+1)$ parenkamas taip, kad esant teisingai hipotezei $H_0$, asimptotiškai (kai $n \to \infty$) statistikos $S_F$ skirstinys artėtų prie chi kvadrato skirstinio.

**Frydmano kriterijus:** hipotezė $H_0$ atmetama kriterijumi su reikšmingumo lygmeniu $\alpha$, kai $S_F \geq S_{F,\alpha}$; čia $S_{F,\alpha}$ yra mažiausias skaičius $c$, tenkinantis nelygybę $\mathbf{P}\{S_F \geq c | H_0\} \leq \alpha$.

$P$ reikšmė yra $pv = \mathbf{P}\{S_F \geq s\}$; čia $s$ yra gautoji statistikos $S_F$ realizacija. Kai $n$ nėra dideli, Frydmano statistikos kritines reikšmes (arba $P$ reikšmes) galima rasti remiantis rangų skirstiniais, pateikiamais 4.1 skyrelyje.

**Didelės imtys.** Rasime Frydmano statistikos asimptotinį skirstinį, kai imties didumas $n \to \infty$.

**5.9.1 teorema.** *Kai hipotezė $H_0$ teisinga, tai*

$$S_F \xrightarrow{d} S \sim \chi^2(k-1), \quad n \to \infty. \tag{5.9.4}$$

**Įrodymas.** Tegu

$$\bar{\mathbf{R}} = (\bar{R}_{\cdot 1}, \ldots, \bar{R}_{\cdot k})^T.$$

Pagal (5.2.2) a. v. $\bar{\mathbf{R}}$ vidurkis $\mathbf{E}(\bar{\mathbf{R}})$ ir kovariacinė matrica $\boldsymbol{V}(\bar{\mathbf{R}}) = \boldsymbol{\Sigma}_n$ yra

$$\mathbf{E}(\bar{\mathbf{R}}) = ((k+1)/2, \ldots, (k+1)/2), \quad \boldsymbol{\Sigma}_n = [\sigma_{ls}]_{k\times k};$$

čia

$$\sigma_{ls} = \mathbf{Cov}(\bar{R}_{\cdot l}, \bar{R}_{\cdot s}) = \frac{1}{n}\mathbf{Cov}(R_{1l}, R_{1s}) = \begin{cases} (k^2-1)/(12n), & \text{kai} \quad l = s, \\ -(k+1)/(12n), & \text{kai} \quad l \neq s. \end{cases}$$

Taigi
$$\mathbf{\Sigma}_n = \frac{k(k+1)}{12n}(\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T);$$
čia $\mathbf{E}_k$ yra vienetinė matrica, $\mathbf{1} = (1,\dots,1)^T$. Matricos $\mathbf{\Sigma}_n$ rangas yra $k-1$, nes
$$k\mathbf{E}_k - \mathbf{1}\mathbf{1}^T = \begin{pmatrix} k-1 & -1 & \cdots & -1 \\ -1 & k-1 & \cdots & -1 \\ \cdots & \cdots & \cdots & \cdots \\ -1 & -1 & \cdots & k-1 \end{pmatrix} \sim$$
$$\begin{pmatrix} k-1 & -1 & \cdots & -1 \\ -k & k & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ -k & 0 & \cdots & k \end{pmatrix} \sim \begin{pmatrix} 0 & -1 & \cdots & -1 \\ 0 & k & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & k \end{pmatrix}$$
Visų pirma pridėjome pirmąją eilutę prie kitų eilučių, paskui prie pirmojo stulpelio pridėjome likusius.

Pagal CRT atsitiktinių vektorių sumoms,
$$\sqrt{n}(\bar{\mathbf{R}} - \mathbf{E}(\bar{\mathbf{R}})) \xrightarrow{d} \boldsymbol{Z} \sim N_k(\mathbf{0}, \mathbf{\Sigma});$$
čia
$$\mathbf{\Sigma} = n\mathbf{\Sigma}_n = k(k+1)(\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T)/12, \quad Rang(\mathbf{\Sigma}) = k-1.$$
Kaip ir Kruskalo ir Voliso kriterijaus atveju remsimės teorema: jeigu $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \mathbf{\Sigma})$, tai $(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{\Sigma}^-(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2(r)$, čia $\mathbf{\Sigma}^-$ yra apibendrintoji atvirkštinė matrica, o $r$ yra matricos $\mathbf{\Sigma}$ rangas. Pagal šią teoremą
$$Q_n = \sqrt{n}(\bar{\mathbf{R}} - \mathbf{E}\bar{\mathbf{R}})\mathbf{\Sigma}^-\sqrt{n}(\bar{\mathbf{R}} - \mathbf{E}\bar{\mathbf{R}})^T \xrightarrow{d} S \sim \chi^2(k-1).$$
Įrodysime, kad $Q_n = S_F$. Gauname
$$(\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T)^- = \mathbf{E}_k + k\mathbf{1}\mathbf{1}^T,$$
nes
$$\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T = k\mathbf{1}\mathbf{1}^T$$
ir
$$(\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T)(\mathbf{E}_k + k\mathbf{1}\mathbf{1}^T)(\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T) = (\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T)(\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T) = (\mathbf{E}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^T)$$
Taigi
$$\mathbf{\Sigma}^- = \frac{12}{k(k+1)}(\mathbf{E}_k + k\mathbf{1}\mathbf{1}^T)$$
ir
$$Q_n = \frac{12n}{k(k+1)}\left(\sum_{j=1}^k(\bar{R}_{\cdot j} - \frac{k+1}{2})^2 + k[\sum_{j=1}^k(\bar{R}_{\cdot j} - \frac{k+1}{2})]^2\right) =$$

$$\frac{12n}{k(k+1)}\sum_{j=1}^{k}\left(\bar{R}_{\cdot j}-\frac{k+1}{2}\right)^2=S_F.$$

▲

**Asimptotinis Frydmano kriterijus:** jei $n$ yra didelis, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$S_F>\chi^2_\alpha(k-1). \tag{5.9.5}$$

Tada $P$ reikšmės aproksimacija $pv_a=1-F_{\chi^2_{k-1}}(s)$; čia $s$ yra stebėtoji statistikos $S_F$ reikšmė.

**Sutampančios reikšmės.** Jeigu yra sutampančių stebėjimų, tai Frydmano statistika modifikuojama. Įrodysime 5.9.1 teoremos analogą bendruoju atveju, kai skirstiniai nebūtinai absoliučiai tolydūs. Pažymėkime

$$S_F^*=\frac{S_F}{1-\sum_{i=1}^n T_i/(n(k^3-k))};$$

čia

$$T_i=\sum_{j=1}^{k_i}(t_{ij}^3-t_{ij}),$$

$k_i$ yra sutampančių reikšmių grupių skaičius $i$-ajam objektui (t. y. $i$-ojoje matricos (5.9.1) eilutėje), $t_{ij}$ yra $j$-osios grupės elementų skaičius.

**5.9.2 teorema.** *Kai hipotezė $H_0$ teisinga, tai*

$$S_F^*\xrightarrow{d}S\sim\chi^2(k-1),\quad n\to\infty. \tag{5.9.6}$$

**Įrodymas.** Jis nedaug skiriasi nui 5.9.1 teoremos įrodymo. Naudosime tuos pačius žymėjimus, kaip ir šiai teoremai įrodyti. be to, pažymėkime $t=\mathbf{E}T_i$. Pagal (5.2.6) a. v. $\bar{\mathbf{R}}$ vidurkis toks pat, o kovariacinės matricos $\mathbf{\Sigma}_n$ elementai yra

$$\sigma_{ls}=\begin{cases}\frac{1}{12n}(k^2-1-\frac{t}{k}), & \text{kai}\quad l=s,\\ \frac{1}{12n}(-(k+1)+\frac{t}{k(k-1)}), & \text{kai}\quad l\neq s.\end{cases}=$$

$$\begin{cases}\frac{k(k+1)}{12n}(1-\frac{t}{k^3-k})(1-\frac{1}{k}), & \text{kai}\quad l=s,\\ \frac{k(k+1)}{12n}(1-\frac{t}{k^3-k})(-\frac{1}{k}), & \text{kai}\quad l\neq s.\end{cases}$$

taigi

$$\mathbf{\Sigma}_n=\frac{k(k+1)}{12n}(1-\frac{t}{k^3-k})(\mathbf{E}_k-\frac{1}{k}\mathbf{1}\mathbf{1}^T),$$

$$\mathbf{\Sigma}^-=\frac{12}{k(k+1)}(1-\frac{t}{k^3-k})(\mathbf{E}_k+k\mathbf{1}\mathbf{1}^T),$$

todėl

$$\frac{S_F}{1-\frac{t}{k^3-k}}\xrightarrow{d}S\sim\chi^2(k-1).$$

Konvergavimas į chi kvadrato skirstinį išlieka pakeitus $t$ į $\frac{1}{n}\sum_{i=1}^n T_i$, nes

$$\frac{1}{n}\sum_{i=1}^n T_i\xrightarrow{P}t=\mathbf{E}T_1.$$

▲

**5.9.3 pavyzdys.** Trimis skirtingais metodais ($k = 3$) išmatuotas amilazės kiekis tiems patiems $n = 9$ pankreatitu sergantiems pacientams. Duomenys pateikti lentelėje.

| Pacientas | Metodas | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 4000 | 3210 | 6120 |
| 2 | 1600 | 1040 | 2410 |
| 3 | 1600 | 647 | 2210 |
| 4 | 1200 | 570 | 2060 |
| 5 | 840 | 445 | 1400 |
| 6 | 352 | 156 | 249 |
| 7 | 224 | 155 | 224 |
| 8 | 200 | 99 | 208 |
| 9 | 184 | 70 | 227 |

Patikrinsime hipotezę, kad šie trys metodai ekvivalentūs.

Amilazės kiekiai nustatomi tuo pačiu metodu skirtingiems pacientams, gali būti laikomi nepriklausomais a. d. Jie gali būti skirtingai pasiskirstę (pvz., jei ligos stadijos yra skirtingos). Skirtingais metodais gauti amilazės kiekiai tam pačiam pacientui yra priklausomi atsitiktiniai dydžiai.

Rangų lentelė

| Pacientas | Metodas | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 2 | 1 | 3 |
| 2 | 2 | 1 | 3 |
| 3 | 2 | 1 | 3 |
| 4 | 2 | 1 | 3 |
| 5 | 2 | 1 | 3 |
| 6 | 3 | 1 | 2 |
| 7 | 2,5 | 1 | 2,5 |
| 8 | 2 | 1 | 3 |
| 9 | 2 | 1 | 3 |
| | $R_{.1} = 19,5$ | $R_{.2} = 9$ | $R_{.3} = 25,5$ |

Turime $n = 9$, $k = 3$,

$$S_F = \frac{12}{9 \cdot 3 \cdot 4}(19,5^2 + 9^2 + 25,5^2) - 3 \cdot 9 \cdot 4 = 15,5.$$

Šiame pavyzdyje yra viena grupė sutampančių reikšmių septintajam pacientui:

$$k_7 = 1,\ t_7 = 2,\ T_7 = 2^3 - 2 = 6,\ \sum_{i=1}^{9} T_i = 6,$$

taigi

$$1 - \frac{\sum_{i=1}^{n} T_i}{nk(k^2 - 1)} = 1 - \frac{6}{9 \cdot 3 \cdot 8} = 0,97222$$

ir

$$S_F^* = \frac{S_F}{0,97222} = 15,943.$$

Tiksli $P$ reikšmė yra $pv = 0,00034518$. Hipotezė atmestina. Asimptotinė $P$ reikšmė $pv_a = \mathbf{P}\{\chi_2^2 > 15,943\} = 0,00001072$ gerokai skiriasi nuo tikslios, nes imtys nėra didelės.

**5.9.4 pavyzdys.** (**5.9.2** pavyzdžio tęsinys ). Kiekvieną savaitės dieną ($k = 7$) buvo nustatomas $n = 6$ darbininkų grupių darbo efektyvumas. Duomenys pateikti lentelėje.

| | Diena | | | | | | |
|---|---|---|---|---|---|---|---|
| Grupė | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 60 | 62 | 58 | 52 | 31 | 23 | 26 |
| 2 | 64 | 63 | 63 | 36 | 34 | 32 | 27 |
| 3 | 14 | 46 | 47 | 39 | 42 | 43 | 57 |
| 4 | 30 | 41 | 40 | 41 | 37 | 17 | 12 |
| 5 | 72 | 38 | 46 | 47 | 38 | 60 | 41 |
| 6 | 35 | 35 | 33 | 46 | 47 | 47 | 38 |

Tikriname hipotezę, kad darbo efektyvumas nepriklauso nuo savaitės dienos.

Rangų lentelė

| | Diena | | | | | | |
|---|---|---|---|---|---|---|---|
| Grupė | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 6 | 7 | 5 | 4 | 3 | 1 | 2 |
| 2 | 7 | 5,5 | 5,5 | 4 | 3 | 2 | 1 |
| 3 | 1 | 5 | 6 | 2 | 3 | 4 | 7 |
| 4 | 3 | 6,5 | 5 | 6,5 | 4 | 2 | 1 |
| 5 | 7 | 1,5 | 4 | 5 | 1,5 | 6 | 3 |
| 6 | 2,5 | 2,5 | 1 | 5 | 6,5 | 6,5 | 4 |
| | 26,5 | 28 | 26,5 | 26,5 | 21 | 21,5 | 18 |

Turime $n = 6$, $k = 7$,

$$S_F = (26,5^2 + 28^2 + 26,5^2 + 26,5^2 + 21^2 + 21,5^2 + 18^2)/28 - 3 \cdot 6 \cdot 8 = 3,07143.$$

Šiame pavyzdyje sutampančių grupių daugiau:

$$k_2 = k_4 = k_5 = 1, \quad k_6 = 2, \quad t_2 = t_4 = t_5 = t_{61} = t_{61} = 2,$$

$$T_3 = T_4 = T_5 = 6, \quad T_6 = 6 + 6 = 12, \quad \sum_{i=1}^{6} T_i = 30.$$

Koreguojantis daugiklis

$$1 - \frac{\sum_{i=1}^{n} T_i}{nk(k^2-1)} = 1 - \frac{30}{6 \cdot 7 \cdot 48} = 0,985119,$$

ir

$$S_F^* = 3,07143/0,985119 = 3,1178.$$

Asimptotinė P reikšmė $pv_a = 0,7939$. Atmesti hipotezę nėra pagrindo.

**5.9.1 pastaba.** Nagrinėta situacija su skirtingais a. v. $\mathbf{X}_1, ..., \mathbf{X}_n$ skirstiniais aptinkama ir tada, kai a. v. $\mathbf{X}_i = (X_{i1}, ..., X_{ik})^T$ koordinatės nepriklausomos. Kadangi iš $j$-ųjų koordinačių sudaryta imtis $X_{1j}, ..., X_{nj}$ nėra paprastoji, tai vietoje Kruskalo ir Voliso kriterijaus reikėtų naudoti Frydmano kriterijų.

**5.9.5 pavyzdys.** Tiriamas 9 skirtingų metalo lydinių atsparumas korozijai. Lydiniai bandomi 8 skirtingomis sąlygomis. Duomenys pateikti lentelėje.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1,40 | 1,45 | 1,91 | 1,89 | 1,77 | 1,66 | 1,92 | 1,84 | 1,54 |
| 2 | 1,35 | 1,57 | 1,48 | 1,48 | 1,73 | 1,54 | 1,93 | 1,79 | 1,43 |
| 3 | 1,62 | 1,82 | 1,89 | 1,39 | 1,54 | 1,68 | 2,13 | 2,04 | 1,70 |
| 4 | 1,31 | 1,24 | 1,51 | 1,67 | 1,23 | 1,40 | 1,23 | 1,58 | 1,64 |
| 5 | 1,63 | 1,18 | 1,58 | 1,37 | 1,40 | 1,45 | 1,51 | 1,63 | 1,07 |
| 6 | 1,41 | 1,52 | 1,65 | 1,11 | 1,53 | 1,63 | 1,44 | 1,28 | 1,38 |
| 7 | 1,93 | 1,43 | 1,38 | 1,72 | 1,32 | 1,63 | 1,33 | 1,69 | 1,70 |
| 8 | 1,40 | 1,86 | 1,36 | 1,37 | 1,34 | 1,36 | 1,38 | 1,80 | 1,84 |

Šiame pavyzdyje natūralu tarti, kad a. v. $(X_1, ..., X_9)^T$ koordinatės yra nepriklausomos, tačiau dėl nevienodų bandymo sąlygų gali turėti skirtingus skirstinius. Hipotezę, kad visi lydiniai vienodai atsparūs korozijai, galima suformuluoti šitaip:

$$H_0 : F_{i1}(x) = ... = F_{i9}(x), \quad \forall x \in \mathbf{R}, \quad \forall i = 1, ..., 8.$$

Vietoje Kruskalo ir Voliso kriterijaus reikėtų taikyti Frydmano kriterijų.

Gauname $S_F = 31,1417$ ir, atsižvelgiant į sutampančias reikšmes, $S_F^* = 31,2720$. Asimptotinė $P$ reikšmė $pv_a = \mathbf{P}\{\chi_9^2 > 31,2720\} = 0,00027$. Hipotezė atmestina.

**Frydmano kriterijaus ASE Fišerio kriterijaus atžvilgiu: nepriklausomos imtys.** Jei su kiekvienu $i$ a. d. $X_{i1}, ..., X_{ik}$ yra nepriklausomi, o alternatyva yra

$$F_{ij}(x) = F(x - \alpha_i - \beta_j), \quad \exists j, j' : \beta_j \neq \beta_{j'},$$

tai hipotezė $H_0$ ekvivalenti parametrinei hipotezei

$$H_0' : \beta_1 = \cdots = \beta_k.$$

Parametriniu atveju, kai skirstiniai yra normalieji, turime dvifaktorės dispersinės analizės modelį. Minėtoji hipotezė tikrinama naudojant Fišerio kriterijų, kurio statistika

$$F = \frac{n(n-1)\sum_{j=1}^{k}(\bar{X}_{.j} - \bar{X}_{..})^2}{\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2}$$

esant teisingai hipotezei turi Fišerio skirstinį su $k-1$ ir $(n-1)(k-1)$ laisvės laipsnių. Hipotezė atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$F > F_\alpha(k-1, (n-1)(k-1)).$$

Sudarydami kriterijų neparametriniu atveju remiamės tuo, kad statistika $(k-1)F$ asimptotiškai $(n \to \infty)$ turi chi kvadrato skirstinį su $k-1$ laisvės laipsnių. Hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$(k-1)F > \chi^2_\alpha(k-1).$$

Įrodyta, kad Frydmano kriterijaus ASE Fišerio kriterijaus atžvilgiu yra

$$e(S_F, F) = \frac{12k\sigma^2}{k+1}\left[\int_{-\infty}^{\infty} f^2(x)dx\right]^2, \quad f = F', \quad \sigma^2 = \boldsymbol{V}(X_{ij}).$$

čia $f(x)$ yra stebimų a. d. tankio funkcija, $\sigma^2 = \boldsymbol{V}(X_{ij})$.

Kai skirstinys normalusis, gauname

$$e(S_F, F) = \frac{3k}{(k+1)\pi}.$$

Efektyvumas labai priklauso nuo $k$. Jeigu $k = 2$, tai ASE yra mažiausias: $ASE = 2/\pi \approx 0,637$, t. y. sutampa su ASE ženklų kriterijaus. Kai $k = 5$, tai $ASE = 0,796$, ir jei $k$ didėja, tai ASE artėja prie $3/\pi = 0,955$. Skirstiniams, kurių „uodegos“ gęsta lėčiau, ASE gali netgi viršyti 1. Pavyzdžiui, Laplaso skirstinio ASE $= 2 - 2/(k+1)$.

## 5.10. Ranginis kelių imčių nepriklausomumo kriterijus

Apibendrinsime Spirmeno ranginio koreliacijos koeficiento sąvoką, kai imčių skaičius $k > 2$.

Tarkime, kad turime paprastąją imtį

$$(X_{11}, ..., X_{1k})^T, \ldots, (X_{n1}, ..., X_{nk})^T,$$

gautą stebint a. v, $(X_1, ..., X_k)^T$.

Tokie duomenys gaunami, pavyzdžiui, kai $k$ ekspertų vertina $n$ objektų kokybę: atsitiktinį vektorių $(X_{j1}, ..., X_{jk})^T$ sudaro visų ekspertų $j$-ojo objekto įvertinimai, $j = 1, ..., n$. Atsitiktinį vektorių $(X_{1j}, ..., X_{nj})^T$ sudaro visų objektų įvertinimai atlikti $j$-ojo eksperto.

Pažymėkime $(R_{ij}, ..., R_{nj})^T$ rangų vektorių, gautą ranguojant a. v. $(X_{ij}, ..., X_{nj})^T$ elementus. Kiekvieno eksperto rangų suma yra ta pati: $R_{\cdot j} = \sum_{i=1}^n R_{ij} = n(n+1)/2$, $j = 1, ..., k$.

Rangų suma, tekusi $i$-ajam objektui, yra $R_{i\cdot} = \sum_{j=1}^k R_{ij}$. Rangų sumų $R_{1\cdot}, ..., R_{n\cdot}$ aritmetinis vidurkis

$$\bar{R}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^n R_{i\cdot} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k R_{ij} = \frac{1}{n}\sum_{j=1}^k\sum_{i=1}^n R_{ij} = \frac{1}{n}k\frac{n(n+1)}{2} = \frac{k(n+1)}{2}.$$

**5.10.1 apibrėžimas.** Atsitiktinis dydis

$$W = \frac{12}{k^2 n(n^2-1)}\sum_{i=1}^n (R_{i\cdot} - k(n+1)/2)^2 \tag{5.10.1}$$

vadinamas *Kendalo konkordancijos koeficientu.*

Pažymėkime

$$R_S^{(jl)} = \frac{12}{n^3 - n}\sum_{i=1}^n (R_{ij} - (n+1)/2)(R_{il} - (n+1)/2)$$

Spirmeno koreliacijos koeficientą, sudarytą remiantis $j$-uoju ir $l$-uoju atsitiktiniais vektoriais $(X_{1j}, ..., X_{nj})^T$ ir $(X_{1l}, ..., X_{nl})^T$.

**5.10.1 teorema.** *Kendalo konkordancijos koeficientas yra Spirmeno koreliacijos koeficientų aritmetinio vidurkio*

$$R_S^{vid} = \frac{2}{k(k-1)}\sum_{j<l} R_S^{(jl)},$$

*tiesinė funkcija*

$$W = \frac{1}{k}(1 + (k-1)R_S^{vid}).$$

**Įrodymas.** Gauname

$$\sum_{i=1}^{n}(R_{i\cdot}-\frac{k(n+1)}{2})^2=\sum_{i=1}^{n}\{\sum_{j=1}^{k}(R_{ij}-\frac{n+1}{2})^2+2\sum_{j<l}(R_{ij}-\frac{n+1}{2})(R_{il}-\frac{n+1}{2})\}=$$

$$=\frac{k(n^3-n)}{12}+2\frac{(n^3-n)}{12}\sum_{j<l}R_S^{(jl)},$$

iš kur ir gaunamas teoremos tvirtinimas. ▲

Remdamiesi teorema gauname, kad kai $k=2$, koeficientas $W$ yra Spirmeno koreliacijos koeficiento tiesinė funkcija:

$$W=(R_S+1)/2.$$

Jei visi Spirmeno koreliacijos koeficientai yra lygūs 1, t. y. visų ekspertų vertinimai sutampa, tai $W$ įgyja didžiausią galimą reikšmę: $W=1$.

Jeigu visi Spirmeno koeficientai lygūs 0, tai $W=1/k$. Jei jie visi neneigiami, tai $W\geq 1/k$ .

Jei dalis Spirmeno koreliacijos koeficientų yra neigiami, tai iš apibrėžimo ir teoremos išplaukia, kad $W$ gali įgyti reikšmes kiek mažesnes, tiek ir didesnes už $1/k$ .

Reikia pažymėti, kad ir priklausomų stebėjimų atveju statistika $W$ gali įgyti reikšmę $1/k$, netgi su tikimybe 1. Pavyzdžiui, jei $k=4$, $X_i=-iX_1$, $i=2,3,4$, tai

$$R_S^{(12)}=R_S^{(13)}=R_S^{(14)}=-1,\quad R_S^{(23)}=R_S^{(24)}=R_S^{(34)}=1,$$

todėl $R_S^{vid}=0$ ir $W=1/k$.

Konkordancijos koeficientas naudojamas hipotezei dėl vektoriaus koordinačių nepriklausomumo tikrinti, kai alternatyva yra, kad visų porų koreliacijos yra teigiamos.

Tarkime, kad a. v. $(X_{i1},...,X_{ik})^T$ yra absoliučiai tolydus, o visi koreliacijos koeficientai $\rho_{jl}=\mathbf{E}R_S^{(jl)}$ neneigiami.

Pažymėkime $F_i$ atsitiktinio vektoriaus $X_{i1},\dots,X_{ik}$ pasiskirstymo funkciją ir $F_{i1},...,F_{ik}$ marginaliąsias pasiskirstymo funkcijas.

**A. d.** $X_{i1},\dots,X_{ik}$ **nepriklausomumo hipotezė:**

$$H_0: F_i(x_1,\dots,x_k)=F_{11}(x_1)\dots F_{ik}(x_k),\quad \forall x_1,\dots,x_k\in\mathbf{R},\quad i=1,...,n.$$

Tariame, kad alternatyva yra

$$\bar{H}:\ \rho_{jl}\geq 0,\ \ \forall\, 1\leq j<l\leq k\ ,\ \exists\rho_{i_0,l_0}>0.$$

**Nepriklausomumo kriterijus, grindžiamas konkordancijos koeficientu:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $W>W_\alpha$; čia $W_\alpha$ yra minimalus skaičius $c$, tenkinantis nelygybę $\mathbf{P}\{W>c|H_0\}\leq\alpha$.

Reikia pažymėti, kad statistikos $k(n-1)W$ skaičiavimo formulė sutampa su Frydmano statistikos formule, jei $n$ ir $k$ sukeisti vaidmenimis. Todėl, jei nepriklausomumo hipotezė teisinga, tai $k(n-1)W$ skirstinys sutampa su Frydmano statistikos skirstiniu, kai teisinga suderinamumo hipotezė, o $n$ ir $k$ sukeisti vaidmenimis. Taigi galima naudotis Frydmano statistikos kritinėmis reikšmėmis.

Remdamiesi sąryšiu su Frydmano statistika ir šios statistikos savybėmis gauname, kad kai dideli $k$, statistikos $k(n-1)W$ skirstinį galima aproksimuoti chi kvadrato skirstiniu su $n-1$ laisvės laipsniu. Tačiau tokia situacija nėra dažna. Reikia manyti, kad dažnesnė situacija, kai $n$ didelis, o $k$ mažesnis.

Kadangi statistikos $W$ skirstinys sukoncentruotas intervale $[0,\ 1]$, tai jo skirstinį galima aproksimuoti beta skirstiniu parenkant pastarojo parametrus taip, kad jo vidurkis ir dispersija sutaptų su statistikos $W$ atitinkamais momentais:

$$\mathbf{E}W = \frac{1}{k}, \quad \boldsymbol{V}W = \frac{2(k-1)}{k^3(n-1)}.$$

Gauname aproksimaciją

$$W \approx Y \sim Be(\gamma, \eta), \quad \gamma = \frac{n-1}{2} - \frac{1}{k}, \quad \eta = (k-1)\gamma. \tag{5.10.2}$$

**5.10.1 pavyzdys.** Lentelėje pateikta 20 studentų tikimybių teorijos, matematinės analizės ir matematinės statistikos egzaminų pažymiai $X_{1i}, X_{2i}, X_{3i}$, $i = 1, 2, ..., 20$.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_{1i}$ | 7 | 5 | 6 | 6 | 8 | 4 | 5 | 5 | 8 | 8 |
| $X_{2i}$ | 7 | 5 | 8 | 5 | 8 | 3 | 5 | 4 | 7 | 6 |
| $X_{3i}$ | 9 | 5 | 6 | 5 | 8 | 5 | 5 | 5 | 8 | 7 |

| $i$ | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_{1i}$ | 7 | 5 | 6 | 3 | 7 | 7 | 8 | 7 | 6 | 5 |
| $X_{2i}$ | 6 | 3 | 7 | 4 | 5 | 8 | 7 | 5 | 5 | 5 |
| $X_{3i}$ | 8 | 5 | 5 | 5 | 7 | 7 | 7 | 6 | 7 | 5 |

Tikėtina kad a.d. $X_{1i}, X_{2i}, X_{3i}$ yra teigiamai koreliuoti. Patikrinsime nepriklausomumo hipotezę naudodami Kendalo konkordancijos koeficientą. Atlikdami skaičiavimus SPSS paketu randame $W = 0,859$ ir $pv_a = 0,00026$. Nepriklausomumo hipotezė atmetama.

## 5.11. Pratimai

**5.1.** Įrodykite, kad absoliučiai tolydžių skirstinių atveju Spirmeno koreliacijos koeficientas $r_S = 1 - 12V/(n(n^2-1))$, čia $V = \sum_{i<j} h_{ij}(j-i)$, $h_{ij} = 1$, kai $R_i > R_j$ ir $h_{ij} = 0$, kai $R_i < R_j$.

**5.2.** Įrodykite, kad koeficientų $R_S$ ir $R_K$ Pirsono koreliacijos koeficientas yra

$$\rho(R_S, R_K) = 2(n+1)/\sqrt{2n(2n+5)}.$$

**5.3.** Krakmolo kiekis bulvėse nustatomas dviem būdais. Norint palyginti tuos būdus, buvo paimta 16 bulvių ir kiekvienos iš jų krakmolo kiekis nustatytas abiem būdais. Gauti rezultatai surašyti lentelėje ($X_i$ – pirmu būdu, o $Y_i$ – antruoju).

| $i$ | $X_i$ | $Y_i$ | $i$ | $X_i$ | $Y_i$ |
|---|---|---|---|---|---|
| 1 | 21,7 | 21,5 | 9 | 14,0 | 13,9 |
| 2 | 18,7 | 18,7 | 10 | 17,2 | 17,0 |
| 3 | 18,3 | 18,3 | 11 | 21,7 | 21,4 |
| 4 | 17,5 | 17,4 | 12 | 18,6 | 18,6 |
| 5 | 18,5 | 18,3 | 13 | 17,9 | 18,0 |
| 6 | 15,6 | 15,4 | 14 | 17,7 | 17,6 |
| 7 | 17,0 | 16,7 | 15 | 18,3 | 18,5 |
| 8 | 16,6 | 16,9 | 16 | 15,6 | 15,5 |

Patikrinkite hipotezę apie a. d. $X$ ir $Y$ nepriklausomumą, remdamiesi Spirmeno ir Kendalo ranginiais koreliacijos koeficientais.

**5.4.** Tiriant specialios sėjamosios efektyvumą, 10 sklypelių buvo sėjama paprasta sėjamąja ir 10 sklypelių – specialia sėjamąja, paskui buvo lyginamas derlingumas. Norint eliminuoti dirvožemio įtaką, 20 vienodo ploto sklypelių buvo taip sugrupuota poromis, kad jie būtų greta vienas kito. Metant monetą, buvo nusprendžiama, kuriame iš dviejų gretimų sklypelių sėti specialia sėjamąja. Rezultatai pateikti lentelėje ($X_i$ – derlingumas sėjant specialia sėjamąja, $Y_i$ – paprasta sėjamąja).

| $i$ | $X_i$ | $Y_i$ | $i$ | $X_i$ | $Y_i$ |
|---|---|---|---|---|---|
| 1 | 8,0 | 5,6 | 6 | 7,7 | 6,1 |
| 2 | 8,4 | 7,4 | 7 | 7,7 | 6,6 |
| 3 | 8,0 | 7,3 | 8 | 5,6 | 6,0 |
| 4 | 6,4 | 6,4 | 9 | 5,6 | 5,5 |
| 5 | 8,6 | 7,5 | 10 | 6,2 | 5,5 |

Patikrinkite hipotezę apie a. d. $X$ ir $Y$ nepriklausomumą, remdamiesi Spirmeno ir Kendalo ranginiais koreliacijos koeficientais.

**5.5.** Patikrinkite atsitiktinumo hipotezę pagal **2.16** pratimo duomenis.

**5.6.** Patikrinkite atsitiktinumo hipotezę pagal **2.17** pratimo duomenis.

**5.7.** Remdamiesi Vilkoksono ir Van der Vardeno kriterijais patikrinkite hipotezę apie nuodų poveikio vienodumą pagal **4.14** pratimo duomenis.

**5.8.** Įrodykite, kad gama skirstinio $G(1,\eta)$ atveju Vilkoksono kriterijaus ASE, palyginti su Stjudento kriterijumi, kai alternatyvos poslinkio yra

$$e(W,t) = A(\eta) = \frac{3\eta}{2^{4(\eta-1)}(2(\eta-1)B(\eta,\eta))^2}.$$

Patikrinkite, kad $A(\eta) > 1,25$, kai $\eta \leq 3$; $A(\eta) \to \infty$, kai $\eta \to 1/2$; $A(\eta) \to 3/\pi$, kai $\eta \to \infty$.

**5.9.** Tikrinama hipotezė, kad impulso atpažinimo paklaida nepriklauso nuo jo intensyvumo. Buvo atlikti du nepriklausomi eksperimentai. Impulsas, kurio intensyvumas 10 sąlyginių vienetų, buvo įvertintas 9, 9, 8, 10, 12, 13, 10, 11 vienetų; impulsas, kurio intensyvumas 20 sąlyginių vienetų, – 15, 16, 17, 23, 22, 20, 21, 24, 27. Remdamiesi ranginiais 4.5.5 skyrelio kriterijais, kai alternatyvos yra mastelio, patikrinkite, ar gauti duomenys neprieštarauja iškeltai hipotezei.

**5.10.** Suskaidykime **2.16** pratimo duomenis į 10 vienodo didumo imčių (duomenys surašyti skirtingose eilutėse). Remdamiesi Kruskalo ir Voliso ir inversijų skaičiumi grindžiamais kriterijais, patikrinkite hipotezę, kad visais atvejais buvo stebimas tas pats atsitiktinis dydis.

**5.11.** Trijose gamyklose buvo testuojami kineskopai. Jų darbo trukmė (mėnesiais iki pirmo gedimo) surašyti pateikiamoje lentelėje. Ar galima tvirtinti, kad visose trijose gamyklose gaminamų kineskopų darbo trukmė yra vienoda?

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 gamykla | 41 | 70 | 26 | 89 | 62 | 54 | 46 | 77 | 34 | 51 | | |
| 2 gamykla | 30 | 69 | 42 | 60 | 44 | 74 | 32 | 47 | 45 | 37 | 52 | 81 |
| 3 gamykla | 23 | 35 | 29 | 38 | 21 | 53 | 31 | 25 | 36 | 50 | 61 | |

**5.12.** Lentelėje pateikti avarijų Lietuvos keliuose 1990 – 1999 metų duomenys apie avarijas Lietuvos keliuose.

| Metai | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ | 5135 | 6067 | 4049 | 4319 | 3902 | 4144 | 4579 | 5319 | 6445 | 6356 |
| $Y$ | 933 | 1093 | 779 | 893 | 765 | 672 | 667 | 725 | 829 | 748 |
| $Z$ | 5491 | 6638 | 4251 | 4555 | 4146 | 4508 | 5223 | 6198 | 7669 | 7696 |

Apskaičiuokite konkordancijos koeficientą ir patikrinkite hipotezę dėl a. d. $X$ (avarijos), $Y$ (žuvusieji) ir $Z$ (sužeistieji) priklausomybės. Remdamiesi Spirmeno ir Kendalo ranginiais koreliacijos koeficientais patikrinkite hipotezes dėl a. d. $X$ ir $Y$ priklausomybės; dėl a. d. $X$ ir $Z$ priklausomybės.

**5.13.** Remdamiesi Vilkoksono ženklų kriterijumi priklausomoms imtims, patikrinkite hipotezę dėl a. d. X ir Y skirstinių vienodumo pagal **4.3** pratimo duomenis.

**5.14.** Remdamiesi Vilkoksono ženklų kriterijumi priklausomoms imtims, patikrinkite hipotezę dėl a. d. X ir Y skirstinių vienodumo pagal **4.4** pratimo duomenis.

**5.15.** Lentelėje pateikti duomenys apie 3 tiekėjų siūlomų 12 skirtingų tipų spausdintuvų kainas.

| Tipas | Tiekėjas | | | Tipas | Tiekėjas | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | | 1 | 2 | 3 |
| 1 | 660 | 673 | 658 | 7 | 1980 | 1950 | 1970 |
| 2 | 790 | 799 | 785 | 8 | 2300 | 2295 | 2310 |
| 3 | 590 | 580 | 599 | 9 | 2500 | 2480 | 2490 |
| 4 | 950 | 945 | 960 | 10 | 2190 | 2190 | 2210 |
| 5 | 1290 | 1280 | 1295 | 11 | 5590 | 5500 | 5550 |
| 6 | 1550 | 1500 | 1499 | 12 | 6000 | 6100 | 6090 |

Remdamiesi Frydmano kriterijumi patikrinkite hipotezę, kad skirtingų tiekėjų spausdintuvų kainos nesiskiria.

## 5.12. Atsakymai

**5.3.** Spirmeno ir Kendalo koreliacijos koeficientai įgijo reikšmes $r_S = 0,9889$ ir $\tau_b = 0,9422$. Abiejų kriterijų atveju $pv < 0,0001$. Hipotezė atmetama. **5.4.** Spirmeno ir Kendalo koreliacijos koeficientai įgijo reikšmes $r_S = 0,7822$ ir $\tau_b = 0,6746$. Atitinkamos P reikšmės $pv = 0,0075$ ir $pv = 0,0084$. Hipotezė atmetama. **5.5.** Spirmeno ir Kendalo koreliacijos koeficientai įgijo reikšmes $r_S = 0,0935$ ir $\tau_b = 0,0621$. Atitinkamos P reikšmės $pv = 0,3547$ ir $pv = 0,3663$. Atmesti hipotezę nėra pagrindo. **5.6.** Spirmeno ir Kendalo koreliacijos koeficientai įgijo reikšmes $r_S = 0,0643$ ir $\tau_b = 0,0390$. Atitinkamos P reikšmės $pv = 0,4343$ ir $pv = 0,4852$. Atmesti hipotezę nėra pagrindo. **5.7.** Vilkoksono statistika įgijo reikšmę $W = 239$, $Z_{m,n} = -0,0198$; asimptotinė P reikšmė yra $pv_a = 2\Phi(-0,0198) = 0,9842$. Van der Vardeno statistikos reikšmė $V = -0,1228$ ir $pv_a = 0,9617$. Atmesti hipotezę nėra pagrindo. **5.9.** Statistikų reikšmės yra: $S_{ZT} = 91,1667$; $S_{AB} = 47,5$; $S_K = 2,2603$; $S_M = 8,9333$ ir atitinkamos P reikšmės 0,0623; 0,0713; 0,0320; 0,0342. Asimptotinės P reikšmės: 0,0673; 0,0669; 0,0371; 0,0382. Homogeniškumo hipotezė atmestina. **5.10.** Kruskalo ir Voliso statistika įgijo reikšmę $F_{KW} = 3,9139$ ir $pv_a = 0,9170$. Atmesti hipotezę nėra pagrindo. 5.11. Kruskalo ir Voliso statistika įgijo reikšmę $F_{KW} = 6,5490$ ir $pv_a = 0,0378$. Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,0378. **5.12.** Kendalo konkordancijos koeficiento reikšmė 0,6444 ir $pv_a = 0,0428$; nepriklausomumo hipotezė atmetama, kai reikšmingumo lygmuo viršija 0,0428. a) Spirmeno ir Kendalo koreliacijos koeficientai įgijo reikšmes $r_S = 0,2242$ ir $\tau_b = 0,2$; P reikšmės yra 0,5334 ir 0,4208. Nepriklausomumo hipotezė neatmetama. b) Spirmeno ir Kendalo koreliacijos koeficientai įgijo reikšmes $r_S = 0,9879$ ir $\tau_b = 0,9556$; P reikšmės abiem atvejais yra $pv < 0,00001$. Nepriklausomumo hipotezė atmetama. **5.13.** Ranginio ženklų kriterijaus statistikos reikšmė yra $T^+ = 69,5$ ir $pv = 0,0989$. Homogeniškumo hipotezė atmetama, jei reikšmingumo lygmuo viršija 0,0989. **5.14.** Ranginio ženklų kriterijaus statistikos reikšmė yra $T^+ = 43$ ir $pv = 0,0117$. Homogeniškumo hipotezė atmetama, jei reikšmingumo lygmuo viršija 0,0117. **5.15.** Frydmano statistika įgijo reikšmę $S_F = 2,5957$; asimptotinė P reikšmė $pv_a = 0,2731$. Duomenys neprieštarauja iškeltai hipotezei.

# 6 skyrius

# Kiti neparametriniai kriterijai

## 6.1. Ženklų kriterijus

### 6.1.1. Įvadas: parametrinis ženklų kriterijus

Nagrinėsime Bernulio bandymų schemą. Kiekviename bandyme įvyksta įvykis $\{+\}$ arba jam priešingas įvykis $\{-\}$. Sakykime, atlikome $n$ bandymų. Įvykių $\{+\}$ ir $\{-\}$ pasirodymo skaičių pažymėkime atitinkamai $S_1$ ir $S_2 = n - S_1$.

Atsitiktinis dydis $S_1$ turi binominį skirstinį $S_1 \sim B(n,\, p)$; čia $p = \mathbf{P}\{+\}$.

Ženklų kriterijus skirtas hipotezei apie įvykių $\{+\}$ ir $\{-\}$ pasirodymo tikimybių lygybę, t. y. hipotezei $H_0 : p = 0,5$, tikrinti.

**Ženklų kriterijus:** kai alternatyva yra dvipusė $H_3 : p \neq 0,5$, tai hipotezė $H_0$ atmetama ne didesnio kaip $\alpha$ reikšmingumo lygmens kriterijumi, kai

$$S_1 \leq c_1 \quad \text{arba} \quad S_1 \geq c_2, \tag{6.1.1}$$

čia $c_1$ yra maksimalus sveikasis skaičius, tenkinantis nelygybę

$$\mathbf{P}\{S_1 \leq c_1\} = \sum_{k=0}^{c_1} C_n^k (1/2)^n = 1 - I_{0,5}(c_1+1, n-c_1) = I_{0,5}(n-c_1, c_1+1) \leq \alpha/2,$$

ir $c_2$ – minimalus sveikasis skaičius, tenkinantis nelygybę

$$\mathbf{P}\{S_1 \geq c_2\} = \sum_{k=c_2}^{n} C_n^k (1/2)^n = I_{0,5}(c_2, n-c_2+1) \leq \alpha/2;$$

čia

$$I_{0,5}(a,b) = \frac{1}{B(a,b)} \int_0^{0,5} x^{a-1}(1-x)^{b-1}dx, \quad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx.$$

Jei alternatyvos vienpusės: $H_1 : p > 0,5$ arba $H_2 : p < 0,5$, tai hipotezė atmetama, kai atitinkamai $S_1 \geq d_2$ arba $S_1 \leq d_1$; čia $d_1$ ir $d_2$ tenkina analogiškas nelygybes, kaip ir $c_1$ ir $c_2$ pakeičiant $\alpha/2$ į $\alpha$.

$P$ reikšmės yra

$$pv = 2\min\{I_{0,5}(n-S_1, S_1+1), I_{0,5}(S_1, n-S_1+1)\} \quad \text{(dvipusė alternatyva } H_3\text{)};$$

$$pv = I_{0,5}(S_1, n-S_1+1) \quad \text{(vienpusė alternatyva } H_1\text{)};$$

$$pv = I_{0,5}(n-S_1, S_1+1) \quad \text{(vienpusė alternatyva } H_2\text{)}.$$

**Didelės imtys.** Jeigu $n$ didėja, tai remiantis Muavro ir Laplaso CRT

$$(S_1 - n/2)/\sqrt{n/4} \xrightarrow{d} U \sim N(0,\, 1).$$

Kadangi $S_1 - S_2 = 2S_1 - n$, $S_1 + S_2 = n$, tai

$$Z = \frac{(S_1 - S_2)}{\sqrt{S_1 + S_2}} = \frac{(S_1 - n/2)}{\sqrt{n/4}} \xrightarrow{d} N(0,\, 1).$$

**Asimptotinis ženklų kriterijus:** jei $n$ didelis, tai dvipusės hipotezės $H_3$ atveju hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $|Z| > z_{\alpha/2}$.

Vienpusių alternatyvų $H_1$ arba $H_2$ asimptotinių kriterijų kritinės sritys apibrėžiamos atitinkamai nelygybėmis $Z > z_\alpha$ arba $Z < -z_\alpha$.

Vidutinio didumo imtims rekomenduojama naudoti tolydumo pataisą.

**Asimptotinis ženklų kriterijus su tolydumo pataisa:** Kai alternatyva dvipusė, hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$Z^* = \frac{(|S_1 - S_2| - 1)^2}{S_1 + S_2} > \chi^2_\alpha(1).$$

**6.1.1 pavyzdys.** Psichologas apklausia 39 blogai besimokančių vaikų tėvus, norėdamas išsiaiškinti, kaip gerai jie supranta problemas, kurios laukia jų vaikų užaugus. Jis nustatė, kad $S_2 = 22$ atvejais problemas geriau suprato tėvas, o $S_1 = 17$ atvejų – motina. Ar yra pagrindo daryti išvadą, kad tėvai šiuo požiūriu supratingesni už motinas? Reikšmingumo lygmuo $\alpha = 0,1$. Kokia būtų išvada, jei psichologas nustatytų, kad $S_2 = 32$ atvejais problemas geriau suprato tėvas, o $S_1 = 7$ atvejais – motina?

Patikrinsime hipotezę, kad motinos ne mažiau supratingos už tėvus, kai alternatyva yra vienpusė, tvirtinanti, kad tėvai supratingesni. Jei hipotezė bus atmesta, galėsime tvirtinti, kad duomenys neprieštarauja tėvų „pranašumui“. Pažymėkime $p$ tikimybę, kad motinos supratingesnės už tėvus. Reikia patikrinti hipotezę $H : p \geq 0,5$, kai alternatyva yra $H_2 = p < 0,5$. Hipotezė atmetama, kai $S_1$ įgyja mažas reikšmes.

Pirmuoju atveju $n = 39$, $S_1 = 17$, taigi $P$ reikšmė yra

$$pv = I_{0,5}(39 - 17,\, 17 + 1) = I_{0,5}(22,\, 18) = 0,261 > 0,1.$$

Duomenys neprieštarauja hipotezei, t. y. vyrų pranašumo hipotezę atmetame. Kadangi

$$Z = (17 - 22)/\sqrt{17 + 22} = -0,8006 > z_{0,9} = -1,282,$$

tai gauname tą pačią išvadą; asimptotinė $P$ reikšmė $pv_a = \Phi(-0,8006) = 0,217$.

Antru atveju $n = 39$, $S_1 = 7$, taigi $P$ reikšmė yra

$$pv = I_{0,5}(39 - 7,\ 7 + 1) = I_{0,5}(32,\ 8) = 0,000035 < 0,1.$$

Kriterijus atmeta hipotezę, taigi duomenys neprieštarauja vyrų „pranašumo“ alternatyvai. Kadangi

$$Z = (7 - 32)/\sqrt{7 + 32} = -4,0032 < z_{0,9} = -1,282,$$

tai asimptotinė $P$ reikšmė yra $pv_a = \Phi(-4,0032) = 0,000031$.

Ženklų kriterijus taikomas ir kai kurioms neparametrinėms hipotezėms tikrinti. Paminėsime keletą tipinių situacijų (taip pat žr. Maknemaros kriterijų, 6.3 skyrelį).

## 6.1.2. Hipotezė dėl a. d. skirtumo medianos

Sakykime, kad

$$(X_1,\ Y_1)^T, \ldots, (X_n,\ Y_n)^T \tag{6.1.2}$$

yra nepriklausomi a. v., turintys absoliučiai tolydų tikimybinį skirstinį.

Pažymėkime $\theta_i$ skirtumo $D_i = X_i - Y_i$ medianą.

**Hipotezė dėl skirtumo $D_i = X_i - Y_i$ medianos lygybės nuliui:**

$$H_0:\ \theta_1 = \cdots = \theta_n = 0.$$

Ši hipotezė nėra ekvivalenti a. d. $X_i$ ir $Y_i$ medianų lygybei. Yra skirstinių, kai a. d. $X_i$ ir $Y_i$ turi vienodas medianas, tačiau jų skirtumo $D_i = X_i - Y_i$ mediana nelygi nuliui. Skirtumų medianos lygios 0, jeigu a. v. $(X_i, Y_i)^T$ skirstiniai yra simetriški koordinačių atžvilgiu.

Esant teisingai hipotezei teigiamų skirtumų skaičiaus $S = \sum\limits_{D_i>0} 1$ skirstinys sutampa su parametrinio ženklų kriterijaus $S_1$ skirstiniu. Todėl hipotezei $H_0$ tikrinti tinka tikslūs arba asimptotiniai kriterijai, pateikti pirmesniame skyrelyje pakeičiant $S_1$ į $S$.

Jeigu skirtumų $D_i = X_i - Y_i$ skirstiniai, kai teisinga hipotezė ir kai teisinga alternatyva yra vienodi ir simetriški, tai hipotezei dėl skirtumų medianos reikšmės tikrinti geriau naudoti Vilkoksono ranginį ženklų kriterijų, grindžiamą statitika

$$T^+ = \sum_{i:D_i>0} R_i.$$

Ši statistika pilniau panaudoja informaciją, nes ji priklauso ne tik nuo skirtumų $D_i$ ženklų, bet ir nuo jų didumų $|D_i|$.

Tačiau, kaip buvo minėta, Vilkoksono ranginis ženklų kriterijus gali būti neefektyvus, jei skirtumų $D_i$ skirstiniai yra asimetriški. Tokioje situacijoje ženklų kriterijus gali pasirodyti efektyvesnis, nes jis tinka platesnei alternatyvų klasei.

**6.1.2 pavyzdys.** (**5.1.1** pavyzdžio tęsinys). Naudodami **5.1.1** pavyzdžio duomenis patikrinsime hipotezę, kad stebimo a. v. $(X, Y)^T$ koordinačių skirtumo mediana lygi nuliui. Gauname, kad iš 50 skirtumų teigiamų yra 28. Kai alternatyva dvipusė, $P$ reikšmė

$$pv = 2\min\{I_{0,5}(22,\ 29),\ I_{0,5}(28,\ 23)\} = 0,4798.$$

Taikydami normaliąją aproksimaciją su tolydumo pataisa gauname $Z^* = (|S_1 - S_2| - 1)^2/(S_1 + S_2) = 0,48$ ir $pv_a = \mathbf{P}\{\chi_1^2 > 0,48\} = 0,4884$. Duomenys neprieštarauja iškeltai hipotezei.

**6.1.1 pastaba.** Frydmano kriterijus $k = 2$ atveju yra ekvivalentus ženklų kriterijui.

Iš tikrųjų $D_i > 0$ yra ekvivalentu tam, kad $R_{i1} = 2$, $R_{i2} = 1$, ir $D_i < 0$ yra ekvivalentu tam, kad $R_{i1} = 1$, $R_{i2} = 2$. Taigi

$$\bar{R}_{.1} = 1 + S_1/n, \quad \bar{R}_{.2} = 2 - S_1/n,$$

ir

$$S_F = \frac{12n}{6} 2(S_1/n - 1/2)^2 = \frac{(S_1 - n/2)^2}{n/4}.$$

### 6.1.3. Hipotezė dėl medianos reikšmės

Sakykime, $X_1, ..., X_n$ yra paprastoji imtis absoliučiai tolydžiojo a. d. $X$. Pažymėkime $\theta$ atsitiktinio dydžio $X$ medianą.

**Hipotezė dėl medianos reikšmės:** $H_0 : \theta = \theta_0$.

Jeigu hipotezė $H_0$ teisinga, tai skirtumai $D_i = X_i - \theta_0$ įgyja teigiamas ir neigiamas reikšmes su vienodomis tikimybėmis $1/2$. Imkime statistiką

$$S = \sum_{X_i - \theta_0 > 0} 1,$$

kuri reiškia teigiamų skirtumų $D_i$ skaičių. Statistikos $S$ skirstinys sutampa su parametrinio ženklų kriterijaus statistikos $S_1$ skirstiniu. Taigi hipotezei tikrinti pritaikomi 6.1.1 skyrelyje pateikti tikslūs ar asimptotiniai kriterijai, pakeičiant $S_1$ į $S$.

**6.1.3 pavyzdys.** (**5.6.2** pavyzdžio tęsinys). Pagal **5.6.2** pavyzdžio duomenis patikrinkime hipotezę, kad stebimo a. d. mediana a) didesnė už 15; b) lygi 15. Teigiamų skirtumų skaičius $S = 15$. $P$ reikšmė yra a) $pv = I_{0,5}(n-S, s+1) = 0,0047$; b) $pv = 2I_{0,5}(n-S, s+1) = 0,0094$. Taikydami normaliąją aproksimaciją su tolydumo pataisa gauname asimptotines $P$ reikšmes a) $pv_a = \Phi(-2,5714) = 0,0051$; b) $pv_a = \mathbf{P}\{\chi_1^2 > 6,6122\} = 0,0101$. Hipotezė atmestina.

## 6.2. Serijų kriterijus

**6.2.1 apibrėžimas.** *Serija* vadinama vieno tipo įvykių seka, prieš kurią ir po kurios įvyksta kitokio tipo arba joks įvykis.

Nagrinėsime ilgio $N = m + n$ seką, sudarytą iš $m$ įvykių $A$ ir $n$ jam priešingų įvykių $\bar{A}$. Skirtingų tokio tipo sekų skaičius yra $C_N^m = C_N^n$.

Žymėsime $V$ serijų skaičių minėtoje sekoje. Pavyzdžiui, sekoje

$$A\,A\,\bar{A}\,A\,A\,\bar{A}\,\bar{A}\,A\,\bar{A}$$

yra $V = 6$ serijos; $m = 5$, $n = 4$, $N = 9$.

**Dviejų įvykių atsitiktinio išsidėstymo hipotezė:** jeigu $m$ ir $n$ fiksuoti, tai kiekvienos iš galimų $C_N^m$ sekos pasirodymas yra vienodai galimas.

Rasime serijų skaičiaus skirstinį, kai įvykiai išsidėsto atsitiktinai.

**6.2.1 teorema.** *Kai įvykių atsitiktinio išsidėstymo hipotezė teisinga, serijų skaičiaus skirstinys turi tokį pavidalą:*

$$\mathbf{P}\{V = 2i\} = \frac{2C_{m-1}^{i-1}C_{n-1}^{i-1}}{C_N^m}, \quad i = 1, ..., \min(m, n),$$

$$\mathbf{P}\{V = 2i+1\} = \frac{C_{m-1}^{i-1}C_{n-1}^{i} + C_{m-1}^{i}C_{n-1}^{i-1}}{C_N^m}, \quad i = 1, ..., \min(m, n). \tag{6.2.1}$$

**Įrodymas.** Visų galimų sekų su $m$ įvykių $A$ ir $n$ įvykių $\bar{A}$ skaičius yra $C_N^m$.

Ieškosime skaičiaus tokių sekų, kurioms $V = v$. Tarkime, kad $v = 2i$ yra lyginis. Tada turime $i$ serijų, sudarytų iš įvykių $A$, ir $i$ serijų, sudarytų iš įvykių $\bar{A}$. Keliais būdais galima sudalinti seką iš $m$ įvykių $A$ į $i$ dalių? Įsivaizduokime, kad dalijant seką į dalis yra pastatomos pertvaros tarp simbolių $A$. Tokių pertvarų reikia pastatyti $i-1$, o vietų joms pastatyti yra $m-1$. Taigi seką iš $m$ simbolių $A$ galima sudalinti į $i$ dalių $C_{m-1}^{i-1}$ būdais. Analogiškai, seką iš $n$ simbolių $\bar{A}$ galima sudalinti į $i$ dalių $C_{n-1}^{i-1}$ būdais. Elementariųjų įvykių, palankių įvykiui $V = 2i$, skaičius yra $2C_{m-1}^{i-1}C_{n-1}^{i-1}$. Iš dviejų dauginama, nes pirmoji serija gali prasidėti įvykiu $A$ arba įvykiu $\bar{A}$. Pagal klasikinės tikimybės apibrėžimą

$$\mathbf{P}\{V = 2i\} = \frac{2C_{m-1}^{i-1}C_{n-1}^{i-1}}{C_N^m}, \quad i = 1, ..., \min(m, n).$$

Analogiškai skaičiuojame elementariųjų įvykių, palankių įvykiui $V = 2i + 1$, skaičių. Galimi du atvejai: yra $i + 1$ serija, sudaryta iš simbolių $A$, ir $i$ serijų, sudarytų iš simbolių $\bar{A}$, arba $i$ serijų iš $A$ ir $i + 1$ serija iš $\bar{A}$. Taigi

$$\mathbf{P}\{V = 2i+1\} = \frac{C_{m-1}^{i-1}C_{n-1}^{i} + C_{m-1}^{i}C_{n-1}^{i-1}}{C_N^m}, \quad i = 1, ..., \min(m, n).$$

▲

Serijų skaičiaus vidurkis ir dispersija yra

$$\mathbf{E}V = \frac{2mn}{N} + 1, \quad \boldsymbol{V}V = \frac{2mn(2mn - N)}{N^2(N-1)}.$$

Kai $m, n \to \infty$, $m/n \to p \in (0,\ 1)$,

$$Z_{m,n} = \frac{V - \mathbf{E}V}{\sqrt{\boldsymbol{V}V}} \xrightarrow{d} Z \sim N(0,\ 1). \tag{6.2.2}$$

## 6.2.1. Dviejų įvykių atsitiktinio išsidėstymo hipotezė

Serijų skaičiaus statistiką naudosime suformuluotai atsitiktinio įvykių išsidėstymo hipotezei tikrinti.

Tarkime, kad hipotezė neteisinga ir įvykiai $A$ ir $B$ išsidėsto neatsitiktinai. Apie tai liudytų, pavyzdžiui, tokio tipo sekų $AAAAAAABBBBBBB$,

$ABBBBBBAAAAAAB$, kuriose sekų skaičius mažas, pasirodymas. Kita vertus, tokio tipo sekos $ABABABABABABAB$ pasirodymas taip pat rodo, kad įvykiai kaitaliojasi determinuota tvarka, o ne išsidėsto atsitiktinai. Taigi hipotezę dėl atsitiktinio įvykių išsidėstymo reikėtų atmesti, kai serijų skaičius yra per daug didelis arba per daug mažas.

**Serijų kriterijus hipotezei dėl atsitiktinio įvykių išsidėstymo tikrinti:** hipotezė atmetama ne didesnio reikšmingumo lygmens kaip $\alpha$ kriterijumi, kai $V \leq c_1$ arba $V \geq c_2$; čia $c_1$ yra maksimalus svekasis skaičius, tenkinantis nelygybę $\mathbf{P}\{V \leq c_1 | H_0\} \leq \alpha/2$, ir $c_2$ – minimalus sveikasis skaičius, tenkinantis nelygybę $\mathbf{P}\{V \geq c_2 | H_0\} \leq \alpha/2$.

**6.2.1 pavyzdys.** Tarkime, ilgio $k = 11$ sekoje įvykių $A$ ir $B$ skaičiai yra $k_1 = 5$ ir $k_2 = 6$. Tikrinama hipotezė dėl įvykių atsitiktinio išsidėstymo. Apskaičiuosime $P$ reikšmę, kai stebėta statistikos $V$ reikšmė $v$ yra a) $v = 2$; b) $v = 3$; c) $v = 4$. Remdamiesi $P$ reikšmės apibrėžimu, kai alternatyva dvipusė, gauname

$$pv = 2\min\{F_V(v), 1 - F_V(v-)\} = 2\min\{\mathbf{P}\{V \leq v\}, 1 - \mathbf{P}\{V < v\}\}.$$

Pagal (6.2.1)

$$\mathbf{P}\{V = 2\} = 2\frac{C_4^0 C_5^0}{C_{11}^5} = \frac{2}{462} = 0,004329, \quad \mathbf{P}\{V = 3\} = \frac{C_4^1 C_5^0 + C_4^0 C_5^1}{C_{11}^5} = \frac{9}{462} = 0,019481,$$

$$\mathbf{P}\{V = 4\} = 2\frac{C_4^1 C_5^1}{C_{11}^5} = \frac{40}{462} = 0,086580,$$

taigi

a) $pv = 2\min\{0,004329,\ 1\} = 0,008658$;
b) $pv = 2\min\{0,023810,\ 0,9956716\} = 0,047619$;
c) $pv = 2\min\{0,110390,\ 0,952381\} = 0,220780$.

Tokias pačias $P$ reikšmes gautume, jei $V = 11,\ 10,\ 9$, atitinkamai.

Kai $m$ ir $n$ yra dideli, kriterijų sudarome remdamiesi (6.2.2) normaliąja aproksimacija.

**Asimptotinis serijų kriterijus:** atsitiktinio įvykių išsidėstymo hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ serijų kriterijumi, kai $|Z_{k_1,k_2}| \geq z_{\alpha/2}$.

Jeigu $k$ nėra labai didelis, rekomenduojama naudoti tolydumo pataisą.

**Asimptotinis serijų kriterijus su tolydumo pataisa:** hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|Z_{k_1,k_2}^*| = |\frac{|V - \mathbf{E}V| - 0,5}{\sqrt{\mathbf{V}V}}| \geq z_{\alpha/2}.$$

Ši aproksimacija gali būti netiksli, jeigu vienas iš skaičių $m$ arba $n$ yra mažas. Tokioje situacijoje vietoje aproksimacijos normaliuoju skirstiniu rekomenduojama naudoti aproksimaciją binominiu skirstiniu (6.2.4).

**6.2.2 pavyzdys.** Atlikus $n = 40$ nepriklausomų eksperimentų, kurių metu galėjo įvykti įvykis $A$ arba jam priešingas įvykis $B$, gauti tokie rezultatai:

$$ABBBBBBBABBBABBBBBBBBABBBAABBAABABAAAAAA$$

Reikia patikrinti hipotezę, kad įvykiai $A$ ir $B$ išsidėsto atsitiktinai. Randame, kad serijų skaičius $V = 15$, $k_1 = 15, k_2 = 25$. Pagal normaliąją aproksimaciją su tolydumo pataisa gauname $Z^*_{k_1,k_2} = 1,4549$ ir $pv_a = 2(1 - \Phi(1,4549)) = 0,1457$. Pagal aproksimaciją binominiu skirstiniu (6.2.4) gauname $N = 34,185$, $p = 0,5192$ ir $0,1134 = 2I_{0,4808}(22,\,14) < pv_a < I_{0,4808}(21,\,14) = 0,1536$. Tiesiškai interpoliuodami gauname $pv_a \approx 0,1462$. Duomenys neprieštarauja iškeltai hipotezei.

## 6.2.2. Serijų kriterijus atsitiktinumo hipotezei tikrinti

Tarkime, kad stebimas a. v. $\mathbf{X} = (X_1, ..., X_n)^T$.

**Imties atsitiktinumo hipotezė $H_0^*$:** atsitiktinis vektorius $\mathbf{X}$ yra paprastoji imtis, t. y. atsitiktiniai dydžiai $X_1, ..., X_n$ vienodai pasiskirstę n. a. d.

Atsitiktinumo hipotezės alternatyva gali būti egzistavimas trendo, arba cikliškas stebėjimų kitimas, ir pan.

Pažymėkime $D_i = X_i - \hat{M}$; čia $M$ yra a. d. $X_1$ mediana esant teisingai hipotezei, o $\hat{M}$ – empirinė mediana:

$$\hat{M} = \begin{cases} (X_{(\frac{n}{2}+1)} + X_{(\frac{n}{2})})/2, & \text{kai } n \text{ lyginis} \\ X_{(\frac{n+1}{2})}, & \text{kai } n \text{ nelyginis;} \end{cases}$$

čia $X_{(i)}$ yra $i$-oji pozicinė statistika. Jeigu atsitiktinumo hipotezė teisinga, tai a. d. $D_i = X_i - \hat{M}$ yra vienodai pasiskirstę.

Išmeskime iš imties tuos elementus, kuriems $D_i = 0$. Pažymėkime $k_1$ ir $k_2$ įvykių $A_i = \{D_i > 0\}$ ir $B_i = \{D_i < 0\}$ skaičių; $k = k_1 + k_2$. Jeigu teisinga hipotezė $H_0^*$, tai gautoje sekoje įvykiai $A$ ir $B$ išsidėsto atsitiktinai. Jeigu hipotezė dėl atsitiktinio įvykių išsidėstymo atmetama, tai reikėtų atmesti ir hipotezę $H_0^*$.

**6.2.3 pavyzdys.** (**5.4.1** ir **5.4.2** pavyzdžio tęsinys). Pagal **5.4.1** pratimo duomenis patikrinsime atsitiktinumo hipotezę naudodami serijų kriterijų.

Natūralu tarti, kad a. d. $X_i$ mediana yra žinoma: $M_i = 0$. Tada šiame pavyzdyje $k_1 = 17$, $k_2 = 13$, $k = k_1 + k_2 = 30$.

Įvykių $A$ ir $B$ seka yra

$$AAAAAABABAABBBBBAAAAAAAABBBBBB.$$

Taigi

$$V = 8, \quad \mathbf{E}V = \frac{2 \cdot 17 \cdot 13}{30} + 1 = 15,733333;$$

$$\boldsymbol{V}V = \frac{2 \cdot 17 \cdot 13(2 \cdot 17 \cdot 13 - 30)}{30^2 \cdot 29} = 6,977165.$$

Modifikuota statistika (su tolydumo pataisa) įgijo reikšmę

$$|Z^*_{k_1,k_2}| = |\frac{|V - \mathbf{E}V| - 0,5}{\sqrt{\boldsymbol{V}V}}| = 2,738413.$$

Asimptotinė $P$ reikšmė

$$pv_a = 2(1 - \Phi(|Z^*_{k_1,k_2}|)) = 2(1 - \Phi(2,738413)) = 0,006174.$$

Atsitiktinumo hipotezė atmetama, nes $P$ reikšmė yra maža.

Reikia pažymėti, kad šiame pavyzdyje serijų kriterijus pasirodė galingesnis už Spirmeno ar Kendalo atsitiktinumo kriterijus, tačiau mažiau galingas už Bartelio ir Neimano atsitiktinumo kriterijų.

Jeigu šiame pavyzdyje medianą vertintume pagal turimą imtį:

$$\hat{M} = (X_{(15)} + X_{(16)})/2 = (1+2)/2 = 1,5,$$

tai gautume, kad teigiamų ir neigiamų skirtumų skaičius atitinkamai yra $k_1 = 15$, $k_2 = 15$. Įvykiai $A$ ir $B$ išsidėsto taip:

$$AAAAAABABAABBBBBAAAABBAABBBBBB.$$

Taigi serijų skaičius $V = 10$. Statistika su tolydumo pataisa įgijo reikšmę

$$|Z^*_{k_1k_2}| = 2,043864, \quad \text{ir} \quad pv_a = 0,040967.$$

Atsitiktinumo hipotezė atmetama asimptotiniu serijų kriterijumi, jeigu reikšmingumo lygmuo yra $0,05$.

### 6.2.3. Valdo ir Volfovičiaus dviejų imčių homogeniškumo kriterijus

Tarkime, kad turime dvi paprastąsias imtis $\mathbf{X} = (X_1, ..., X_m)^T$ ir $\mathbf{Y} = (Y_1, ..., Y_n)^T$, gautas stebint nepriklausomus tolydžiuosius a. d. $X \sim F$ ir $Y \sim G$.

**Homogeniškumo hipotezė**: $H_0 : F(x) = G(x), \ \forall x \in \mathbf{R}$.

Surašę abi imtis į vieną bendrą variacinę eilutę ir praleidę indeksus gausime seką, susidedančią iš $m$ simbolių $X$ ir $n$ simbolių $Y$. Valdo ir Volfovičiaus kriterijus grindžiamas serijų skaičiumi $V$ minėtoje sekoje.

Jei teisinga homogeniškumo hipotezė, tai simboliai $X$ ir $Y$ išsidėstę atsitiktinai ir serijų skaičius turi tendenciją įgyti reikšmes, artimas vidurkiui $\mathbf{E}(V|H_0)$. Jei teisinga alternatyva $H_3 = H_1 \cup H_2$; čia

$$\bar{H}_1 : \ F(x) \le G(x), \ \exists x_0 : F(x_0) < G(x_0),$$

$$\bar{H}_2 : \ F(x) \ge G(x), \ \exists x_0 : F(x_0) > G(x_0),$$

tai vienos rūšies simboliai turi tendenciją koncentruotis vienoje, o kitos rūšies – kitoje variacinės eilutės pusėje, taigi serijų skaičius dažniau įgis mažesnes reikšmes, negu kad esant teisingai hipotezei.

**Valdo ir Volfovičiaus kriterijus:** hipotezė $H_0$ atmetama ne didesnio už $\alpha$ reikšmingumo lygmens kriterijumi, kai $V \le k$; čia $k$ – didžiausias sveikasis skaičius, tenkinantis nelygybę

$$\mathbf{P}\{V \le k|H_0\} \le \alpha.$$

Jeigu $V = v$, tai $P$ reikšmė $pv = F_V(v) = \mathbf{P}\{V \le v|H_0\}$ gali būti surasta naudojantis (6.2.1) formulėmis. Mažiems $m$ ir $n$ kritinės reikšmės yra tabuliuotos [7]. Jų reikšmes taip pat galima rasti naudojant daugumą matematinės statistikos programų paketų.

**6.2.3 pavyzdys.** Tarkime, kad $m = 5$, $n = 6$, $N = 5 + 6 = 11$. Apskaičiuosime $P$ reikšmę kai stebėtos serijų skaičiaus $V$ yra a) $v = 2$; b) $v = 3$; c) $v = 4$.

Pasinaudoję (6.2.1), gauname

$$pv = F_V(v) = \mathbf{P}\{V \le v\},$$

$$\mathbf{P}\{V=2\} = 2\frac{C_4^0 C_5^0}{C_{11}^5} = \frac{2}{462} = 0,004329, \quad \mathbf{P}\{V=3\} = \frac{C_4^1 C_5^0 + C_4^0 C_5^1}{C_{11}^5} = \frac{9}{462} = 0,019481,$$

$$\mathbf{P}\{V=4\} = 2\frac{C_4^1 C_5^1}{C_{11}^5} = \frac{40}{462} = 0,086580. \tag{6.2.3}$$

Taigi

a) $pv = 0,004329$; b) $pv = 0,023810$; c) $pv = 0,110390$.

Kai $m$ ir $n$ dideli, naudojame (6.2.2) normaliąją aproksimaciją.

**Asimptotinis Valdo ir Volfovičiaus kriterijus:** jei $m$ ir $n$ yra dideli, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $Z_{mn} \le -z_\alpha$.

Kai $N = m + n$ yra vidutinio didumo sudarant kriterijų rekomenduojama naudoti tolydumo pataisą. Pažymėkime

$$Z_{mn}^* = \begin{cases} \frac{V-\mathbf{E}V-0,5}{\sqrt{\mathbf{V}_V}}, & \text{if } V-\mathbf{E}V > 0,5, \\ \frac{V-\mathbf{E}V+0,5}{\sqrt{\mathbf{V}_V}}, & \text{if } V-\mathbf{E}V < -0,5, \\ 0, & \text{if } \mid V-\mathbf{E}V \mid \le 0,5. \end{cases}$$

**Asimptotinis Valdo ir Volfovičiaus kriterijus su tolydumo pataisa:** hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $Z_{mn}^* \le -z_\alpha$.

Jeigu $m$ ir $n$ (m<n) yra vidutinio didumo arba kai santykis $m/n$ yra mažas, tiksliau aproksimuojama naudojant binominį skirstinį (žr. [7]): a. d. $V-2$ skirstinys aproksimuojamas binominiu $B(N,p)$ su parametrais

$$N = \frac{(m+n-1)(2mn-m-n)}{m(m-1)+n(n-1)}, \quad p = 1 - \frac{2mn}{(m+n)(m+n-1)}. \tag{6.2.4}$$

**Asimptotinis Valdo ir Volfovičiaus kriterijus naudojant binominę aproksimaciją:** hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai serijų skaičius $V$ yra mažesnis už atitinkamą binominio skirstinio $B(N,p)$ kritinę reikšmę.

Jeigu $V = v$, tai asimptotinė šio kriterijaus $P$ reikšmė yra

$$pv_a = \sum_{i=0}^{v} C_{N+2}^i p^i (1-p)^{N+2-i} = I_{1-p}(N+2-v, v+1) = 1 - I_p(v+1, N+2-v).$$

**6.2.1 pastaba.** Kai alternatyva yra poslinkio, Valdo ir Volfovičiaus kriterijaus galia mažesnė už Vilkoksono kriterijaus galią, todėl rekomenduotinas Vilkoksono kriterijus. Tačiau kai alternatyva yra ir mastelio, ir poslinkio ar dar bendresnė, tai Valdo ir Volfovičiaus kriterijus gali būti priimtinesnis.

**6.2.4 pavyzdys (4.5.1 pavyzdžio tęsinys).** Pagal **4.5.1** pratimo duomenis Valdo ir Volfovičiaus kriterijumi patikrinsime hipotezę, kad fungicidų naudojimas neturi įtakos medelių sergamumui.

Šiuo duomenis jau analizavome naudodami Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Vilkoksono kriterijus.

Imčių didumai $m = n = 7$, $N = m + n = 14$.

Išdėstę visus stebėjimus į vieną bendrą variacinę eilutę ir pažymėję pirmos imties stebėjimus simboliu (A), o antros imties simboliu (B), gauname tokią simbolių seką:

$$AAAAABBAABBBBB$$

Serijų skaičius $V = 4$. Naudodami paketą SPSS gauname $P$ reikšmę $pv = 0,025058$.

Esant teisingai hipotezei serijų skaičiaus vidurkis ir dispersija yra

$$\mathbf{E}V = \frac{2mn}{m+n} + 1 = 8, \quad \boldsymbol{V}V = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)} = 3,230769.$$

Kadangi $V - \mathbf{E}V = -4 < -0,5$, tai

$$Z^*_{m,n} = \frac{V - \mathbf{E}V + 0,5}{\sqrt{\boldsymbol{V}V}} = -1,94722.$$

Asimptotinė $P$ reikšmė $pv_a = \Phi(-1,94722) = 0,025754$. Hipotezė $H_0$ atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,025058.

Šiame pavyzdyje Valdo ir Volfovičiaus kriterijus yra galingesnis už Kolmogorovo ir Smirnovo dviejų imčių kriterijų, tačiau mažiau galingas, negu Kramero ir Mizeso arba Vilkoksono kriterijai.

## 6.3. Maknemaros kriterijus

Pateiksime ženklų kriterijaus modifikaciją, kai imtys priklausomos.

Tarkime, marginalieji a. v. $(X_i, Y_i)^T$, $i = 1, \ldots, n$, koordinačių skirstiniai yra *Bernulio*, t. y.

$$X_i \sim B(1, p_{i1}), \quad Y_i \sim B(1, p_{i2}), \quad p_{i1} = \mathbf{P}\{X_i = 1\}, \quad p_{i2} = \mathbf{P}\{Y_i = 1\};$$

čia a. d. $X_i$ ir $Y_i$ įgyja reikšmę 1, kai tam tikras įvykis $A$ įvyksta, ir reikšmę 0 priešingu atveju.

**Homogeniškumo hipotezė:**

$$H_0: \; p_{i1} = p_{i2} \quad \text{su visais} \quad i = 1, \ldots, n.$$

Homogeniškumo hipotezei tikrinti galima naudoti modifikuotąjį Frydmano kriterijų (atvejis $k = 2$), grindžiamą statistika $S^*_F$ (žr. (5.9.6)). Matysime, kad šį kriterijų galima suformuluoti ne rangų, o a. d. $X_i, Y_i$ terminais. Taip užrašytas modifikuotasis Frydmano kriterijus vadinamas Maknemaros kriterijumi. Nors formaliai žiūrint šis kriterijus yra parametrinis, tačiau yra nusistovėjusi tradicija jį priskirti neparametrinių kriterijų klasei.

Pateiksime tipines situacijas, kuriomis naudojamas šis kriterijus.

**6.3.1 pavyzdys.** Tarkime, kad prieš rinkimų debatus $n$ įvairių visuomenės sluoksnių žmonių paprašyta atsakyti į klausimą: „ar Jūs balsuosite už kandidatą N?“ Po debatų tų pačių žmonių prašoma dar kartą atsakyti į tą patį klausimą. Reikia nuspręsti, ar žmonių nuomonė pakito. Įvykis $A$ yra

$$A = \{\text{aš balsuosiu už kandidatą N}\},$$

$$p_{i1} = \mathbf{P}\{X_{i1} = 1\} \quad \text{ir} \quad p_{i2} = \mathbf{P}\{X_{i2} = 1\}$$

tikimybės, kad atitinkamai prieš debatus ir po jų $i$-asis apklausos dalyvis numato balsuoti už kandidatą $N$.

**6.3.2 pavyzdys.** Tiriamas galvos skausmą mažinančių vaistų efektyvumas. Apie kiekvieną iš dviejų vaistų tie patys įvairaus amžiaus asmenys atsako į klausimą: „Ar vaistas palengvina galvos skausmą?" Norima įsitikinti, ar abu vaistai vienodai efektyvūs. Šiuo atveju įvykis

$$A = \{\text{galvos skausmas palengvėjo}\},$$

o $p_{i1}$ ir $p_{i2}$ yra tikimybės, kad $i$-asis apklausos dalyvis mano, jog atitinkamai pirmasis ir antrasis vaistas palengvina galvos skausmus.

Nagrinėsime alternatyvą

$$H_3 = H_1 \cup H_2;$$

čia

$$H_1 : \ p_{i1} \le p_{i2} \quad \text{su visais} \quad i = 1, \dots, n, \text{egzistuoja} \quad i_0, \text{kad} \quad p_{i_0 1} < p_{i_0 2},$$

$$H_2 : \ p_{i1} \ge p_{i2} \quad \text{su visais} \quad i = 1, \dots, n, \text{egzistuoja} \quad i_0, \text{kad} \quad p_{i_0 1} > p_{i_0 2}.$$

**6.3.1 teorema.** *Hipotezė $H_0$ ekvivalenti tvirtinimui*

$$\mathbf{P}\{X_i = 1, Y_i = 0\} = \mathbf{P}\{X_i = 0, Y_i = 1\} \quad \textit{su visais} \quad i = 1, \dots, n.$$

*Alternatyvos $H_1$ ir $H_2$ ekvivalenčios analogiškam tvirtinimui, pakeičiant lygybę atitinkamomis nelygybėmis.*

**Įrodymas.** Nagrinėkime sąlygines tikimybes

$$\gamma_i = \mathbf{P}\{Y_i = 1 | X_i = 1\} \quad \text{ir} \quad \beta_i = \mathbf{P}\{Y_i = 0 | X_i = 0\}.$$

Tada pagal pilnosios tikimybės formulę

$$p_{i2} = \mathbf{P}\{X_i = 1\} = \gamma_i p_{i1} + (1 - \beta_i)(1 - p_{i1}).$$

Taigi lygybė $p_{i1} = p_{i2}$ ekvivalenti tvirtinimui

$$(1 - \gamma_i) p_{i1} = (1 - \beta_i)(1 - p_{i1}) \Leftrightarrow \mathbf{P}\{Y_i = 0 | X_i = 1\} \mathbf{P}\{X_i = 1\} =$$

$$\mathbf{P}\{Y_i = 1 | X_i = 0\} \mathbf{P}\{X_i = 0\} \Leftrightarrow \mathbf{P}\{X_i = 1, Y_i = 0\} = \mathbf{P}\{X_i = 0, Y_i = 1\}.$$

Antroji teoremos dalis įrodoma analogiškai, keičiant lygybę atitinkamomis nelygybėmis. ▲

Remiantis teorema pakanka nagrinėti tik tuos objektus, kuriems pirmojo ir antrojo bandymo rezultatai yra skirtingi, t. y. įvyko įvykis

$$\{X_i = 1, Y_i = 0\} \cup \{X_i = 0, Y_i = 1\}.$$

Kai hipotezė teisinga

$$\mathbf{P}\{X_i = 1, Y_i = 0 | \{X_i = 1, Y_i = 0\} \cup \{X_i = 0, Y_i = 1\}\} =$$

$$\mathbf{P}\{X_i = 0, Y_i = 1 | \{X_i = 1, Y_i = 0\} \cup \{X_i = 0, Y_i = 1\}\} = 0,5.$$

Todėl kriterijus sudaromas taip: pažymėkime $U_{kl}$ skaičių tokių objektų, kuriems

$$(X_i, Y_i) = (k, l),\ k,\ l = 0,\ 1; \quad U_{00} + U_{01} + U_{10} + U_{11} = n.$$

Stebėjimo rezultatus galime surašyti į 6.3.1 lentelę.

**6.3.1 lentelė.** Statistiniai duomenys

| k/l | 0 | 1 | |
|---|---|---|---|
| 0 | $U_{00}$ | $U_{01}$ | $U_{0.}$ |
| 1 | $U_{10}$ | $U_{11}$ | $U_{1.}$ |
| | $U_{.0}$ | $U_{.1}$ | $n$ |

Remdamiesi teorema ir diskusija po jos, sudarydami kriterijų naudojame tik $m = U_{10} + U_{01}$ objektų stebėjimus.

Esant teisingai hipotezei $H_0$ sąlyginis statistikos $U_{10}$ skirstinys, kai $m = U_{10} + U_{01}$ fiksuotas, yra binominis $B(m, 1/2)$. Taigi šis skirstinys yra lygiai toks pat kaip ženklų kriterijaus statistikos $S_1$, imant $n = m$.

**Maknemaros kriterijus:** homogeniškumo hipotezė, kai alternatyva dvipusė, yra atmetama ne didesnio už $\alpha$ reikšmingumo lygmens kriterijumi, kai

$$U_{10} \le c_1 \quad \text{arba} \quad U_{10} \ge c_2, \tag{6.3.1}$$

čia $c_1$ yra didžiausias sveikasis skaičius, tenkinantis nelygybę

$$\mathbf{P}\{U_{10} \le c_1\} = \sum_{k=0}^{c_1} C_m^k (1/2)^m = 1 - I_{0,5}(c_1+1, m-c_1) = I_{0,5}(m-c_1, c_1+1) \le \alpha/2,$$

o $c_2$ minimalus sveikasis skaičius, tenkinantis nelygybę

$$\mathbf{P}\{U_{10} \ge c_2\} = \sum_{k=c_2}^{m} C_m^k (1/2)^m = I_{0,5}(c_2,\ m - c_2 + 1) \le \alpha/2.$$

Jei $u$ yra stebėtoji statistikos $U_{10}$ reikšmė, tai $P$ reikšmė (žr. 1.4 skyrelį) yra

$$pv = 2\min(F_{U_{10}}(u), 1 - F_{U_{10}}(u)).$$

Remdamiesi 6.1.1 skyrelio rezultatais, gauname

$$Q_2 = \frac{(U_{10} - U_{01})^2}{U_{10} + U_{01}} \xrightarrow{d} \chi^2(1), \quad \text{kai} \quad m \to \infty.$$

**Asimptotinis Maknemaros kriterijus:** jei $m$ yra didelis, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$Q_2 > \chi^2_\alpha(1). \tag{6.3.2}$$

Vidutinėms $m$ reikšmėms chi kvadrato skirstiniu geriau aproksimuojama modifikuota statistika, gaunama atsižvelgiant į tolydumo pataisą:

$$Q_2^* = \frac{(|U_{10} - U_{01}| - 1)^2}{U_{10} + U_{01}}.$$

**Asimptotinis Maknemaros kriterijus su tolydumo pataisa:** jei $m$ yra didelis, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$Q_2^* > \chi^2_\alpha(1). \tag{6.3.3}$$

**6.3.3 pavyzdys.** (**6.3.1** pavyzdžio tęsinys). Tarkime, prieš debatus 1 000 rinkėjų buvo užduotas klausimas: „Ar Jūs balsuosite už kandidatą N?“ Po debatų tiems patiems rinkėjams vėl buvo užduotas tas pats klausimas. Rezultatai pateikti lentelėje.

| | 0 | 1 | $\sum$ |
|---|---|---|---|
| 0 | 421 | 115 | 536 |
| 1 | 78 | 386 | 464 |
| $\sum$ | 499 | 501 | 1000 |

Gauname $m = U_{10} + U_{01} = 193$, $U_{10} = 78$ ir $pv = \mathbf{P}\{U_{10} \leq 78|p = 0,5\} = 0,00938$. Hipotezė atmestina.

**6.3.4 pavyzdys.** Dviem skirtingais klasifikatoriais suskirstoma 500 objektų į dvi grupes. Pirmasis neteisingai suklasifikavo 80 objektų, o antrasis – 60 objektų. Kartais apie klasifikatorių kokybę sprendžiama tiesiog palyginant klaidų dažnius. Tačiau kadangi klasifikuojami tie patys objektai, tai stebėjimai yra priklausomi ir toks palyginimas nėra korektiškas. Šiam tikslui reikėtų naudoti Maknemaros kriterijų.

Tarkime, papildomai yra žinoma, kad abu klasisfikatoriai teisingai suklasifikavo tuos pačius $U_{00} = 400$ objektų ir klaidingai suklasifikavo tuos pačius $U_{11} = 40$ objektų. Hipotezė, kad antrasis klasifikatorius geresnis, gali būti suformuluota kaip parametrinė hipotezė $H : p = 0,5$, kai alternatyva vienpusė $\bar{H} : p < 0,5$ dėl binominio skirstinio tikimybės. Eksperimentų skaičiumi reikia imti $m = U_{10} + U_{01} = 60$, o dominančio įvykio įvykimų skaičius $U_{10} = 20$. Gauname $P$ reikšmę $pv = \mathbf{P}\{U_{10} \leq 20\} = 0,0067$. Pagal normaliąją aproksimaciją su tolydumo pataisa gauname $pv_a = 1 - \Phi(2,4529) = 0,0071$. Darome išvadą, kad antrasis klasifikatorius yra geresnis.

**6.3.5 pavyzdys.** 10 ekspertų tikrina dviejų tipų produktus ir padaro išvadas: atitinka standartą ar jo neatitinka. Reikia patikrinti hipotezę, kad abu produktai yra vienodos kokybės. Gauti rezultatai pateikti lentelėje (1 – produktas atitinka standartą; 0 – neatitinka).

| Ekspertas | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Pirmasis tipas | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| Antrasis tipas | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |

Gauname:

$$U_{10} = 1, \quad U_{01} = 4, \quad U_{10} - U_{01} = -3, \quad m = U_{10} + U_{01} = 5.$$

Maknemaros kriterijus suvedamas į hipotezės $H : p = 1/2$ dėl binominio skirstinio tikimybės reikšmės grindžiamą sėkmių skaičiumi $U_{10}$, kai Bernulio eksperimentų skaičius yra $m$.

Dvipusio kriterijaus atveju $P$ reikšmė yra

$$pv = 2\min(F_{U_{10}}(1), 1 - F_{U_{10}}(0)).$$

Kadangi

$$F_{U_{10}}(0) = \frac{1}{2^5}C_5^0 = \frac{1}{32}, \quad F_{U_{10}}(1) = \frac{1}{2^5}(C_5^0 + C_5^1) = \frac{3}{16},$$

gauname $pv = \frac{3}{16} = 0,1875$.

Pagal asimptotinį Maknemaros kriterijų gauname $Q = \frac{9}{5} = 1,8$ ir

$$pv_a = 1 - F_{\chi_1^2}(1,8) = 0,1797.$$

Abiem atvejais atmesti hipotezę nėra pagrindo, nes $P$ reikšmė nėra maža.

## 6.4. Kochrano kriterijus

Apibendrinsime Maknemaros kriterijų tuo atveju, kai stebimo a. v. dimensija $k > 2$.

Tarkime, kad stebimi nepriklausomi a. v. $(X_{i1}, \ldots, X_{ik})^T$, $i = 1, \ldots, n$, kurių marginalieji skirstiniai yra Bernulio:

$$X_{i1} \sim B(1, p_{i1}), \ldots, X_{ik} \sim B(1, p_{ik}); \quad p_{i1} = \mathbf{P}\{X_{i1} = 1\}, \ldots, p_{ik} = \mathbf{P}\{X_{ik} = 1\};$$

čia bet kuris a. d. $X_{i1}$ įgyja reikšmę 1, jei įvyksta tam tikras įvykis $A$, ir reikšmę 0, jei įvykis $A$ neįvyksta.

Visus stebėjimus $X_{ij}$ surašykime į 6.4.1 lentelę.

**6.4.1 lentelė.** Statistiniai duomenys

| $i$ $j$ | 1 | 2 | ... | $k$ | $\Sigma$ |
|---|---|---|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ | ... | $X_{1k}$ | $X_{1.}$ |
| 2 | $X_{21}$ | $X_{22}$ | ... | $X_{2k}$ | $X_{2.}$ |
| . | ... | ... | ... | ... | ... |
| $n$ | $X_{n1}$ | $X_{n2}$ | ... | $X_{nk}$ | $X_{n.}$ |
| $\Sigma$ | $X_{.1}$ | $X_{.2}$ | ... | $X_{.k}$ | |

Šioje lentelėje naudojami žymėjimai:

$$X_{i.} = \sum_{j=1}^{k} X_{ij}, \quad i = 1, ..., n, \quad \text{ir} \quad X_{.j} = \sum_{i=1}^{n} X_{ij}, \quad j = 1, ..., k.$$

**Homogeniškumo hipotezė:**

$$H_0: \ p_{i1} = ... = p_{ik}, \quad \text{su visais } i = 1, \ldots, n. \qquad (6.4.1)$$

Pateiksime keletą tipinių situacijų, kai reikia tikrinti tokio tipo hipotezę.

**6.4.1 pavyzdys.** Tarkime, jog $k$ skirtingais metodais nustatoma, kad virusą turi $n$ įvairaus amžiaus individų. Naudojant $j$-ąjį metodą $i$-ajam individui gaunamas vienas iš dviejų rezultatų: $X_{ij} = 1$ (įvyksta įvykis $A = \{\text{virusas rastas}\}$) arba $X_{ij} = 0$ (įvyksta įvykis $\bar{A} = \{\text{viruso nerasta}\}$).

Reikia patikrinti hipotezę, kad visi $k$ metodai yra ekvivalentūs. Šiame pavyzdyje $p_{ij}$ yra tikimybė, kad $j$-uoju metodu virusas surastas $i$-ajam individui.

**6.4.2 pavyzdys.** Lyginamas galvos skausmą raminančių $k$ vaistų efektyvumas $n$ pacientams, kurių profesijos gali būti skirtingos. Kiekvienam pacientui pateikiami klausimai: „Ar $j$-ojo tipo vaistas palengvina galvos skausmą?“, $j = 1, 2, ..., k$. Jei $j$-asis vaistas ($j = 1, ..., k$) palengvina galvos skausmą $i$-ajam individui, tai kintamasis $X_{ij}$ įgyja reikšmę 1, antraip jis įgyja reikšmę 0. Reikia patikrinti hipotezę, kad visų $k$ vaistų efektyvumas vienodas. Šiuo atveju įvykis $A$ yra {galvos skausmas palengvėja}, ir $p_{ij}$ yra tikimybė, kad $i$-ajam individui $j$-asis vaistas palengvina galvos skausmą.

**6.4.3 pavyzdys.** Dešimt ekspertų tikrina 4 tipų produktus ir „atitinka standartui“ (kintamasis įgijo reikšmę 1) arba „neatitinka standartui“ (kintamasis įgijo reikšmę 0. Reikia patikrinti hipotezę, kad visų tipų produktai vienodos kokybės.

Hipotezės $H_0$ alternatyva:

$$H_1: \quad \text{egzistuoja} \quad i, j, l: \; p_{ij} \neq p_{il}.$$

Jeigu a. d. $X_{ij}$ skirstiniai nepriklauso nuo indeksų $i$, t. y. $p_{ij} = p_j$ su visais $i = 1, ..., n$, tai alternatyva įgauna paprastesnį pavidalą:

$$H_1: \quad \text{egzistuoja} \quad j, l: p_j \neq p_l.$$

Kochrano kriterijus sudaromas taip. Jei teisinga hipotezė $H_0$, tai a. d. $X_{.1}, ..., X_{.k}$ yra vienodai pasiskirstę, taigi skirtumai $X_{.j} - \bar{X}$ įgyja artimas 0 reikšmes; čia

$$\bar{X} = \frac{1}{k} \sum_{j=1}^{k} X_{.j}\, .$$

Skirtumų $X_{.j} - \bar{X}$ sklaidą apie 0 apibūdina

$$\sum_{j=1}^{k} (X_{.j} - \bar{X})^2 = \sum_{j=1}^{k} X_{.j}^2 - k\bar{X}^2.$$

Ši statistika proporcinga Kochrano statistikai:

$$Q = \frac{k(k-1)\left(\sum_{j=1}^{k} X_{.j}^2 - k\bar{X}^2\right)}{k\sum_{i=1}^{n} X_{i.} - \sum_{i=1}^{n} X_{i.}^2}. \tag{6.4.2}$$

**6.4.1 pastaba.** Kochrano statistika sutampa su modifikuotąja Frydmano statistika (5.9.6).

Iš tikrųjų, kadangi $X_{ij}$ įgyja tik dvi reikšmes: 1 arba 0, tai kiekvienoje eilutėje yra tiktai viena arba dvi sutampančios duomenų grupės. Jei grupės dvi, tai $k_i = 2$, o grupių dydžiai yra $t_{i1} = X_{i.}$ ir $t_{i2} = k - X_{i.}$ Jei grupė viena, tai $k_i = 1$, o jos dydis $t_{i1} = k$.

Jei $k_i = 1$, tai $T_i = k^3 - k$. Jei $k_i = 2$, tai

$$T_i = \sum_{j=1}^{2}(t_{ij}^3 - t_{ij}) = X_{i.}^3 + (k - X_{i.})^3 - (X_{i.} + (k - X_{i.})) = k(3X_{i.}^2 - 3kX_{i.} + k^2 - 1).$$

Abiem atvejais $T_i = k(3X_{i.}^2 - 3kX_{i.} + k^2 - 1)$, nes jei $k_i = 1$, tai $X_{i.} = 0$ arba $X_{i.} = k$. Taigi

$$T_i = k(3\cdot 0^2 - 3k\cdot 0 + k^2 - 1) = k^3 - k \quad \text{ir} \quad T_i = k(3k^2 - 3kk + k^2 - 1) = k^3 - k.$$

Todėl

$$1 - \frac{1}{n(k^3 - k)}\sum_{i=1}^{n} T_i = \frac{3}{n(k^2 - 1)}(k\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i^2). \tag{6.4.3}$$

Surikiavę lentelės 6.4.1 $i$-ąją eilutę gausime, kad $k - X_{i.}$ nuliukai yra pozicijose nuo 1 iki $k - X_{i.}$, o vienetukai tolimesnėse pozicijose iki $X_{i.}$.

Jei $X_{ij} = 0$, tai

$$R_{ij} = \frac{1 + 2 + \cdots + (k - X_{i.}}{k - X_{i.}} = (k - X_{i.} + 1)/2.$$

Jei $X_{ij} = 1$, tai

$$R_{ij} = \frac{(k - X_{i.}) + (k - X_{i.} + 1) + \cdots + k}{X_{i.}} = k/2 + (k - X_{i.} + 1)/2.$$

Taigi

$$\sum_{j=1}^{k} R_{.j}^2 = \sum_{j=1}^{k}(\frac{1}{2}\sum_{i=1}^{n}(k - X_{i.} + 1) + \frac{k}{2}X_{.j})^2 = \sum_{j=1}^{k}(\frac{k}{2}(X_{.j} - \bar{X}) + n\frac{k+1}{2})^2$$

$$= \frac{k^2}{4}\sum_{j=1}^{k}(X_{.j} - \bar{X})^2 + \frac{n^2k(k+1)^2}{4}.$$

Gauname

$$S_F = \frac{12}{k(k+1)n}\sum_{j=1}^{k} R_{.j}^2 - 3n(k+1) = \frac{3k}{n(k+1)}\sum_{j=1}^{k}(X_{.j} - \bar{X})^2.$$

Remdamiesi (5.9.6) ir $S_F$ apibrėžimu, turime

$$S_F^* = \frac{S_F}{1 - \frac{1}{n(k^3-k)}\sum_{i=1}^{n} T_i} = \frac{k(k-1)\left(\sum_{j=1}^{k} X_{.j}^2 - k\bar{X}^2\right)}{k\sum_{i=1}^{n} X_{i.} - \sum_{i=1}^{n} X_{i.}^2} = Q.$$

**Kochrano kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens ne didesnio už $\alpha$ kriterijumi, kai $Q \geq Q_\alpha$; čia $Q_\alpha$ yra mažiausias realus skaičius $c$, tenkinantis nelygybę $\mathbf{P}\{Q \geq c | H_0\} \leq \alpha$.

Kriterijaus $P$ reikšmė yra $pv = \mathbf{P}\{Q \geq q\}$; čia $q$ yra statistikos $Q$ realizacija.

Jeigu $n$ yra didelis, tai remiantis 5.9.2 teorema Kochrano statistikos skirstinys aproksimuojamas chi kvadrato skirstiniu su $k-1$ laisvės laipsniu.

**Asimptotinis Kochrano kriterijus:** jei $n$ yra didelis, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$Q > \chi^2_\alpha(k-1). \tag{6.4.4}$$

**6.4.4 pavyzdys.** 10 ekspertų vertina 4 tipų produktų kokybę ir padaro išvadas „atitinka standartą“ (kintamojo reikšmė 1) arba „neatitinka standarto“ (kintamojo reikšmė 0). Duomenys surašyti į 6.4.2 lentelę.

**6.4.2 lentelė.** Statistiniai duomenys

| $i\,j$ | 1 | 2 | 3 | 4 | $X_{i.}$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 3 |
| 2 | 1 | 1 | 1 | 0 | 3 |
| 3 | 0 | 1 | 0 | 1 | 2 |
| 4 | 0 | 0 | 1 | 1 | 2 |
| 5 | 1 | 1 | 0 | 0 | 2 |
| 6 | 1 | 1 | 1 | 0 | 3 |
| 7 | 1 | 1 | 1 | 0 | 3 |
| 8 | 0 | 0 | 0 | 1 | 1 |
| 9 | 1 | 1 | 0 | 0 | 2 |
| 10 | 0 | 1 | 0 | 1 | 2 |
| $X_{.j}$ | 6 | 8 | 5 | 4 | 23 |

Reikia patikrinti hipotezę, kad visų tipų produktai yra vienodos kokybės. Šiame pavyzdyje $k = 4$, $n = 10$,

$$\sum_{i=1}^{10} X_{i.}^2 = 4 \cdot 3^2 + 5 \cdot 2^2 + 1^2 = 57, \quad \sum_{j=1}^{4} X_{.j}^2 = 6^2 + 8^2 + 5^2 + 4^2 = 141,$$

$$\sum_{j=1}^{4} X_{.j} = \sum_{i=1}^{10} X_{i.} = 23, \quad \bar{X} = \frac{23}{4} = 5,75,$$

$$Q = \frac{k(k-1)\left(\sum_{j=1}^{k} X_{.j}^2 - k\bar{X}^2\right)}{k\sum_{i=1}^{n} X_{i.} - \sum_{i=1}^{n} X_{i.}^2} = \frac{4 \cdot 3(141 - 4 \cdot 5,75^2)}{4 \cdot 23 - 57} = 3.$$

Naudodami SPSS paketą gauname $P$ reikšmę $pv = 0,466732$. Asimptotinė $P$ reikšmė $pv_a = \mathbf{P}\{\chi^2_3 > 3\} = 0,391625$. Duomenys neprieštarauja iškeltai hipotezei.

**6.4.2 pastaba.** Maknemaros kriterijus yra atskiras atvejis Kochrano kriterijaus, kai $k = 2$.

Iš tikrųjų, jei $k = 2$, tai

$$X_{i.} = \begin{cases} 0, & \text{kai} \quad X_{i1} = 0, X_{i2} = 0, \\ 1, & \text{kai} \quad X_{i1} = 1, X_{i2} = 0 \text{ arba } X_{i1} = 0, X_{i2} = 1, \\ 2, & \text{kai} \quad X_{i1} = 1, X_{i2} = 1. \end{cases}$$

Taigi galioja tokie 6.3.1 ir 6.4.2 lentelių duomenų ryšiai:

$$\sum_{i=1}^{n} X_{i.} = X_{1.} + X_{2.} = U_{10} + U_{01} + 2U_{11}, \quad \sum_{i=1}^{n} X_{i.}^2 = U_{10} + U_{01} + 4U_{11}.$$

Gauname, kad (6.4.2) vardiklis yra

$$2\sum_{j=1}^{n} X_{j.} - \sum_{j=1}^{n} X_{j.}^2 = U_{10} + U_{01},$$

o skaitiklis

$$X_{.1}^2 + X_{.2}^2 - 2\bar{X}^2 = X_{.1}^2 + X_{.2}^2 - \frac{1}{2}(X_{.1} + X_{.2})^2 = \frac{(X_{.1} - X_{.2})^2}{2} = \frac{(U_{10} - U_{01})^2}{2},$$

Įstatę į statistikos $Q$ išraišką (6.4.3), gauname

$$Q = \frac{(U_{10} - U_{01})^2}{U_{10} + U_{01}},$$

o tai ir yra Maknemaros kriterijaus statistika.

## 6.5. Specialieji suderinamumo kriterijai

Tegu $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis, gauta stebint a. d. $X$, kurio pasiskirstymo funkcija $F$ priklauso neparametrinei šeimai $\mathcal{F}$. Nagrinėsime sudėtines suderinamumo hipotezes.

**Sudėtinė suderinamumo hipotezė:**

$$H_0 : F \in \mathcal{F}_0 = \{F_0(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subset \mathbf{R}^s\} \subset \mathcal{F}, \tag{6.5.1}$$

čia $F_0(x|\boldsymbol{\theta})$ yra žinoma specialaus pavidalo pasiskirstymo funkcija, priklausanti nuo nežinomo baigtinės dimensijos parametro $\boldsymbol{\theta} \in \Theta$.

Specialiais šeimų $\mathcal{F}_0$ atvejais kartais pavyksta rasti tokias statistikas, kurių skirstiniai esant teisingai hipotezei $H_0$ nepriklauso nuo nežinomų parametrų ir tam tikra prasme apibūdina šeimą $\mathcal{F}_0$. Tokias statistikas galima panaudoti sudarant kriterijus sudėtinėms suderinamumo hipotezėms tikrinti.

Būtent taip sudaryti modifikuotasis chi kvadrato, modifikuotieji Neimano ir Bartono, Kolmogorovo ir Smirnovo, Kramero ir Mizeso ir Anderseno ir Darlingo kriterijai (žr. 2.3, 3.5, 4.4 skyrelius) tikrinti hipotezes dėl skirstinio priklausymo specialioms poslinkio ir mastelio šeimoms (normalusis, logistinis, ekstremalių reikšmių, Koši skirstiniai), bei specialioms mastelio ir laipsnio šeimoms (lognormalusis, Veibulo, loglogistinis skirstiniai).

Šiame skyrelyje pateikiame dar keletą tokio tipo kriterijų.

### 6.5.1. Normalusis skirstinys

Tarkime, kad $\mathcal{F}_0$ yra normaliųjų skirstinių šeima $\{N(\mu, \sigma^2), -\infty < \mu < \infty, 0 < \sigma < \infty\}$.

**1. Kriterijai grindžiami empirinių momentų funkcijomis.** Normaliojo a. d. $X \sim N(\mu, \sigma^2)$ asimetrijos ir eksceso koeficientai lygūs 0:

$$\gamma_1 = \mathbf{E}(X - \mu)^3/\sigma^3 = 0, \quad \gamma_2 = \mathbf{E}(X - \mu)^4/\sigma^4 - 3 = 0. \tag{6.5.2}$$

Pirmasis centrinis absoliutinis momentas

$$\gamma_3 = \frac{\mathbf{E}|X-\mu|}{\sigma} = \sqrt{\frac{2}{\pi}}.$$

Asimetrijos ir eksceso koeficientų empiriniai analogai yra

$$g_1 = m_3/m_2^{3/2}, \quad g_2 = m_4/m_2^2 - 3, \quad m_k = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^k, \quad k = 2, 3, 4, \tag{6.5.3}$$

o parametro $\gamma_3$ empirinis analogas

$$g_3 = \frac{1}{n}\sum_{j=1}^{n}|X_j - \bar{X}|/\sqrt{m_2}. \tag{6.5.4}$$

Atsitiktinio dydžio $Y = (X-\mu)/\sigma$, o ir statistikų $g_j$ skirstiniai nepriklauso nuo nežinomų parametrų $\mu$ ir $\sigma$. Jeigu $g_1, g_2$ arba $g_3$ daug skiriasi atitinkamai nuo $\gamma_1 = 0$, $\gamma_2 = 0$ arba $\gamma_3 = \sqrt{2/\pi}$, tai hipotezę $H_0$ reikėtų atmesti.

Tiksliau tikriname hipotezes

$$H_1 : \gamma_1 = 0, \quad H_2 : \gamma_2 = 0, \quad H_3 : \gamma_3 = \sqrt{2/\pi},$$

kad stebimo dydžio parametrai $\gamma_j$ yra tokie kaip normaliojo skirstinio.

**Kriterijai grindžiami empirinių momentų funkcijomis:** hipotezė $H_j$ atmetama, kai $g_j < c_{1j}$ arba $g_j > c_{2j}$; čia $c_{1j}$ ir $c_{2j}$ yra statistikos $g_j$, $j = 1, 2, 3$, atitinkamai $(1-\alpha/2)$ ir $\alpha/2$ kritinės reikšmės. Kaip specialūs normališkumo kriterijai pirmieji du buvo pasiūlyti D'Agostinjo [10]; trečiasis pasiūlytas Geri [12]. Literatūroje kriterijai grindžiami statistikomis $g_1$ ir $g_2$ vadinami D'Agostinjo kriterijais, o kriterijus, grindžiamas statistika $g_3$, – Geri kriterijumi.

Nedideliems $n$ statistikų $g_j$ kritinės reikšmės yra tabuliuotos (žr. [7]); jų reikšmes taip pat galima rasti naudojant kai kuriuos matematinės statistikos paketus, arba modeliuojant.

Kai $n$ yra didelis, statistikų $g_j$ skirstiniai aproksimuojami normaliuoju skirstiniu.

Statistikų vidurkiai ir dispersijos yra

$$\mathbf{E}g_1 = 0, \quad \mathbf{E}g_2 = -\frac{6}{n+1}, \quad \mathbf{E}g_3 = \frac{2\Gamma((n+1)/2)}{\sqrt{\pi(n-1)}\Gamma(n/2)},$$

$$\boldsymbol{V}(g_1) = \frac{6(n-2)}{(n+1)(n+3)}, \quad \boldsymbol{V}(g_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}.$$

$$\boldsymbol{V}(g_3) = \frac{1}{n}\left\{1 + \frac{2}{\pi}[\sqrt{n(n-2)} + \arcsin\frac{1}{n-1}]\right\} - \frac{n-1}{\pi}\left[\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\right]^2.$$

**Asimptotinis kriterijus hipotezėms $H_j$ tikrinti:** jei $n$ yra didelis, tai hipotezė $H_j$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$|\bar{g}_j| = \frac{|g_j - \mathbf{E}g_j|}{\sqrt{\boldsymbol{V}(g_j)}} > z_{\alpha/2}, \quad j = 1, 2, 3. \tag{6.5.5}$$

**6.5.1 pastaba.** Pateiktus kriterijus reikėtų interpretuoti kaip nukrypimų nuo normaliojo skirstinio kriterijus. Jeigu hipotezės $H_1$, $H_2$ arba $H_3$ atmetamos, tai tuo labiau reikia atmesti normalumo hipotezę. Šių hipotezių priėmimas nereiškia, kad skirstinys yra normalusis, nes egzistuoja skirtingi nuo normaliojo skirstiniai, turintys tokius pačius parametrus $\gamma_j$, $j = 1, 2, 3$.

**6.5.2 pastaba.** Nustatyta, kad statistikos $g_2$ skirstinio konvergavimas į normalųjį yra kur kas lėtesnis negu kitų dviejų statistikų.

**6.5.1 pavyzdys.** (**2.4.2** pavyzdžio tęsinys). Pagal **2.4.2** pratimo duomenis, naudodami kriterijus, grindžiamus empirinių momentų funkcijomis, patikrinsime hipotezes, kad stebimojo a. d. skirstinys yra a) normalusis; b) lognormalusis.

a) Randame $\bar{g}_1 = 4,0697$; $\bar{g}_2 = 2,7696$; $\bar{g}_3 = -0,4213$. Atitinkamos asimptotinės $P$ reikšmės yra $4,7110^{-5}; 0,0056; 0,6735$. Remiantis kriterijais, grindžiamais empiriniais asimetrijos ir eksceso koeficientais, normalumo hipotezė atmetama.

b) Atlikę transformaciją $Y_i = \ln X_i$, $i = 1, ..., 49$, randame $\bar{g}_1 = -1,8692$; $\bar{g}_2 = 0,2276$; $\bar{g}_3 = 0,8915$ ir atitinkamos asimptotinės $P$ reikšmės yra $0,0616; 0,8200; 0,3727$. Jei parinktas reikšmingumo lygmuo $\alpha = 0,05$, tai nė vienas iš pateiktų kriterijų hipotezės neatmeta.

Kartais naudojama modifikuota D'Agostinjo statistika

$$T = n\left(\frac{g_1^2}{6} + \frac{(g_2-3)^2}{24}\right),$$

kurios skirstinys aproksimuojamas chi kvadrato skirstiniu su 2 laisvės laipsniais. Hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $T > \chi^2_\alpha(2)$.

**2. Sarkadi kriterijus**

Tegu $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis a. d. $X \sim N(\mu, \sigma^2)$. Pažymėkime

$$Y_j = X_{j+1} - \frac{1}{1+\sqrt{n}}X_1 - \frac{n}{n+\sqrt{n}}\bar{X}, \quad j = 1, ..., n-1.$$

Tada a. d. $Y_1, ..., Y_{n-1}$ (žr. 6.9 pratimą) yra vienodai pasiskirstę nepriklausomi a. d. $Y_j \sim N(0, \sigma^2)$. Apibrėžkime a. d.

$$Z_j = \frac{Y_j}{\sqrt{\frac{1}{n-j-1}(Y_{j+1}^2 + ... + Y_{n-1}^2)}}, \quad j = 1, ..., n-2,$$

kurie yra nepriklausomi ir turi Stjudento skirstinius: $Z_j \sim S(n-j-1), j = 1, ..., n-2$ (žr. 6.10 pratimą).

Pažymėkime $S(x|\nu)$ Stjudento skirstinio su $\nu$ laisvės laipsnių pasiskirstymo funkciją. Tada atsitiktiniai dydžiai $U_j = S(Z_j|n-j-1), j = 1, ..., n-2$, yra nepriklausomi ir vienodai tolygiai pasiskirstę intervale $[0,\ 1]$, t. y. $U_1, ..., U_{n-2}$ yra didumo $n-2$ paprastoji imtis a. d. $U \sim U(0,\ 1)$.

Tikrinsime hipotezę $H_0^*$, kad $(U_1, ..., U_{n-2})^T$ yra paprastoji imtis a. d. $U \sim U(0,\ 1)$. Šią hipotezę galima tikrinti naudojant bet kurį paprastosios suderinamumo hipotezės tikrinimo kriterijų: Pirsono chi kvadrato, Neimano ir Bartono, Kramero ir Mizeso, Anderseno ir Darlingo ir pan.

Jeigu hipotezė $H_0^*$ atmetama, tai normalumo hipotezę taip pat reikia atmesti.

**6.5.3 pastaba.** Apibrėžiant a. d. $Y_j$ vietoje $X_1$ galima imti bet kurį $X_m$, kai $m$ fiksuotas. Jeigu alternatyva poslinkio, tai natūralu imti $m = 1$ arba $m = n$. Jeigu žinoma, kad momentu $s$ eksperimento sąlygos galėjo pakisti, tai natūralu imti $m = s$.

**6.5.2 pavyzdys.** (**2.4.2** pavyzdžio tęsinys). Pagal **2.4.2** pratimo duomenis, naudodami Sarkadi kriterijų, patikrinsime hipotezes, kad stebimojo a. d. skirstinys yra a) normalusis; b)lognormalusis.

a) Atlikę nurodytas transformacijas gauname imtį $Z_1, ..., Z_{47}$, kuri esant teisingai hipotezei yra paprastoji imtis a. d. $Z \sim U(0,1)$. Tikrindami paprastąją hipotezę $H : Z \sim U(0,1)$ Pirsono chi kvadrato kriterijumi ir parinkę $k = 8$ vienodų tikimybių intervalus, gauname a. d. $X_n^2$ realizaciją 14,9149 ir asimptotinę $P$ reikšmę $pv_a = \mathbf{P}\{\chi_7^2 > 14,9149\} = 0,0107$. Normalumo hipotezė atmestina. Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo statistikos įgijo reikšmes 0,1508, 0,2775, 1,7396, ir joms atitinka $P$ reikšmės 0,2203, 0,1621, 0,1299. Šie kriterijai normalumo hipotezės neatmeta.

b) Perėję prie logaritmų ir paskui atlikę nurodytas transformacijas gauname imtį $Z_1, ..., Z_{47}$, kuri esant teisingai hipotezei yra paprastoji imtis a. d. $Z \sim U(0,1)$. Tikrindami paprastąją hipotezę $H : Z \sim U(0,1)$ Pirsono chi kvadrato kriterijumi ir parinkę $k = 8$ vienodų tikimybių intervalus, gauname a. d. $X_n^2$ realizaciją 1,8936 ir asimptotinę $P$ reikšmę $pv_a = \mathbf{P}\{\chi_7^2 > 1,8936\} = 0,8637$. Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo statistikos įgijo reikšmes 0,1086, 0,0999, 0,6934 ir joms atitinka $P$ reikšmės 0,6364, 0,5854, 0,5644. Visi kriterijai lognormalumo hipotezės neatmeta.

### 3. Šapiro ir Vilkso kriterijus.

Tarkime, kad $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis absoliučiai tolydžiojo a. d. su vidurkiu $\mu = \mathbf{E}X_i$ ir dispersija $\boldsymbol{V}(X_i) = \sigma^2$.

Nepaslinktieji parametrų $\mu$ ir $\sigma^2$ įvertiniai yra

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Esant normaliajam skirstiniui kvadratinis nuokrypis $\sigma$ gali būti įvertintas ir kitokiu būdu.

Pažymėkime $\mathbf{X}^{(\cdot)} = (X_{(1)}, \ldots, X_{(n)})^T$ pozicinių statistikų vektorių.

Tegu $Z_{(i)} = (X_{(i)} - \mu)/\sigma$. Kai hipotezė teisinga, $Z_{(i)}$ yra standartinio normaliojo skirstinio $i$-oji pozicinė statistika ir a. v.

$$\mathbf{Z}^{(\cdot)} = (Z_{(1)}, ..., Z_{(n)})^T$$

skirstinys nepriklauso nuo nežinomų parametrų $\mu$ ir $\sigma$.

Pažymėkime

$$\mathbf{m} = (m_{(1)}, ..., m_{(n)})^T \quad \text{ir} \quad \mathbf{\Sigma} = [\sigma_{ij}]_{n \times n},$$

atsitiktinio vektoriaus $\mathbf{Z}^{(\cdot)}$ vidurkį ir kovariacijų matricą, kurie nepriklauso nuo nežinomų parametrų $\mu$ ir $\sigma$ esant teisingai hipotezei $H_0$.

Tegu

$$\boldsymbol{\theta} = (\mu, \sigma)^T, \quad \boldsymbol{C} = \left( \ \mathbf{1} \mid \mathbf{m} \ \right) = \begin{pmatrix} 1 & ... & 1 \\ m_{(1)} & ... & m_{(n)} \end{pmatrix}^T$$

Remdamiesi lygybėmis $X_{(i)} = \mu + \sigma Z_{(i)}$ gauname, kad a. v. $\mathbf{X}^{(\cdot)}$ vidurkis ir kovariacinė matrica turi tokį pavidalą:

$$\mathbf{E}\mathbf{X}^{(\cdot)} = \boldsymbol{C}\boldsymbol{\theta}, \quad \mathbf{V} = \sigma^2 \boldsymbol{\Sigma}.$$

Turime tiesinį modelį, iš kurio mažiausiųjų kvadratų metodu galime įvertinti nežinomus parametrus:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{B}\mathbf{X}^{(\cdot)}, \quad \boldsymbol{B} = (\boldsymbol{C}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{C})^{-1} \boldsymbol{C}^T \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} c_1 & ... & c_n \\ a_1 & ... & a_n \end{pmatrix}.$$

Turime

$$\hat{\sigma} = \sum_{i=1}^n a_i X_{(i)}, \quad \hat{\sigma}^2 = (\sum_{i=1}^n a_i X_{(i)})^2,$$

čia vektorius $\mathbf{a} = (a_1, ..., a_n)^T$ turi tokį pavidalą

$$\mathbf{a} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{m}}{\mathbf{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} \mathbf{m}}, \quad \mathbf{a}^T \mathbf{a} = \sum_{i=1}^n a_i^2 = 1.$$

Reikia pažymėti, kad $\boldsymbol{BC} = \mathbf{E}$, čia $\mathbf{E}$ yra $2 \times 2$ vienetinė matrica, taigi $\sum_{i=1}^n c_i = 1$, $\sum_{i=1}^n a_i = 0$.

Šapiro ir Vilkso kriterijus grindžiamas statistika $W$, kuri proporcinga santykiui $\hat{\sigma}^2 / S_n^2$ dviejų parametro $\sigma^2$ įvertinių:

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}.$$

Remdamiesi lygybėmis $X_{(i)} = \mu + \sigma Z_{(i)}$ ir $\sum_{i=1}^n a_i = 0$ gauname, kad statistikos $W$ skirstinys nepriklauso nuo nežinomų parametrų, kai tikrinamoji hipotezė teisinga.

Kai normalumo hipotezė teisinga, tai du dispersijos įvertiniai yra artimi. Priešingu atveju šie įvertiniai dažniau įgyja tolimas reikšmes.

**Šapiro ir Vilkso kriterijus:** normalumo hipotezė atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $W < c_1$, čia $c_1$ yra maksimalus realus skaičius, tenkinantis nelygybę $\mathbf{P}\{W < c_1 | H_0\} \le \alpha$.

Daugelyje matematinės statistikos programų paketų yra numatyta rasti koeficientus $a_i$ ir statistikos $W$ kritines reikšmes radimas.

**6.5.3 pavyzdys.** (**2.4.2** pavyzdžio tęsinys). Pagal **2.4.2** pratimo duomenis, naudodami Šapiro ir Vilkso kriterijų, patikrinsime hipotezes, kad stebimojo a. d. skirstinys yra a) normalusis; b) lognormalusis.

a) Naudodami SPSS programų paketą gauname $W = 0,8706$ ir $P$ reikšmė $pv < 0,0001$. Normalumo hipotezė atmetama.

b) Perėję prie logaritmų $\ln X_i$ ir naudodami SPSS programų paketą gauname $W = 0,9608$ ir $P$ reikšmė $pv = 0,1020$. Lognormalumo hipotezė neatmetama.

### 6.5.2. Eksponentinis skirstinys

Tegu $X_1, ..., X_n$ yra paprastoji imtis absoliučiai tolydžiojo a. d. $X$ su pasiskirstymo funkcija $F$ ir tankio funkcija $f$.

**Eksponentiškumo hipotezė:** $H_0 : X \sim \mathcal{E}(\lambda)$, $\lambda > 0$.

Esant teisingai hipotezei a. d. $X$ pasiskirstymo funkcija yra $F(x) = 1 - e^{-\lambda x}, x \geq 0$.

Pateiksime hipotezės $H_0$ tikrinimo kriterijus, kurie tinka ir cenzūruotiems duomenims, kai stebima tik $r$ pirmųjų pozicinių statistikų, taip pat kriterijus, kurie pritaikomi tik pilnoms imtims.

Pirmųjų pozicinių statistikų vektoriaus $(X_{(1)}, X_{(2)}, \dots, X_{(r)})^T$ tankio funkcija turi tokį pavidalą

$$f_{X_{(1)},...,X_{(r)}}(x_1, ..., x_r) = \frac{n!}{(n-r)!}(1 - F(x_r))^{n-r} f(x_1), ..., f(x_r). \qquad (6.5.6)$$

Ji apibrėžta srityje $Q_r = \{(x_1, ..., x_r) : \quad -\infty < x_1 \leq ... \leq x_r < +\infty\}$.

Atskiru eksponentinio skirstinio atveju

$$f_{X_{(1)},...,X_{(r)}}(x_1, ..., x_r) = \frac{n!}{(n-r)!}\lambda^r e^{-\lambda(\sum_{i=1}^{r} x_i + (n-r)x_r)}, \quad 0 \leq x_1 \leq ... \leq x_r < +\infty. \qquad (6.5.7)$$

Pažymėkime

$$Z_1 = nX_{(1)}, Z_2 = (n-1)(X_{(2)} - X_{(1)}), ..., Z_r = (n-r+1)(X_{(r)} - X_{(r-1)}).$$

**6.5.1 teorema.** *Kai hipotezė $H_0$ teisinga, atsitiktiniai dydžiai $Z_1, .., Z_r$ yra nepriklausomi ir vienodai pasiskirstę. Be to, $Z_i \sim \mathcal{E}(\lambda)$.*

**Įrodymas.** Atlikę kintamųjų keitimą

$$z_1 = nx_1, \ z_2 = (n-1)(x_2 - x_1), \ \dots, \ z_r = (n-r+1)(x_r - x_{r-1})$$

gauname

$$\sum_{i=1}^{r} x_i + (n-r)x_r = \sum_{i=1}^{r} z_i, x_1 = \frac{z_1}{n}, x_2 = \frac{z_2}{n-1}, ..., x_r = \frac{z_1}{n} + \frac{z_2}{n-1} + ... + \frac{z_r}{n-r+1}$$

ir pakeitimo jakobianą $(n-r)!/n!$. Įstatę į tankio išraišką (6.5.7) ir padauginę iš jakobiano gauname a. v. $(Z_1, .., Z_r)^T$ tankio funkciją:

$$f_{Z_1,...,Z_r}(z_1, \dots, z_r) = \frac{n!}{(n-r)!}\lambda^r e^{-\lambda \sum_{i=1}^{r} z_i}, \quad z_1, \dots, z_r \geq 0, \qquad (6.5.8)$$

▲

**1. Gnedenkos kriterijus**

Tegu

$$G = \frac{r_2 \sum\limits_{i=1}^{r_1} z_i}{r_1 \sum\limits_{i=r_1+1}^{r} z_i};$$

čia $r_1 = [r/2]$, $r_2 = r - [r/2]$.

**6.5.2 teorema.** *Kai hipotezė $H_0$ teisinga, statistikos $G$ skirstinys yra Fišerio skirstinys su $2r_1$ ir $2r_2$ laisvės laipsnių.*

**Įrodymas.** Remiantis 6.5.1 teorema nepriklausomi a. d. $2\lambda Z_1, ..., 2\lambda Z_r$ turi chi kvadrato skirstinį su 2 laisvės laipsniais. Taigi a. d. $2\lambda \sum\limits_{i=1}^{r_1} z_i$ ir $2\lambda \sum\limits_{i=r_1+1}^{r} z_i$ yra nepriklausomi ir turi chi kvadrato skirstinius su $2r_1$ ir $2r_2$ laisvės laipsnių. Iš čia gauname teoremos tvirtinimą.

▲

**Gnedenkos kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $G < F_{1-\alpha/2}(2r_1, 2r_2)$ arba $G > F_{\alpha/2}(2r_1, 2r_2)$.

Kriterijaus $P$ reikšmė yra

$$pv = 2\min\{F_G(t), 1 - F_G(t)\},$$

čia $F_G$ yra pasiskirstymo funkcija a. d., turinčio Fišerio skirstinį su $2r_1$ ir $2r_2$ laisvės laipsnių.

Gnedenko kriterijus yra galingas tada, kai gedimų intensyvumo funkcija $\lambda(t)$ yra didėjanti ar mažėjanti intervale $(0, \infty)$. Tada statistika $G$ turi tendenciją įgyti didesnes arba mažesnes reikšmes, negu tuo atveju, kai teisinga eksponentiškumo hipotezė.

**6.5.4 pavyzdys.** (**2.4.1** pavyzdžio tęsinys). Pagal **2.4.1** pratimo duomenis patikrinsime hipotezę, kad buvo stebėtas eksponentinis a. d.

Statistika $G$ įgijo reikšmę 2,1350 ir ją atitinkanti $P$ reikšmė yra $pv = 2\min(\mathbf{P}\{F_{35,35} < 2,1350\}, \mathbf{P}\{F_{35,35} > 2,1350\}) = 0,0277$. Eksponentiškumo hipotezė atmestina.

**2. Bolševo kriterijus**

Pažymėkime

$$W_1 = \frac{Z_1}{Z_1 + Z_2}, \quad W_2 = \frac{Z_1 + Z_2}{Z_1 + Z_2 + Z_3}, \ \ldots, \quad W_{r-1} = \frac{\sum\limits_{i=1}^{r-1} Z_i}{\sum\limits_{i=1}^{r} Z_i}, \quad W_r = \sum_{i=1}^{r} Z_i,$$

$$U_i = W_i^i, \quad i = 1, ..., r-1.$$

**6.5.3 teorema.** *Kai hipotezė $H_0$ teisinga, a. d. $U_1, .., U_{r-1}$ yra nepriklausomi ir vienodai tolygiai pasiskirstę: $U_j \sim U(0,1), j = 1, ..., r$.*

**Įrodymas.** Atlikime kintamųjų keitimą:

$$w_1 = \frac{z_1}{z_1+z_2},\ w_2 = \frac{z_1+z_2}{z_1+z_2+z_3},\ \dots,\ w_{r-1} = \frac{\sum\limits_{i=1}^{r-1} z_i}{\sum\limits_{i=1}^{r} z_i}, w_r = \sum_{i=1}^{r} z_i.$$

Gauname

$$z_1 = w_1 w_2 ... w_r,\quad z_2 = (1-w_1)w_2 \cdots w_r, \cdots,$$

$$z_{r-1} = (1-w_{r-2})w_{r-1}w_r,\quad z_r = (1-w_{r-1})w_r.$$

Pakeitimo jakobianas yra $w_2 w_3^2 ... w_{r-1}^{r-2} w_r^{r-1}$. Įrašę kintamųjų $z_i$ išraiškas į (6.5.8) ir padauginę iš jakobiano, gausime a. v. $(W_1,..,W_r)^T$ tankio funkciją:

$$f_{W_1,..,W_r}(w_1,\dots,w_r) = \frac{\lambda^r w_r^{r-1}}{(r-1)!} e^{-\lambda w_r} 1(2w_2)(3w_3^2)...(r-1)w_{r-1}^{r-2}.$$

Taigi a. d. $W_1, W_2 ..., W_{r-1}$ yra nepriklausomi, o jų tankio funkcijos

$$f_{W_i}(w_i) = i w_i^{i-1}, \quad 0 \le w_i \le 1, \quad i = 1, ..., r-1.$$

Iš čia gauname teoremos tvirtinimą.

▲

Įrodyta [6], kad suformuluotas rezultatas yra charakteringoji eksponentinio skirstinio savybė.

Taigi hipotezė $H_0$ yra ekvivalenti hipotezei, kad $U_1, U_2 ..., U_{r-1}$ yra paprastoji didumo $r-1$ imtis, gauta stebint a. d. $U \sim U(0,1)$. Šiai hipotezei tikrinti galime naudoti bet kurį paprastosios suderinamumo hipotezės tikrinimo kriterijų: chi kvadrato, Neimano ir Bartono, Kolmogorovo ir Smirnovo, Anderseno ir Darlingo, Kramero ir Mizeso.

Pavyzdžiui, Kolmogorovo ir Smirnovo kriterijus būtų grindžiamas statistika $D_{r-1} = \max(D_{r-1}^+, D_{r-1}^-)$;

$$D_{r-1}^+ = \max_{\le k \le r-1}\left(\frac{k}{r-1} - U_{(k)}\right), \quad D_{r-1}^- = \max_{\le k \le r-1}\left(U_{(k)} - \frac{k-1}{r-1}\right),$$

o $U_{(k)}$ yra $k$-oji pozicinė statistika paprastosios imties $U_1, U_2 ..., U_{r-1}$.

**Bolševo, Kolmogorovo ir Smirnovo kriterijus:** hipotezė $H_0$ yra atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $D_{r-1} > D_\alpha(r-1)$; čia $D_\alpha(r-1)$ yra statistikos $D_{r-1}$ lygmens $\alpha$ kritinė reikšmė.

*Bolševo, Neimano ir Bartono, Bolševo, Anderseno ir Darlingo, Bolševo, Kramero ir Mizeso* kriterijai formuluojami analogiškai.

**6.5.5 pavyzdys.** (**2.4.1** pavyzdžio tęsinys). Pagal **2.4.1** pratimo duomenis patikrinsime hipotezę, kad buvo stebėtas eksponentinis a. d. Atlikę Bolševo transformaciją gauname imtį $U_1, ..., U_{69}$, kuri esant teisingai hipotezei yra paprastoji a. d. $U \sim U(0,1)$ imtis. Tikrindami šią hipotezę Pirsono chi kvadrato kriterijumi ir parinkę $k = 6$ vienodų tikimybių intervalus, gauname statistikos $X_n^2$ reikšmę 11,6087 ir ją atitinkančią asimptotinę $P$ reikšmę

$pv_a = \mathbf{P}\{\chi^2_5 > 11,6087\} = 0,0406$. Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo statistikų realizacijos yra 0,2212, 0,9829 ir 4,8406, o joms atitinkančios $P$ reikšmės yra 0,0025, 0,0034 ir 0,0041. Eksponentiškumo hipotezė atmestina.

**3. Barnardo kriterijus**

Pažymėkime

$$V_1 = \frac{Z_1}{\sum\limits_{i=1}^{r} Z_i}, \quad V_2 = \frac{Z_1 + Z_2}{\sum\limits_{i=1}^{r} Z_i}, \quad \dots \quad , \quad V_{r-1} = \frac{\sum\limits_{i=1}^{r-1} Z_i}{\sum\limits_{i=1}^{r} Z_i}, \quad V_r = \sum_{i=1}^{r} Z_i,$$

**6.5.4 teorema.** *Kai hipotezė $H_0$ teisinga, atsitiktinio vektoriaus $(V_1, .., V_{r-1})^T$ skirstinys sutampa su pozicinių statistikų vektoriaus didumo $r-1$ imtyje, gautoje stebint a.d. $V \sim U(0,1)$, skirstiniu.*

**Įrodymas.** Atlikime kintamųjų keitimą:

$$v_1 = \frac{z_1}{\sum\limits_{i=1}^{r} z_i}, \quad v_2 = \frac{z_1 + z_2}{\sum\limits_{i=1}^{r} z_i}, \quad \dots \quad , \quad v_{r-1} = \frac{\sum\limits_{i=1}^{r-1} z_i}{\sum\limits_{i=1}^{r} z_i}, \quad v_r = \sum_{i=1}^{r} z_i.$$

Gauname

$$z_1 = v_1 v_r,\ z_2 = v_r(v_2 - v_1),\ z_{r-1} = v_r(v_{r-1} - v_{r-2}),\ z_r = v_r(1 - v_{r-1}),$$

pakeitimo jakobianas yra $v_r^{r-1}$. Įstatę kintamųjų $z_i$ išraiškas į (6.5.8) ir padauginę iš jakobiano, gauname a. v. $(V_1, .., V_r)^T$ tankio funkciją:

$$f_{V_1,..,V_r}(v_1, \dots, v_r) = \frac{\lambda^r v_r^{r-1}}{(r-1)!} e^{-\lambda v_r} (r-1)!, \quad 0 \le v_1 \le \dots \le v_{r-1},\ v_r \ge 0.$$

Iš čia išplaukia teoremos rezultatas.

▲

Remdamiesi šia teorema galima tvirtinti, kad pakeitus $U_{(k)}$ į $V_k$, $k = 1, \dots, r-1$, vietoje Bolševo-Kolmogorovo-Smirnovo statistikos $D_{r-1}$ gausime statistiką $\tilde{D}_{r-1}$, kuri esant teisingai hipotezei turi tą patį skirstinį.

**Barnardo, Kolmogorovo ir Smirnovo kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $\tilde{D}_{r-1} > D_\alpha(r-1)$.

Analogiškai galime suformuluoti Pirsono chi kvadrato, Kramero ir Mizeso, Anderseno ir Darlingo, Neimano ir Bartono kriterijus, grindžiamus atsitiktiniais dydžiais, gautais atlikus Barnardo transformaciją.

**6.5.6 pavyzdys.** (**2.4.1** pavyzdžio tęsinys). Pagal **2.4.1** pratimo duomenis patikrinsime hipotezę, kad buvo stebėtas eksponentinis a. d. Atlikę Barnardo transformaciją gauname imtį $V_1, \dots, V_{69}$, kuri, kai teisinga hipotezė, yra paprastosios a. d. $V \sim U(0,\ 1)$ imties variacinė eilutė. Tikrindami šią hipotezę Pirsono chi kvadrato kriterijumi ir parinkę $k = 6$ vienodų tikimybių intervalus, gauname statistikos $X_n^2$ reikšmę 13,0 ir ją atitinkančią asimptotinę $P$ reikšmę $pv_a = \mathbf{P}\{\chi_5^2 > 13,0\} = 0,0234$. Kolmogorovo ir Smirnovo, Kramero ir Mizeso ir Anderseno, Darlingo statistikų realizacijos yra 0,2407, 1,3593, 6,6310, o joms atitinkančios $P$ reikšmės yra 0,0008, 0,0004, 0,0007. Eksponentiškumo hipotezė atmestina.

### 6.5.3. Veibulo skirstinys

Tegu $X_1, ..., X_n$ yra paprastoji imtis absoliučiai tolydžiojo a. d. $X$, kurio pasiskirstymo funkcija $F$ ir tankio funkcija $f$.

**Hipotezė dėl Veibulo skirstinio:** $H_0 : X \sim W(\theta, \nu), \theta, \nu > 0$.

Hipotezė $H_0$ ekvivalenti hipotezei, kad a. d. $Y = \ln X$ turi ekstremalių reikšmių skirstinį, kurio pasiskirstymo funkcija $F_Y(y) = 1 - e^{-e^{(x-\mu)/\sigma}}, x \in \mathbf{R}$; čia $\mu = \ln\theta, \sigma = 1/\nu$.

**1. Mano kriterijus**

Pateikiamas kriterijus tinka ir antro tipo cenzūruotoms imtims, t. y. kai stebima tik $r$ pirmųjų pozicinių statistikų. Kai imtys pilnos, pateikiamose formulėse reikia imti $r = n$.

Tegu $(X_{(1)}, X_{(2)}, ..., X_{(r)})^T$ yra $r$ pirmųjų pozicinių statistikų vektorius. Pažymėkime $Y_{(i)} = \ln X_{(i)}$, $r_1 = [r/2]$ ir

$$S_{rn} = \frac{\sum\limits_{i=r_1+1}^{r-1} (Y_{(i+1)} - Y_{(i)})/E_{in}}{\sum\limits_{i=1}^{r-1} Y_{(i+1)} - Y_{(i)})/E_{in}};$$

čia $Z_{(i)} = (Y_{(i)} - \mu)/\sigma$, $E_{in} = \mathbf{E}(Z_{(i+1)} - Z_{(i)})$. Remdamiesi lygybėmis $Y_{(i+1)} - Y_{(i)} = \sigma(Z_{(i+1)} - Z_{(i)})$ ir tuo, kad $Z_{(i)}$ skirstinys nepriklauso nuo nežinomų parametrų, gauname, kad vidurkiai $E_{in}$ ir statistikos $S_{rn}$ skirstinys taip pat nepriklauso nuo nežinomų parametrų. Mano kriterijus grindžiamas šia statistika. Pažymėkime $S_\alpha(r,n)$ statistikos $S_{rn}$ lygmens $\alpha$ kritinę reikšmę. Nedidelių $n$ koeficientai $E_{in}$ ir kritinės reikšmės $S_\alpha(r,n)$ gali būti surastos naudojant, pavyzdžiui, SAS programų paketą.

**1. Mano kriterijus:** hipotezė $H_0$ atmetama reikšmingumo lygmens $\alpha$ kriterijumi, kai $S_{rn} < S_{1-\alpha/2}(r,n)$.

**Didelės imtys.** Jeigu $n > 25$, tai statistikos $S_{rn}$ skirstinys aproksimuojamas beta skirstiniu su parametrais $r - r_1 - 1$ ir $r_1$, o koeficientai $E_{in}$ aproksimuojami taip: $E_{in} \approx 1/[-(n-i)\ln(1-i/n)]$, $i = 1, ..., r-1$.

Pažymėkime $\beta_{\alpha/2}(r - r_1 - 1, r_1)$ beta skirstinio su parametrais $r - r_1 - 1$ ir $r_1$ lygmens $\alpha$ kritinę reikšmę.

**Asimptotinis Mano kriterijus:** jeigu $n$ yra didelis, tai hipotezė $H_0$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai $S_{rn} < \beta_{1-\alpha/2}(r - r_1 - 1, r_1)$ arba $S_{rn} > \beta_{\alpha/2}(r - r_1 - 1, r_1)$.

**6.5.7 pavyzdys.** (**2.4.1** pavyzdžio tęsinys). Pavyzdžiuose **6.5.4; 6.5.5; 6.5.6** matėme, kad pagal **2.4.1** pavyzdžio duomenis eksponentiškumo hipotezė atmetama. Remdamiesi Mano kriterijumi patikrinsime hipotezę, kad ši imtis gauta stebint a. d., turintį Veibulo skirstinį. Naudodami SAS programų paketą gauname $S_{n,n} = 0,4294$ ir $P$ reikšmė yra $pv = 0,2921$. Duomenys neprieštarauja iškeltai hipotezei.

### 6.5.4. Puasono skirstinys

Tarkime, paprastoji imtis $X_1, ..., X_n$ gauta stebint diskretųjį a. d. $X$, įgyjantį sveikas neneigiamas reikšmes.

**Puasoniškumo hipotezė:** $H_0$: $X \sim \mathcal{P}(\lambda), \lambda > 0$.

Ši hipotezė tvirtina, kad stebimojo a. d. $X$ skirstinys yra Puasono.

**1. Bolševo kriterijus.** Fiksuokime sumą $T_n = X_1 + ... + X_n$. Kai hipotezė $H_0$ teisinga, sąlyginis a. v. $\mathbf{X} = (X_1, ..., X_n)^T$ skirstinys yra polinominis: $\mathbf{X} \sim \mathcal{P}_n(T_n, \boldsymbol{\pi}_0), \boldsymbol{\pi}_0 = (\pi_{i0}, ..., \pi_{n0})^T, \pi_{10} = ... = \pi_{n0} = 1/n$, t. y.

$$\mathbf{P}\{X_1 = x_1, ..., X_n = x_n | T_n\} = \frac{T_n!}{x_1!...x_n!}\left(\frac{1}{n}\right)^{T_n}.$$

Bolševas [5] įrodė, kad ši lygybė yra charakteringoji Puasono skirstinio savybė. Remdamiesi šiuo faktu vietoje hipotezės $H_0$ tikrinsime hipotezę $H_0^* : \boldsymbol{\pi} = \boldsymbol{\pi}_0$, kad polinominio skirstinio tikimybių vektorius $\boldsymbol{\pi}$ yra lygus vektoriui $\boldsymbol{\pi}_0 = (1/n, ..., 1/n)^T$, kai polinominių eksperimentų skaičius yra $T_n$.

Hipotezę $H_0^*$ galima tikrinti Pirsono chi kvadrato kriterijumi (žr. 2.1.1 skyrelį). Kriterijaus statistika

$$X_n^2 = \sum_{j=1}^{n} \frac{(X_j - T_n/n)^2}{T_n/n} = \frac{n}{T_n}\sum_{j=1}^{n} X_j^2 - n.$$

Kai hipotezė $H_0^*$ teisinga ir $T_n$ didelis, statistikos $X_n^2$ skirstinys aproksimuojamas chi kvadrato skirstiniu su $n-1$ laisvės laipsniu.

**Bolševo kriterijus:** jei $T_n$ yra didelis, tai hipotezė $H_0^*$ atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$X_n^2 < \chi^2_{1-\alpha/2}(n-1), \quad \text{arba} \quad X_n^2 > \chi^2_{\alpha/2}(n-1). \tag{6.5.9}$$

Jei hipotezė $H_0^*$ atmetama, tai hipotezė $H_0$ taip pat atmetama.

Kriterijaus statistika yra proporcinga dispersijos $\boldsymbol{V}X_i$ ir vidurkio $\mathbf{E}X_i$ nepaslinktųjų įvertinių santykiui:

$$X_n^2 = (n-1)\frac{s_n^2}{\bar{X}_n}, \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Kai alternatyvos specialios, gali būti tikslinga, kad kriterijaus kritinė sritis būtų vienpusė. Kadangi santykis $\boldsymbol{V}X_i/\mathbf{E}X_i = 1$, kai skirstinys Puasono;

$\boldsymbol{V}X_i/\mathbf{E}X_i < 1$, kai skirstinys binominis; $\boldsymbol{V}X_i/\mathbf{E}X_i > 1$, kai skirstinys neigiamas binominis, o santykis $s_n^2/\bar{X}_n$ yra santykio $\boldsymbol{V}X_i/\mathbf{E}X_i$ įvertinys, tai natūralu, jei alternatyva binominė, atmesti hipotezę,kai

$$X_n^2 < \chi^2_{1-\alpha}(n-1), \tag{6.5.10}$$

o jei yra neigiamojo binominio skirstinio alternatyva, atmesti hipotezę, kai

$$X_n^2 > \chi^2_{\alpha}(n-1).$$

**6.5.8 pavyzdys.** Panagrinėsime Ruterfordo, Geigerio ir Čedviko eksperimento, kurio metu buvo registruojamas radioaktyviosios medžiagos išspinduliuotų $\alpha$ dalelių skaičius, rezultatus (žr. [27]). Skaitiklis registravo žybsnių, kurie pasirodo atsitrenkiant $\alpha$ dalelei į specialią plokštelę, skaičių. Užregistruoti žybsnių skaičiai $X_1, ..., X_n$ per $n = 2608$ vienodo ilgio 7.5 sekundžių intervalų. Tegu $n_j$ yra intervalų, kuriuose užregistruota po $j$ dalelių skaičius, t. y. skaičius tokių a, d. $\{X = j\}$, kurie įgijo reikšmę $j$. Eksperimento rezultatai pateikti lentelėje.

| $j$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_j$ | 57 | 203 | 383 | 525 | 532 | 408 | 273 | 139 | 45 | 27 | 10 | 4 | 2 |

Gauname, kad bendras užregistruotų žybsnių skaičius yra

$$T_n = \sum_{i=1}^{n} X_i = \sum_{j=1}^{\infty} j n_j = 10094$$

ir empiriniai momentai yra:

$$\bar{X}_n = \frac{T_n}{n} = 3,8704, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{j=0}^{\infty} n_j j^2 - n\bar{X}_n^2 = 3,6756.$$

Kadangi $s_n^2/\bar{X}_n = 0,9497 < 1$, tai Mizesas padarė išvadą, kad imtis $X_1, ..., X_n$ gauta stebint binominį a. d. Ar Mizeso išvada korektiška?

Naudosime chi kvadrato kriterijų (6.5.10). Pagal turimus duomenis gauname $X_n^2 = 2475,8$, ir asimptotinė $P$ reikšmė yra

$$pv_a = \mathbf{P}\{\chi^2_{n-1} < 2475,8\} = \mathbf{P}\{\chi^2_{2607} < 2475,8\} = 0,03295.$$

Puasoniškumo hipotezė atmetama binominio skirstinio naudai, jei kriterijaus reikšmingumo lygmuo viršija $0,03295$.

Kita vertus, radioaktyviosios medžiagos išspinduliuojamų $\alpha$ dalelių srautas turėtų būti puasoninis. Šį prieštaravimą Feleris paaiškino šitaip. Naudojamas skaitiklis negali atskirti dviejų žybsnių, jeigu jie abu įvyksta mažame ilgio $\gamma > 0$ laiko intervale. Taigi, jei du ar daugiau žybsnių įvyksta laiko intervale $\gamma > 0$, tai jie registruojami kaip vienas žybsnis. Atsižvelgus į šią aplinkybę gaunama, kad duomenys neprieštarauja dalelių srauto puasoniškumo prielaidai.

## 2. Tuščių dėžių kriterijus

Jei Puasono skirstinio parametras $\lambda$ yra mažas, tai daugelis imties elementų įgis reikšmę 0 ir aproksimacija chi kvadrato skirstiniu gali būti netiksli. Tada sudarant kriterijų kartais naudojamas „tuščių dėžių“ kriterijus. Šis kriterijus grindžiamas statistika

$$Z_0 = \sum_{j=1}^{n} \mathbf{1}_{\{0\}}(X_j), \tag{6.5.11}$$

t. y. skaičiumi imties elementų $X_j$, įgijusių reikšmę 0.

Tarkime, turime $N$ rutulių ir $n$ dėžių. Rutuliai mėtomi į dėžes taip, kad kiekvienas rutulys nepriklausomai nuo kitų gali patekti į kiekvieną iš $n$ dėžių vienoda tikimybe $1/n$. Pažymėkime $Z_i = Z_i(n, N)$ skaičių tokių dėžių, kuriose po tokio eksperimento bus $i$ rutulių. Tada

$$\sum_{i=0}^{N} Z_i = n, \quad \sum_{i=0}^{N} iZ_i = N.$$

Rasime tuščių dėžių skaičiaus $Z_0 = Z_0(n, N)$ skirstinį.

**6.5.5 teorema.** *Atsitiktinio dydžio $Z_0 = Z_0(n, N)$ galimos reikšmės yra* $\max(0, n-N) \leq k \leq n-1$, *o jų įgijimo tikimybės*

$$\mathbf{P}\{Z_0(n, N) = k\} = C_n^k (1 - \frac{k}{n})^N \mathbf{P}\{Z_0(n-k, N) = 0\}, \qquad (6.5.12)$$

$$\mathbf{P}\{Z_0(n-k, N)\} = \sum_{l=0}^{n-k} C_{n-k}^l (-1)^l (1 - \frac{l}{n-k})^N.$$

*Atsitiktinio dydžio $Z_0$ vidurkis ir dispersija yra*

$$\mathbf{E} Z_0(n, N) = n(1 - \frac{1}{n})^N,$$

$$\boldsymbol{V} Z_0(n, N) = n(n-1)(1 - \frac{2}{n})^N + n(1 - \frac{1}{n})^N - n^2(1 - \frac{1}{n})^{2N}. \qquad (6.5.13)$$

**Įrodymas.** Pažymėkime $A_i$ įvykį, kad $i$-oji dėžė yra tuščia, o $\bar{A}_i$ – jam priešingą įvykį. Tada

$$\mathbf{P}\{Z_0(n, N) = k\} = \sum_{1 \leq i_1 < \ldots < i_k \leq n} \mathbf{P}\{A_{i_1} \cap \ldots \cap A_{i_k} \cap \bar{A}_{j_1} \cap \ldots \cap \bar{A}_{j_{n-k}}\};$$

čia $\{j_1, \ldots, j_{n-k}\}$ yra aibės $\{i_1, \ldots, i_k\}$ papildinys iki aibės $\{1, 2, \ldots, n\}$. Sumos dėmenų skaičius yra $C_n^k$, o tikimybės turi tokį pavidalą:

$$\mathbf{P}\{A_{i_1} \cap \ldots \cap A_{i_k}\} \mathbf{P}\{\cap \bar{A}_{j_1} \cap \ldots \cap \bar{A}_{j_{n-k}} | A_{i_1} \cap \ldots \cap A_{i_k}\} =$$

$$= (1 - \frac{k}{n})^N \mathbf{P}\{Z_0(n-k, N) = 0\}.$$

Tikimybė, kad likusios $n-k$ dėžių nėra tuščios

$$\mathbf{P}\{Z_0(n-k, N) = 0\} = 1 - \mathbf{P}\{Z_0(n-k, N) > 0\} = 1 - \mathbf{P}\{\cup_{i=1}^{n-k} A_i\} =$$

$$= 1 - \{\sum_i \mathbf{P}\{A_i\} - \sum_{i<j} \mathbf{P}\{A_i \cap A_j\} + \sum_{i<j<l} \mathbf{P}\{A_i \cap A_j \cap A_l\} - \ldots\} =$$

$$= 1 - (n-k)(1 - \frac{1}{n-k})^N + C_{n-k}^2 (1 - \frac{2}{n-k})^N - C_{n-k}^3 (1 - \frac{3}{n-k})^N + \ldots$$

Iš čia gauname (6.5.12).

Užrašykime a. d. $Z_0$ kaip sumą binominių a. d. $Z_0 = Y_1 + Y_2 + \ldots + Y_n$; atsitiktinis dydis $Y_i$ įgyja reikšmę 1, jei $i$-oji dėžė tuščia, ir reikšmę 0 priešingu atveju. Remdamiesi lygybėmis $\mathbf{P}\{Y_i = 1\} = (1 - 1/n)^N$ ir $\mathbf{P}\{Y_i Y_j = 1\} = (1 - 2/n)^N, i \neq j$, gauname

$$\mathbf{E} Z_0 = n \mathbf{E} Y_i = n(1 - \frac{1}{n})^N,$$

$$\boldsymbol{V} Z_0 = n \mathbf{E} Y_i^2 + \sum_{i \neq j} \mathbf{E}(Y_i Y_j) - (\mathbf{E} Z_0)^2 =$$

$$= n(n-1)(1 - \frac{2}{n})^N + n(1 - \frac{1}{n})^N - n^2(1 - \frac{1}{n})^{2N}. \qquad \blacktriangle$$

Kai hipotezė $H_0$ teisinga, statistika $Z_0(n, S_n)$ turi skirstinį (6.5.13), jei pakeisime $N$ į $S_n$. Jeigu hipotezė neteisinga, tai rutuliai patenka į dėžes su skirtingomis tikimybėmis, taigi statistika $Z_0(n, S_n)$ turės tendenciją įgyti didesnes reikšmes.

**Tuščių dėžių kriterijus:** hipotezė $H_0$ atmetama ne didesniu kaip $\alpha$ reikšmingumo lygmens kriterijumi, kai

$$Z_0(n, S_n) \geq c_n, \tag{6.5.14}$$

čia $c_n$ yra mažiausias sveikasis skaičius, tenkinantis nelygybę

$$\mathbf{P}\{Z_0(n, S_n) \geq c_n\} \leq \alpha.$$

Jei $n$ yra didelis, statistikos $Z_0(n, S_n)$ skirstinys aproksimuojamas normaliuoju.

**Asimptotinis tuščių dėžių kriterijus:** jei $n$ yra didelis, tai hipotezė atmetama asimptotiniu reikšmingumo lygmens $\alpha$ kriterijumi, kai

$$\frac{Z_0 - \mathbf{E}Z_0}{\sqrt{\mathbf{V}Z_0}} > z_\alpha. \tag{6.5.15}$$

**6.5.8 pavyzdys.** Veikiant ląsteles rentgeno spinduliais stebimos chromosomų mutacijos. Lentelėje pateikti chromosomų, kuriose stebėta $i$ mutacijų, skaičiai $n_i$.

| $i$ | 0 | 1 | 2 | 3 | $\sum$ |
|---|---|---|---|---|---|
| $n_i$ | 639 | 141 | 13 | 0 | 793 |

Tikrinsime hipotezę, kad mutacijų skaičius $X$ turi Puasono skirstinį, taikydami tuščių dėžių kriterijų. Turime $Z_0 = Z_0(n, S_n) = 639$, $S_n = \sum_i in_i = 167$; $\mathbf{E}Z_0 = 642,327$, $\mathbf{V}Z_0 = 12,3516$. Pagal normaliąją aproksimaciją gauname asimptotinę $P$ reikšmę $pv_a = 1 - \Phi(0,9466) = 0,1722$. Duomenys neprieštarauja iškeltai hipotezei.

**6.5.3 pastaba.** Tuščių dėžių kriterijų galima taikyti paprastajai hipotezei $H : F(x) \equiv F_0(x)$ tikrinti. Sudalinkime abscisių ašį į $k$ intervalų taškais $-\infty = a_0 < a_1 < ... < a_k = +\infty$ taip, kad $F_0(a_j) - F_0(a_{j-1}) = 1/k$, $j = 1, ..., k$. Jei hipotezė teisinga, $n$ imties elementų (rutulių) atsitiktinai mėtomi į $k$ intervalų (dėžių). Skaičius $Z_0(k, n)$ intervalų, kuriuose nėra imties elementų (tuščių dėžių), turi (6.5.12) skirstinį imant $k$ vietoje $n$ ir $n$ vietoje $S_n$. Hipotezė $H$ atmetama pagal (6.5.14) arba (6.5.15) kriterijus.

## 6.6. Pratimai

**6.1.** Remdamiesi ženklų kriterijumi patikrinkite hipotezę apie dviejų krakmolo kiekio nustatymo būdų ekvivalentiškumą pagal **5.3** pratimo duomenis.

**6.2.** Remdamiesi ženklų kriterijumi patikrinkite hipotezę apie dviejų sėjamųjų vienodą efektyvumą pagal **5.4** pratimo duomenis.

**6.3.** Remdamiesi serijų skaičiumi grindžiamu kriterijumi, patikrinkite hipotezę apie nuodų poveikio vienodumą pagal **1.43** pratimo duomenis.

**6.4.** Įrodykite, kad ženklų kriterijaus hipotezei $H : \theta = \theta_0$ tikrinti ($\theta$ – simetriško tolydžiojo skirstinio, kurio tankis $f(x)$, mediana) ASE, lyginant su Stjudento kriterijumi poslinkio alternatyvų atveju, yra $4\sigma^2(f(\theta_0))^2$

**6.5.** Naudodami serijų kriterijų patikrinkite atsitiktinumo hipotezę pagal **2.16** pratimo duomenis.

**6.6.** Naudodami serijų kriterijų patikrinkite atsitiktinumo hipotezę pagal **2.17** pratimo duomenis.

**6.7.** Visuomenės nuomonės apklausoje tų pačių 3000 rinkėjų buvo klausiama dėl jų nuomonės apie konkrečią parlamentinę partiją prieš rinkimus ir po jų. Prieš rinkimus neigiamą nuomonę išsakė 300 rinkėjų, o praėjus metams po rinkimų – 350. Be to, 150 rinkėjų nepakeitė savo neigiamos nuomonės, 150 rinkėjų neigiama nuomonė pasikeitė į teigiamą, o 200 rinkėjų teigiama nuomonė pasikeitė į neigiamą. Ar pakito partijos reitingas?

**6.8.** (**6.7** pratimo tęsinys). Tarkime, kad tie patys 3000 rinkėjų buvo apklausti prieš kitus rinkimus. 270 rinkėjų nuomonė buvo neigiama. Iš jų 180 rinkėjų buvo tokių, kurie pirmose dviejose apklausose turėjo teigiamą nuomonę, ir 70 rinkėjų, kurių nuomonė buvo teigiama vienoje ir neigiama kitoje iš pirmiau buvusių apklausų. Ar pakito partijos reitingas?

**6.9.** Tegu $X_1, ..., X_n$ yra paprastoji imtis a. d. $X \sim N(\mu, \sigma^2)$ ir $Y_i = X_{i+1} - X_1/(1+\sqrt{n}) - n\bar{X}/(n+\sqrt{n})$. Įrodykite, kad $Y_1, ..., Y_{n-1}$ yra vienodai pasiskirstę n. a. d. ir $Y_j \sim N(0, \sigma^2)$.

**6.10.** (**6.9** pratimo tęsinys). Įrodykite, kad $Z_1, ..., Z_{n-2}$ yra nepriklausomi a. d., turintys Stjudento skirstinius $Z_i \sim S(n-i-1)$, jei $Z_j = Y_j\sqrt{n-j-1}/(Y_{j+1}^2 + ... + Y_{n-1}^2)$.

**6.11.** Patikrinkite normalumo hipotezę naudodami 6.5.1 skyrelio kriterijus pagal **2.16** pratimo duomenis.

**6.12.** Patikrinkite lognormalumo hipotezę naudodami 6.5.1 skyrelio kriterijus pagal **2.16** pratimo duomenis.

**6.13.** Patikrinkite puasoniškumo hipotezę naudodami tuščių dėžių kriterijų pagal **2.18** pratimo duomenis.

## 6.7. Atsakymai

**6.1.** Tarp 13 skirtumų $S = 3$ yra neigiami ir $pv = 2\mathbf{P}\{S \leq 3\} = 0,0923$. Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,0923. **6.2.** Tarp 9 skirtumų $S = 1$ yra neigiamas ir $pv = 2\mathbf{P}\{S \leq 1\} = 0,0195$. Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,0195. **6.3.** Serijų skaičius $V = 29$ ir $P$ reikšmė $pv = \mathbf{P}\{V \geq 29\} = 1,2\dot{1}0^{-6}$. Hipotezė atmetama. **6.5.** Serijų skaičius $V = 54$, $k_1 = 50$, $k_2 = 50$. Gauname $Z^*_{k_1,k_2} = 0,5025$ ir $pv_a = 2(1 - \Phi(0,5025)) = 0,6153$. Atmesti hipotezę nėra pagrindo. **6.6.** Serijų skaičius $V = 90$, $k_1 = 76$, $k_2 = 74$. Gauname $Z^*_{k_1,k_2} = 2,4906$ ir $pv_a = 2(1 - \Phi(2,4906)) = 0,0128$. Hipotezė atmetama, jei kriterijaus reikšmingumo lygmuo viršija 0,0128. **6.7.** Naudojame Maknemaros kriterijų. Rinkėjų, kurių nuomonė pakito iš teigiamos į neigiamą, skaičius yra 200, o iš neigiamos į teigiamą – 150. Tikrinama hipotezė $H : p = 1/2$ apie binominio skirstinio tikimybę, kai Bernulio eksperimentų skaičius yra $U_{01} + U_{10} = 350$, o sėkmių skaičius yra $U_{10} = 150$. Gauname $pv = 2\min(\mathbf{P}\{U_{10 \leq 150}\}, \mathbf{P}\{U_{10} \geq 150\}) = 0,0087$. Hipotezė atmestina. **6.8.** Duomenų pakanka, kad būtų galima apskaičiuoti Kochreno statistikos reikšmę. Gauname $Q = 14,8485$ ir $pv_a = \mathbf{P}\{\chi^2_2 > 14,8485\} = 0,0006$. Hipotezė atmetama. **6.11.** Gauname statistikų realizacijas: $\bar{g}_1 = 2,1598$, $\bar{g}_2 = 0,1524$, $\bar{g}_3 = -1,0849$ ir jas atitinkančias asimptotines $P$ reikšmes 0,0308, 0,8788, 0,2779. Remdamiesi empiriniu asimetrijos koeficientu hipotezę atmetame, jei kriterijaus reikšmingumo lygmuo viršija 0,0308. Atlikę Sarkadi transformaciją gauname paprastąją a. d. $Z \sim U(0, 1)$ imtį $Z_1, ..., Z_{n-2}$. Taikydami Pirsono chi kvadrato kriterijų hipotezei $H : Z \sim U(0, 1)$ tikrinti ir parinkę $k = 8$ vienodų tikimybių intervalus, gauname $X_n^2 = 5,5102$ ir $pv_a = 0,5980$. Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo kriterijų statistikos įgijo reikšmes 0,1071, 0,2629, 1,6185. Normalumo hipotezė neatmetama. Šapiro ir Vilkso statistika įgijo reikšmę 0,9716 ir ją atitinkanti $P$ reikšmė yra 0,0293. Šapiro ir Vilkso kriterijus atmeta normalumo hipotezę, jei kriterijaus reikšmingumo lygmuo viršija 0,0293. **6.12.** Perėję prie logaritmų gauname statistikų realizacijas: $\bar{g}_1 = 0,5901$, $\bar{g}_2 = 0,8785$, $\bar{g}_3 = 0,3765$ ir jas atitinkančias asimptotines $P$ reikšmes 0,5551, 0,3797, 0,7065. Kriterijai, grindžiami empirinių momentų funkcijomis, lognormalumo hipotezės neatmeta. Atlikę Sarkadi transformaciją gauname paprastąją a. d. $Z \sim U(0, 1)$ imtį $Z_1, ..., Z_{n-2}$. Taikydami Pirsono chi kvadrato kriterijų hipotezei $H : Z \sim U(0, 1)$ tikrinti ir parinkę $k = 10$ vienodų tikimybių intervalų, gauname $X_n^2 = 6,1892$ ir $pv_a = 0,7208$. Kolmogorovo ir Smirnovo, Kramero ir Mizeso ir Anderseno ir Darlingo kriterijų statistikos įgijo reikšmes 0,0548, 0,0752, 0,5594. Atitinkamos P reikšmės viršija 0,25. Lognormalumo hipotezė neatmetama. Šapiro ir Vilkso statistika įgijo reikšmę 0,9920 ir ją atitinkanti $P$ reikšmė yra 0,5644. Šapiro ir Vilkso kriterijus lognormalumo hipotezės taip pat neatmeta. **6.13.** a) Tuščių dėžių skaičius $Z_0^{(i)}, i = 1, 2, 3, 4$, yra 280, 593, 639, 359. Apskaičiavę $\mathbf{E}(Z_0^{(i)}), \boldsymbol{V}(Z_0^{(i)})$ randame $\bar{Z}_0^{(i)}$ realizacijas: 0,399; 0,937; -0947; -0,826 ir jas atitinkančias asimptotines $P$ reikšmes 0,345; 0,174; 0,828; 0,796. Atmesti hipotezes nėra pagrindo. b) Tuščių dėžių skaičius $Z_0$ jungtinėje imtyje 1871. Randame $\bar{Z}_0 = -0,107$ ir $pv_a = 0,543$. Hipotezė neatmetama.

# 7 skyrius

# A Priedas

## 7.1. DT įvertinių savybės

Tegu $\mathbf{X} = (X_1, \ldots, X_n)^T$, yra paprastoji imtis, t. y. $\mathbf{X}_1, \ldots, \mathbf{X}_n$ yra vienodai pasiskirstę n. a. d. Tarkime, kad $\mathbf{X}_i \sim p(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbf{R}^m$; čia $p(x, \boldsymbol{\theta})$ yra a. d. $X_i$ tankio funkcija $\sigma$ baigtinio mato $\mu$ atžvilgiu. Tada a. v. $\mathbf{X}$ tankio funkcija yra $f(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^n p(x_i, \boldsymbol{\theta})$, $\boldsymbol{x} \subset \mathbf{R}^n$.

Tikėtinumo funkcija ir jos logaritmas

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n p(X_i, \boldsymbol{\theta}), \quad \ell(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}).$$

Fišerio informacijos matrica

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}}(\dot{\ell}(\boldsymbol{\theta})\dot{\ell}^T(\boldsymbol{\theta})) = -\mathbf{E}_{\boldsymbol{\theta}}\ddot{\ell}(\boldsymbol{\theta}) = n\boldsymbol{i}(\boldsymbol{\theta}),$$

$$\boldsymbol{i}(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}}\ddot{\ell}_1(\boldsymbol{\theta}), \qquad \ell_1(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\theta}, \mathbf{X}_1) = \ln p(X_1, \boldsymbol{\theta}).$$

Atsitiktinio dydžio $X_1$ indukuotą tikimybinį matą žymėsime $\mathbf{P}_{\boldsymbol{\theta}}$, t. y. $\mathbf{P}_{\boldsymbol{\theta}}\{X_1 \subset \boldsymbol{B}\} = \int_{\boldsymbol{B}} p(x, \boldsymbol{\theta}) d\mu(x)$ su bet kuria Borelio aibe $\boldsymbol{B}$.

Minėjome, kad nagrinėjame tik *identifikuojamus* modelius: jei $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, tai $\mathbf{P}_{\boldsymbol{\theta}_1} \neq \mathbf{P}_{\boldsymbol{\theta}_2}$, t. y. egzistuoja tokia Borelio aibė $\boldsymbol{A}$, kad $\mathbf{P}_{\boldsymbol{\theta}_1}(X_1 \in \boldsymbol{A}) \neq \mathbf{P}_{\boldsymbol{\theta}_2}(X_1 \in \boldsymbol{A})$.

Primename, kad kvadratinės matricos $A = (a_{ij})_{n \times n}$ norma vadiname skaičių $|| \ A \ || = (\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} a_{ij}^2)^{1/2}$. Taigi tokių matricų sumos norma $|| \ A_1 + \ldots A_n \ || \leq || \ A_1 \ || + \cdots + || \ A_n \ ||$.

**Sąlygos A:**

1) aibė $\boldsymbol{\Theta}$ atvira;

2) beveik su visais $y \in \mathbf{R}$ parametro $\boldsymbol{\theta}$ tikrosios reikšmės $\boldsymbol{\theta}_0$ aplinkoje $V_\rho = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \leq \rho\}$ egzistuoja tolydžios išvestinės

$$\dot{p}(y, \boldsymbol{\theta}) = (\frac{\partial}{\partial\theta_1} p(y, \boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_m} p(y, \boldsymbol{\theta}))^T,$$

$$\ddot{p}(y, \boldsymbol{\theta}) = \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} p(y, \boldsymbol{\theta})\right]_{m \times m};$$

3) aplinkoje $V_\rho$ galima du kartus diferencijuoti po integralo ženklu, t. y.

$$\int_{\mathbf{R}^r} \dot{p}(y, \boldsymbol{\theta})\mu(dy) = \frac{\partial}{\partial\boldsymbol{\theta}} \int_{\mathbf{R}^r} p(y, \boldsymbol{\theta})\mu(dy) = \mathbf{0},$$

$$\int_{\mathbf{R}^r} \ddot{p}(y, \boldsymbol{\theta})\mu(dy) = \frac{\partial}{\partial\boldsymbol{\theta}} \int_{\mathbf{R}^r} \dot{p}(y, \boldsymbol{\theta})\mu(dy) = \mathbf{0};$$

4) Fišerio informacinė matrica $\boldsymbol{i}(\boldsymbol{\theta})$ teigiamai apibrėžta;

5) egzistuoja tokios neneigiamos funkcijos $h$ ir $b$, kad beveik su visais $y \in \mathbf{R}$ ir visais $\boldsymbol{\theta} \in V_\rho$

$$|| \ddot{\ell}_1(y, \boldsymbol{\theta}) - \ddot{\ell}_1(y, \boldsymbol{\theta}_0) || \leq h(y)\, b(\boldsymbol{\theta}), \quad \mathbf{E}_{\boldsymbol{\theta}_0}\{h(\mathbf{X}_1)\} < \infty, \quad b(\boldsymbol{\theta}_0) = 0,$$

o funkcija $b$ tolydi taške $\boldsymbol{\theta}_0$.

**7.1.1 teorema.** *Kai išpildytos sąlygos A, tai egzistuoja tokia a. d. seka* $\{\hat{\boldsymbol{\theta}}_n\}$, *kad*

$$\mathbf{P}(\dot{\ell}(\hat{\boldsymbol{\theta}}_n) = 0) \to 1, \quad \hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0, \tag{7.1.1}$$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0) \frac{1}{\sqrt{n}} \dot{\ell}(\boldsymbol{\theta}_0) + o_P(1), \tag{7.1.2}$$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N_m(0, \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0)), \tag{7.1.3}$$

$$\frac{1}{\sqrt{n}} \dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} N_m(0, \boldsymbol{i}(\boldsymbol{\theta}_0)), \tag{7.1.4}$$

$$-\frac{1}{n} \ddot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{P} \boldsymbol{i}(\boldsymbol{\theta}_0), \quad -\frac{1}{n} \ddot{\ell}(\hat{\boldsymbol{\theta}}) \xrightarrow{P} \boldsymbol{i}(\boldsymbol{\theta}_0). \tag{7.1.5}$$

**7.1.1 pastaba.** Jei tenkinamos teoremos sąlygos, tai

$$-(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T \ddot{\ell}(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_m, \tag{7.1.6}$$

$$\dot{\ell}^T(\boldsymbol{\theta}_0) \boldsymbol{i}^{-1}(\boldsymbol{\theta}_0) \dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_m, \tag{7.1.7}$$

$$-\dot{\ell}^T(\boldsymbol{\theta}_0) \ddot{\ell}^{-1}(\boldsymbol{\theta}_0) \dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_m, \tag{7.1.8}$$

$$-\dot{\ell}^T(\boldsymbol{\theta}_0) \ddot{\ell}^{-1}(\hat{\boldsymbol{\theta}}_n) \dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_m. \tag{7.1.9}$$

**7.1.2 pastaba.** Jei tenkinamos teoremos sąlygos, tai

$$-2 \ln \frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}}_n)} \xrightarrow{d} \chi^2_m. \tag{7.1.10}$$

**7.1.3 pastaba.** Tegu

$$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta} = \varphi(\boldsymbol{\gamma}), \boldsymbol{\gamma} \in \boldsymbol{G}\}, \quad \boldsymbol{G} \subset \mathbf{R}^{m-k}, \quad k < m,$$

čia $\varphi : |GG \to \boldsymbol{\Theta}_0$ yra tolydžiai diferencijuojamas atvaizdis. Jei tenkinamos teoremos sąlygos ir $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0$, tai

$$R = -2 \ln \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})} \xrightarrow{d} \chi^2_k, \quad n \to \infty. \tag{7.1.11}$$

# 8 skyrius

# B Priedas

## 8.1. Atsitiktinio proceso sąvoka

Baigtinis atsitiktinių dydžių, apibrėžtų toje pačioje tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$, rinkinys $\mathbf{X} = (X_1, \ldots, X_n)^T$ vadinamas atsitiktiniu vektoriumi. Jis indukuoja tikimybinį matą $\mathbf{P}_{\boldsymbol{X}}$ mačiojoje erdvėje $(\mathbf{R}^k, \mathcal{B}^k)$:

$$\mathbf{P}_{\boldsymbol{X}}(A) = \mathbf{P}\{\omega : \mathbf{X}(\omega) \in A\}, \quad A \in \mathcal{B}^k;$$

čia $\mathcal{B}^k$ yra erdvės $\mathbf{R}^k$ Borelio aibių $\sigma$ algebra.

Sąvoka "atsitiktinis procesas" apibendrina sąvoką "atsitiktinis vektorius" tuo atveju, kai atsitiktinių dydžių, apibrėžtų toje pačioje tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$, skaičius gali būti begalinis (netgi nebūtinai skaitus).

Tegu $\mathcal{T}$ yra realiųjų skaičių tiesės poaibis (baigtinis, skaitus, intervalas, visa tiesė).

**8.1.1 apibrėžimas.** Toje pačioje tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$ apibrėžta atsitiktinių dydžių sistema $\{X(t,\omega), t \in T, \omega \in \Omega\}$ vadinama *atsitiktiniu procesu.*

Atskiru $k$-mačio atsitiktinio vektoriaus atveju turime, kad $\mathcal{T}$ yra aibė $\{1, 2, \ldots, k\}$.

Fiksavus elementarųjį įvykį $\omega \in \Omega$ gaunama apibrėžta aibėje $\mathcal{T}$ neatsitiktinė funkcija $x(t) = X(t,\omega)$. Ši funkcija vadinama atsitiktinio proceso *trajektorija* arba *realizacija.*

Žymėsime $D = \{X(\cdot,\omega), \omega \in \Omega\}$ visų trajektorijų erdvę. Atsitiktinį procesą galima traktuoti kaip atsitiktinę funkciją, įgyjančią reikšmes trajektorijų erdvėje. Dar reikia apibrėžti tikimybinį matą, t. y. atsitiktinio proceso patekimo į aibes, priklausančias trajektorijų erdvės $\sigma$ algebrai, tikimybes.

Tegu $\rho(x,y)$ yra atstumas tarp dviejų erdvės $D$ funkcijų $x$ ir $y$. Šioje knygoje atstumu imamas skirtumo modulio supremumas:

$$\rho(x,y) = \sup_{t \in \mathcal{T}} \mid x(t) - y(t) \mid . \tag{8.1.1}$$

Jeigu aibė $G \subset D$ yra atvira, tai su kiekvienu $x \in G$ egzistuoja jo aplinka $B_\varepsilon(x) = \{y : \rho(x,y) < \varepsilon\} \subset G$.

**8.1.2 apibrėžimas.** *Mažiausioji $\sigma$ algebra, kuriai priklauso atviri aibės $D$ poaibiai, vadinama trajektorijų erdvės $D$ Borelio aibių $\sigma$ algebra; žymėsime $\mathcal{B}(D)$.*

**8.1.3 apibrėžimas.** Atsitiktinio proceso $\{X(t), t \in T\}$ tikimybiniu skirstiniu vadiname tikimybinį matą, apibrėžtą mačiojoje erdvėje $(D, \mathcal{B}(D))$ visoms aibėms $A \in \mathcal{B}(\mathcal{X})$:

$$\mathbf{P}^X(A) = \mathbf{P}\{X \in A\} = \mathbf{P}\{\omega : X(t,\omega) \in A\}.$$

# 8.2. Atsitiktinių procesų pavyzdžiai

## 8.2.1. Empirinis procesas

Tegu $\mathbf{X} = (X_1, \ldots, X_n)^T$ yra paprastoji imtis a.d. $X$, kurio pasiskirstymo funkcija $F(t) = \mathbf{P}\{X \leq t\}$, ir

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i)$$

yra empirinė pasiskirstymo funkcija.

**8.2.1 apibrėžimas.** *Atsitiktinis procesas*

$$\mathcal{E}_n(t) = \sqrt{n}(\hat{F}_n(t) - F(t)),\ t \in T = \mathbf{R} \tag{8.2.1}$$

*yra vadinamas empiriniu procesu.*

Jeigu $F(t)$ yra absoliučiai tolydžiojo a.d. pasiskirstymo funkcija, tai pakanka nagrinėti atsitiktinį procesą

$$\mathcal{E}_n^*(y) = \sqrt{n}(\hat{G}_n(y) - y),\ y \in [0,\, 1]; \tag{8.2.2}$$

čia $\hat{G}_n(y)$ yra atsitiktinio dydžio $Y = F(X) \sim U(0,\, 1)$ didumo $n$ paprastosios imties empirinė pasiskirstymo funkcija. Atsitiktinio proceso $\mathcal{E}_n^*(t)$ trajektorijos įgyja reikšmes iš intervalo $[-\sqrt{n}, \sqrt{n}]$, kai $y \in [0,\, 1]$. Trajektorijos yra tolydžios iš dešinės funkcijos, kintančios didumo $1/\sqrt{n}$ šuoliukais. Remiantis 4.3.1 teorema

$$\mathbf{E}(\mathcal{E}_n^*(y)) = 0, \quad \mathbf{Cov}(\mathcal{E}_n^*(y), \mathcal{E}_n^*(z)) = y(1-z), \quad 0 \leq y \leq z \leq 1 \tag{8.2.3}$$

## 8.2.2. Vinerio procesas (Brauno judesys)

**8.2.2 apibrėžimas.** Atsitiktinis procesas vadinamas *Gauso procesu*, jei visi jo baigtiniamačiai skirstiniai yra normalieji.

**8.2.3 apibrėžimas.** Atsitiktinis procesas $W(t), t \in T = [0, \infty)$ vadinamas *Vinerio procesu (Brauno judesiu)*, jei jis tenkina tokias sąlygas:

a) $W(0) = 0$; b) $W(t) - W(s) \sim N(0, t-s)$ su visais $0 \leq s < t < \infty$;

c) $W$ turi ***nepriklausomus pokyčius***, t.y. su visais $0 = t_0 < t_1 < \cdots < t_k$ atsitiktiniai dydžiai $W(t_{j+1}) - W(t_j)$, $j = 1, 2...., k-1$, yra nepriklausomi.

Sąlygos a)-c) vienareikšmiškai nusako proceso baigtiniamačius skirstinius:

$$(W(t_1), \ldots, W(t_n))^T \sim N_n(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = [\sigma_{ij}]_{n\times n}, \quad \sigma_{ij} = t_i \wedge t_j.$$

Skirstiniai yra suderinti, todėl baigtiniamačiai skirstiniai vienareikšmiškai nusako Vinerio proceso tikimybinį skirstinį.

## 8.2.3. Brauno tiltas

**8.2.4 apibrėžimas.** Atsitiktinis procesas

$$B(t) = W(t) - tW(1), \quad t \in [0,\, 1] \tag{8.2.4}$$

vadinamas intervalo $[0,\, 1]$ *Brauno tiltu.*

Baigtiniamačiai skirstiniai yra normalieji: su kiekvienu natūriniu $n$ ir realiais $0 \leq t_1 < \cdots < t_n \leq 1$

$$(B(t_1), \ldots, B(t_n))^T \sim N_n(\mathbf{0}, \Gamma), \quad \Gamma = ||\gamma_{ij}||_{n\times n}, \quad \gamma_{ij} = t_i(1-t_j), \quad 0 \leq t_i \leq t_j \leq 1. \tag{8.2.5}$$

Atkreipsime dėmesį, kad Brauno tilto ir empirinio proceso $\mathcal{E}_n^*(t)$ baigtiniamačių skirstinių vidurkiai ir kovariacinės matricos sutampa.

Brauno judesys ir Brauno tiltas yra Gauso procesai.

## 8.3. Atsitiktinių procesų silpnas konvergavimas

Tarkime, turime atsitiktinių procesų seką $\{X^{(n)}\}$ ir atsitiktinį procesą $X$, apibrėžtus toje pačioje tikimybinėje erdvėje $(\Omega, \mathcal{F}, \mathbf{P})$. Šių procesų tikimybinius skirstinius mačiojoje erdvėje $(D, \mathcal{B}(D))$ žymėsime $\mathbf{P}^{X^{(n)}}$ ir $\mathbf{P}^X$.

Žymėsime $\partial A$ aibės $A \in \mathcal{B}$ kraštą.

**8.3.1 apibrėžimas**. Atsitiktinių procesų seka $\{X^{(n)}\}$ silpnai konverguoja į atsitiktinį procesą $X$, jei su bet kuria $A \in \mathcal{B}^s(D)$, tokia, kad $\mathbf{P}^X(\partial A) = 0$, gauname:

$$\mathbf{P}^{X^{(n)}}(A) \to \mathbf{P}^X(B), \quad n \to \infty.$$

Kaip ir atsitiktinių dydžių ar vektorių silpnas konvergavimas žymimas $X^{(n)} \xrightarrow{d} X$.

Iš silpno konvergavimo išplaukia, kad su visais $x_1, \ldots, x_m$

$$(X^{(n)}(x_1), \ldots, X^{(n)}(x_m)) \xrightarrow{d} (X(x_1), \ldots, X(x_m)),$$

bet nebūtinai atvirkščiai.

## 8.4. Empirinio proceso silpnas invariantiškumas

Tarkime, $\mathbf{X} = (X_1, \ldots, X_n)^T$ yra paprastoji imtis a. d. $X$ su pasiskirstymo funkcija $F$, o $\hat{F}_n$ yra empirinė pasiskirstymo funkcija.

Remdamiesi CRT atsitiktinių vektorių sumoms gauname, kad empirinio proceso

$$\mathcal{E}_n(x) = \sqrt{n}(\hat{F}_n(x) - F(x))$$

baigtiniamačiai vektoriai $(\mathcal{E}_n(x_1), \ldots, \mathcal{E}_n(x_m))^T$ silpnai konverguoja į atsitiktinį vektorių

$$(Z_1, \ldots, Z_m)^T \sim N_m(\mathbf{0}, \Gamma), \quad \Gamma = [\gamma_{ij}]_{m \times m},$$

$$\gamma_{ij} = F(x_i)(1 - F(x_j)), \quad i \le j = 1, \ldots, m,$$

su visais $m$ ir bet kokiais rinkiniais $-\infty < x_1 < \cdots < x_m < \infty$.

Iš šio rezultato išeina, kad empirinio proceso baigtiniamačiai skirstiniai silpnai konverguoja į atsitiktinio proceso $B(F(x))$ baigtiniamačius skirstinius; čia $B$ yra Brauno tiltas.

Apskritai, jei $h$ yra tolydi funkcija, tai

$$(h(\mathcal{E}_n(x_1)), \ldots, h(\mathcal{E}_n(x_m))) \xrightarrow{d} (h(B(F(x_1))), \ldots, h(B(F(x_m)))). \tag{8.4.1}$$

Teisinga ir dar bendresnė teorema.

**8.4.1 teorema.** *(empirinio proceso silpno invariantiškumo principas). Jei $F$ yra absoliučiai tolydi pasiskirstymo funkcija ir $h$ tolydus funkcionalas tolydžių iš dešinės ir turinčių baigtines ribas iš kairės apibrėžtų intervale* $[0, 1]$ *funkcijų klasėje, tai*

$$h(\mathcal{E}_n^*) \xrightarrow{d} h(B);$$

*čia*

$$\mathcal{E}_n^*(y) = \sqrt{n}(\hat{G}_n(y) - y), \quad \hat{G}_n(y) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(-\infty, y]}(F(X_i)),$$

*o $B$ yra intervalo* $[0, 1]$ *Brauno tiltas.*

▲

Silpno invariantiškumo savybė labai naudinga ieškant įvairių empirinio proceso funkcionalų ribinių skirstinių. Pavyzdžiui, galima tvirtinti, kad Kolmogorovo ir Smirnovo, Kramero ir Mizeso, Anderseno ir Darlingo statistikos turi tokias ribas:

$$\sqrt{n}D_n \xrightarrow{d} \sup_{0 \le t \le 1} |B(t), \quad nC_n \xrightarrow{d} \int_0^t B^2(t)dt, \quad nA_n \xrightarrow{d} \int_0^t \frac{B^2(t)}{t(1-t)}dt.$$

Vadinasi, tereikia rasti Brauno tilto funkcionalų tikimybinius skirstinius.

## 8.5. Brauno judesio ir Brauno tilto savybės

**1 savybė.** Jei $\tau = \inf\{t : W(t) \in A\}$, $A$ – realių skaičių tiesės Borelio aibė, o $W(t)$ – Brauno judesys, tai $\tilde{W}(t) = W(t+\tau) - W(\tau)$ taip pat yra Brauno judesys.

**Įrodymas.** Gauname: $\tilde{W}(0) = 0$ ir su visais $0 = t_0 < t_1 < \cdots < t_k$, $x_1, \ldots, x_k$

$$\mathbf{P}\{\tilde{W}(t_1) - \tilde{W}(t_0) \leq x_1, \ldots, \tilde{W}(t_k) - \tilde{W}(t_{k-1}) \leq x_k\} =$$

$$= \int_0^\infty \mathbf{P}\{W(t_1+u) - W(u) \leq x_1, \ldots, W(t_k+u) - W(t_{k-1}+u) \leq x_k | \tau = u\} dF_\tau(u) =$$

$$= \mathbf{P}\{W(t_1+u) - W(u) \leq x_1\} \ldots \mathbf{P}\{W(t_k+u) - W(t_{k-1}+u) \leq x_k\},$$

nes įvykis $\{\tau = u\}$ nusakomas atsitiktiniu procesu $W(s), s \leq u$, taigi sąlyginė tikimybė po integralo ženklu sutampa su besąlygine. ▲

**2 savybė.**(Atspindžio taisyklė). Su visais $x, y \in \mathbf{R}, t \geq t_0 \geq 0$

$$\mathbf{P}\{W(t) > x + y | W(t_0) = x\} = \mathbf{P}\{W(t) < x - y | W(t_0) = x\}.$$

**Įrodymas.** Gauname

$$\mathbf{P}\{W(t) > x+y | W(t_0) = x\} = \mathbf{P}\{W(t_0) + (W(t) - W(t_0)) > x + y | W(t_0) = x\} =$$

$$\mathbf{P}\{W(t) - W(t_0) > y\} = 1 - \Phi(y/\sqrt{t - t_0}).$$

Analogiškai

$$\mathbf{P}\{W(t) < x - y | W(t_0) = x\} = \mathbf{P}\{W(t) - W(t_0) < -y\} = 1 - \Phi(y/\sqrt{t-t_0}).$$

▲

**3 savybė.** Su visais $x \in \mathbf{R}$

$$P(x) = \mathbf{P}\{\exists t \in [0,1] : B(t) = x\} = e^{-2x^2}.$$

**Įrodymas.** $P(0) = 1$, $P(-x) = P(x)$, nes $B$ ir $-B$ turi tuos pačius skirstinius. Todėl pakanka nagrinėti atvejį $x > 0$.

Apibrėžkime atsitiktinį procesą $W(t) = B(t) + tW_1$; čia $W_1 \sim N(0,1)$ – nepriklausantis nuo $B$ atsitiktinis dydis. Tada $W(t), t \in [0,1]$, yra Brauno judesys, nes tai Gauso procesas,

$$\mathbf{E}W(t) = 0, \quad \mathbf{Cov}(W(s), W(t)) = s \wedge t.$$

Kadangi $B(1) = 0$, tai $W(1) = W_1$.

Nagrinėkime sąlyginę tikimybę

$$P(x, \varepsilon) = \mathbf{P}\{\exists t \in [0.1] : W(t) \geq x | \, |W(1)| < \varepsilon\}, \quad 0 < \varepsilon < x.$$

Pažymėkime $\tau = \inf\{t : W(t) > x\}$. Kadangi $W$ trajektorijos tolydžios, $W(\tau) = x$. Pagal 1) savybę

$$\tilde{W}(u) = W(\tau + u) - W(\tau)$$

yra Brauno judesys. Jei $\tau < 1$, tai

$$W(1) = W(\tau + (1-\tau)) = \tilde{W}(1-\tau) + x.$$

Todėl pasinaudoję tuo, kad $\tilde{W}$ ir $-\tilde{W}$ turi vienodus skirstinius, gauname

$$P(x,\varepsilon) = \mathbf{P}\{\tau < 1 | |W(1)| < \varepsilon\} = \mathbf{P}\{\tau < 1, |\tilde{W}(1-\tau) + x| < \varepsilon\} / \mathbf{P}\{|W(1)| < \varepsilon\} =$$

$$\mathbf{P}\{\tau < 1, |\tilde{W}(1-\tau) - x| < \varepsilon\} / \mathbf{P}\{|W(1)| < \varepsilon\} = \mathbf{P}\{\tau < 1, |W(1) - 2x| < \varepsilon\} / \mathbf{P}\{|W(1)| < \varepsilon\} =$$

$$\mathbf{P}\{|W(1) - 2x| < \varepsilon\} / \mathbf{P}\{|W(1)| < \varepsilon\} = \frac{\Phi(2x+\varepsilon) - \Phi(2x-\varepsilon)}{\Phi(\varepsilon) - \Phi(-\varepsilon)}.$$

Taigi

$$\lim_{\varepsilon \downarrow 0} P(x, \varepsilon) = e^{-2x^2}.$$

Kadangi $B$ ir $W(1)$ nepriklausomi, tai gauname

$$P(x) = \mathbf{P}\{\exists t \in [0,\,1] : B(t) = x||W(1)| < \varepsilon\} \le$$

$$\mathbf{P}\{\exists t \in [0,\,1] : B(t) \ge x - \varepsilon - tW(1)|\,|W(1)| < \varepsilon\} = P(x-\varepsilon, \varepsilon).$$

Analogiškai $P(x) \ge P(x+\varepsilon, \varepsilon)$.

Fiksuokime $\delta > \varepsilon > 0$. Kai $\varepsilon \downarrow 0$ tai

$$P(x) \le P(x-\varepsilon, \varepsilon) \le P(x-\delta, \varepsilon) \to e^{-2(x-\delta)^2},$$

$$P(x) \ge P(x+\varepsilon, \varepsilon) \ge P(x+\delta, \varepsilon) \to e^{-2(x+\delta)^2},$$

todėl

$$e^{-2(x+\delta)^2} \le P(x) \le e^{-2(x-\delta)^2}.$$

Perėję prie ribos, kai $\delta \downarrow 0$, gausime $P(x) = e^{-2x^2}$. ▲

**4 savybė.** (Kolmogorovo ir Smirnovo statistikos ribinis skirstinys). Su visais $x > 0$

$$\mathbf{P}\{\sup_{0\le t\le 1} |B_t| \ge x\} = 2\sum_{n=1}^{\infty} (-1)^{n-1} e^{-2n^2x^2}. \tag{8.5.1}$$

**Įrodymas.** Nagrinėkime įvykį

$$A_n(x) = \{\exists\, 0 \le t_1 < \cdots < t_n \le 1 : B(t_j) = (-1)^{j-1}x, j = 1, \ldots, n\}$$

ir atsitiktinį dydžius $\tau = \inf\{t : B(t) = x\}$, $\tau' = \inf\{t : B(t) = -x\}$. Pažymėkime

$$P_n(x) = \mathbf{P}\{A_n(x)\}, \quad Q_n(x) = \mathbf{P}\{A_n(x), \tau < \tau'\}.$$

Gauname

$$Q_n(x) + Q_{n+1}(x) = \mathbf{P}\{A_n(x), \tau < \tau'\} + \mathbf{P}\{A_{n+1}(x), \tau < \tau'\} =$$
$$= \mathbf{P}\{A_n(x), \tau < \tau'\} + \mathbf{P}\{A_n(x), \tau' < \tau\} = P_n(x).$$

Pagal 3) savybę $P_1(x) = e^{-2x^2}$.

Rasime $P_2(x)$. Pagal pilnosios tikimybės formulę

$$\mathbf{P}\{\exists\, 0 < t_1 < t_2 \le 1 : W(t_1) = x, W(t_2) = -x, |W(1)| < \varepsilon\} =$$

$$= \mathbf{P}\{\exists\, 0 \le t_1 < t_2 \le 1 : B(t_1) = x - t_1 W_1, B(t_2) = -x - t_2 W_1, |W(1)| < \varepsilon\} =$$

$$= \int_{-\varepsilon}^{\varepsilon} \mathbf{P}\{\exists\, 0 < t_1 < t_2 \le 1 : B(t_1) = x - t_1 v, B(t_2) = -x - t_2 v\}\varphi(v)dv;$$

čia $\varphi(v)$ – standartinio normaliojo skirstinio tankis. Todėl

$$\lim_{\varepsilon\downarrow 0} \frac{1}{2\varepsilon}\mathbf{P}\{\exists\, 0 < t_1 < t_2 \le 1 : W(t_1) = x, W(t_2) = -x||W(1)| < \varepsilon\} =$$

$$P_2(x) \lim_{\varepsilon\downarrow 0} \frac{\Phi(\varepsilon) - \Phi(-\varepsilon)}{2\varepsilon} = P_2(x)\varphi(0).$$

Kairiąją lygybės pusę pertvarkome naudodami atspindžio taisyklę:

$$\mathbf{P}\{\exists\, 0 < t_1 < t_2 \le 1 : W(t_1) = x, W(t_2) = -x, |W(1)| < \varepsilon\} =$$

$$\mathbf{P}\{\exists\, 0 < t_1 < t_2 < 1 : W(t_1) = x, W(t_2) = 3x, |W(1) - 4x| < \varepsilon\} =$$
$$\mathbf{P}\{|W(1) - 4x| < \varepsilon\} = \Phi(4x + \varepsilon) - \Phi(4x - \varepsilon).$$

Taigi gauname

$$P_2(x) = \lim_{\varepsilon\downarrow 0} \frac{\Phi(4x+\varepsilon) - \Phi(4x-\varepsilon)}{2\varepsilon\varphi(0)} = e^{-8x^2}.$$

Analogiškai, $P_n(x) = e^{-2n^2x^2}$.

Gauname

$$Q_1 = P_1 - Q_2 = P_1 - P_2 + Q_3 = \sum_{k=1}^{n-1} (-1)^k P_k + (-1)^n Q_n.$$

Kadangi $Q_n \leq P_n \to 0$, tai

$$Q_1(x) = \sum_{n=1}^{\infty} (-1)^n e^{-2n^2x^2}.$$

Pagaliau

$$\mathbf{P}\{\sup_{0 \leq t \leq 1} |B_t| \geq x\} = 2Q_1(x) = 2\sum_{n=1}^{\infty} (-1)^n e^{-2n^2x^2}.$$

▲

# Literatūra

**1. Anderson T.W.** On the distribution of the two-sample Cramer-von-Mises criterion. *Ann. Math. Statist.*, vol. **33**, 1962.

**2. Bagdonavičius V., Kruopis J., Nikulin M.** Nonparametric Tests for Complete Data. ISTE: London, 2011.

**3. Bagdonavičius V., Kruopis J., Nikulin M.** *Nonparametric Tests for Censored Data*, ISTE: London, 2011.

**4. Barton D. E.** On Neyman's cmooth test of goodness of fit and its power with respect to a particular system of alternatives. Skand. Aktuartidskr. 36,, 24, 1953.

**5. Bolshev L.N.** On charakterization of the Poisson distribution and its statistical applications. *Theory of Probability and its Applications*, vol. **10**, 1965.

**6. Bolshev L.N., Mirvaliev M.** Chi-square goodness-of-fit tests for the Poisson, binomial and negative binomial distributions. *Theory of Probability and its Applications*, vol. **23**, 1978.

**7. Bolshev L.N., Smirnov N. N.** Tables of Mathematical Statistics. Nauka: Moskow, 1983.

**8. Corder G. W., Foreman D. I.** Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. Wiley: New Jersey, 2009.

**9. Cramer H.** Mathematical Methods of Statistics. Princeton University Press, 1946.

**10. D'Agostino R. B.** Transformation to normality of the null distribution of g1. *Biometrika*, vol. **57**, 1970.

**11. D'Agostino R. B., Stephens M. A.** Goodness-of-fit techniques. Marcel Dekker: New York, 1986.

**12. Geary R. C.** The ratio of the mean deviation to the standart deviation as a test of normality. *Biometrika*, vol. **27**, 1935.

**13. Gibbons J. D., Chakraborti S.** Nonparametric Statistical Inference. CRC Press, 5th edn., 2009.

**14. Govindarajulu Z.** Nonparametric Inference. World Scientific, 2007.

**15. Hollander M., Wolfe D. A.** Nonparametric Statistical Methods, 2nd edn. Wiley: New York, 1999.

**16. Kang S.I** Performance of Generalized Neyman Smooth Goodness of fit Tests. Disertacija, 1978.

**17. Kruopis J.** Matematinė statistika. Mokslo ir enciklopedijų leidykla: Vilnius, 1993.

**18. Lemeshko B. Yu.** Errors when using nonparametric fitting criteria. *Measurement Techniques*, vol. **47**, 2, 2004.

**19. Lemeshko B. Yu., Lemeshko S. B.** Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part I. *Measurement Techniques*, vol. **52**, 6, 2009.

**20. Lemeshko B. Yu., Lemeshko S. B.** Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II. *Measurement Techniques*, vol. **52**, 8, 2009.

**21. Mardia K. V.** Statistics of Directional Data. Academic Press Inc (London), 1972.

**22. Martynov G. V.** Omega-square Criteria. Nauka: Moskow, 1977.

**23. Neyman J.** Smooth test for goodness-of-fit. *Skand. Aktuartidskr*, vol. **20**, 1937.

**24. Nikulin M. S.** Chi-square test for normality. *Proceedings of the International Vilnius Conference on Probability Theory and Mathematical Statistics*, vol. **2**, 1973.

**25. Nikulin M. S.** Chi-square test for continuous distributions with shift and scale parameters. *Theory of Probability and its Applications*, vol. **18**, 1973.

**26. Pitman E. J. G.** Non-parametric Statistical Inference. University of North Carolina Institute of Statistics (lecture notes), 1948.

**27. Rutherford E., Chadwick J., Ellis C. D.** Radiation from Radioactive Substances. Cambridge University Press: London, 1930.

**28. Smirnov N. V.** On estimating the discrepancy between empirical distribution functions in two independent samples. *The Bulletin of the Moscow's Gos.University*, Ser. A, vol. **2**, 1939.

**29. Van der Vaerden B. L.** Order tests for the two-sample problem and their power. *Proc. Kon. Ned. Akad. Wetensch*, A, vol. **55**, 1952.

**30. Van der Vaart A. W.** Asymptotic Statistics. Cambridge University Press, 2000.

**31. Yates F.** Contingency table involving small numbers and the chi square test. *Supplement to the Journal of the Royal Statistical Society*, vol. **1**, 1934.

# Dalykinė rodyklė

**Vilijandas Bagdonavičius, Julius Jonas Kruopis**
Matematinė statistika: vadovėlis
Trečia dalis. Neparametrinė statistika. – Vilnius: Vilniaus universitetas, 2015. – 209 p.
ISBN 978–609–459–517–2

Trečia vadovėlio dalis skiriama hipotezių tikrinimo uždaviniams spręsti, kai statistinis modelis neparametrinis. Pateikiami kriterijai suderinamumo, nepriklausomumo, atsitiktinumo, homogeniškumo hipotezėms tikrinti. Nagrinėjamos trys kriterijų, kurių statistikos tiesiogiai nepriklauso nuo stebimų a. d. skirstinių, klasės. Chi kvadrato tipo kriterijų statistikos sudaromos naudojant grupuotas imtis, t. y. pereinant prie polinominio skirstinio. Antros klasės kriterijų statistikos sudaromos atliekant iš pradžių stebimų a. d. transformacijas, kurios suveda jų skirstinius į tolygiuosius. Trečios klasės kriterijų statistikos sudaromos naudojant rangus, t. y. priklauso tik nuo stebėjimų tarpusavio padėties, o ne nuo jų faktiškų reikšmių.

519.2(075.8)

Vlijandas Bagdonavičius, Julius Jonas Kruopis
Matematinė statistika. III dalis. Neparametrinė statistika
Vadovėlis

Lietuvių kalbos redaktorė *Danutė Petrauskienė*
Maketuotoja *Rūta Levulienė*

Išleido *Vilniaus universiteto leidykla*