IV skyrius. MATEMATINĖS STATISTIKOS PRADMENYS

1. MATEMATINĖS STATISTIKOS UŽDAVINIAI

Matematinė statistika nagrinėja stebėjimo rezultatų matematinio aprašymo ir analizavimo būdus. Tipiškas matematinės statistikos uždavinys gali būti nusakytas šitaip. Sakykime, turime aibę objektų, kuriuos tiriame pagal kuriuos nors požymius, o pačius požymius galime charakterizuoti skaičiais. Tų skaičių arba vektorių (jei tiriame keletą požymių) aibę įprasta vadinti *generaline aibe*. Paprastai tiriamojo požymio pasiskirstymas generalinėje aibėje nėra žinomas. Norint jį nustatyti, reikėtų ištirti visus generalinės aibės objektus. Tai gali pareikalauti daug darbo ir lėšų, o kartais toks tyrimas iš principo nėra galimas. Tarkime, kad automatinės staklės gamina kokias nors pigias detales. Mums rūpi tų detalių svorio pasiskirstymas visoje vienos dienos produkcijoje. Tektų sverti kiekvieną detalę. Šis darbas kainuotų daugiau negu pati detalių partija. Kitas pavyzdys. Fabrikas gamina elektros lemputes. Reikia sužinoti, kiek laiko vidutiniškai jos dega. Norint patikrinti visų per dieną pagamintų lempučių degimo trukmę, reikėtų jas žibinti tol, kol perdegs. Taip sugadintume visą produkciją. Todėl elgiamasi kitaip: atsitiktinai parenkama generalinės aibės objektų dalis, ištiriamas reikiamo požymio pasiskirstymas tame poaibyje ir iš jo sprendžiama apie to požymio pasiskirstymą visoje generalinėje aibėje. Parinktoji generalinės aibės dalis vadinama *imtimi*. Taikant matematinės statistikos metodus, galima įvertinti tiriamojo požymio pasiskirstymą generalinėje aibėje, kai žinomas to požymio pasiskirstymas imtyje.

Matematinės statistikos metodai tinka tik tada, kai imtis yra *reprezentatyvi*, t. y. kai ji teisingai atspindi tiriamojo požymio galimų reikšmių proporcijas generalinėje aibėje.

Imtį galima sudaryti įvairiai. Galima atsitiktinai imti kurį nors generalinės aibės objektą, po to grąžinti jį į generalinę aibę ir toliau vėl atsitiktinai imti bet kurį tos aibės objektą. Tačiau galima kiekvieno parinkto objekto į generalinę aibę nebegrąžinti. Parinkimai gali būti nepriklausomi, bet gali būti ir priklausomi.

Matematinėje statistikoje stengiamasi imtis aprašyti atsitiktinių dydžių terminais. Paprasčiausiu atveju tinka šitokia schema. Tiriame atsitiktinį dydį su nežinoma pasiskirstymo funkcija. Stebime tą dydį n kartų. Gauname n stebėjimo rezultatų – imtį. Iš jos reikia spręsti apie nežinomą pasiskirstymo funkciją. Toliau visur stebėjimus laikysime nepriklausomais. Sudarysime vienamačio atsitiktinio dydžio n nepriklausomų stebėjimų matematinį modelį.

Kaip žinome, atsitiktinis dydis yra mati funkcija, atvaizduojanti pirmykštę tikimybinę erdvę mačioje erdvėje $\{R, \mathcal{B}\}$. Tačiau tiriant to dydžio reikšmių pasiskirstymą, nebūtina žinoti pirmykštės tikimybinės erdvės – galima operuoti erdve $\{R, \mathcal{B}\}$ ir atsitiktinio dydžio indukuotu tikimybiniu pasiskirstymu toje erdvėje (žr. II.2 skyrelį).

Sakykime, stebime atsitiktinį dydį, indukuojantį tikimybinę erdvę $\{R, \mathcal{B}, P_\theta\}$. Apie to dydžio tikimybinį pasiskirtymą P_θ žinome tik tiek, kad jis priklauso pasiskirstymų klasei $\{P_\theta, \theta \in \Theta\}$; čia Θ gali būti realiųjų skaičių, arba erdvės R^s, $s > 1$, taškų, arba ir dar sudėtingesnė aibė. Sakykime, žinome, kad stebimasis atsitiktinis dydis yra pasiskirstęs pagal Puasono dėsnį su nežinomu parametru λ; tada aibe Θ galime laikyti visų realiųjų teigiamų skaičių aibę $(0, \infty)$. Jei atsitiktinis dydis pasiskirstęs pagal normalųjį dėsnį $N(a, \sigma^2)$ su nežinomu vidurkiu a ir nežinoma dispersija σ^2, tai aibė Θ gali būti erdvės R^2 aibė $R \times (0, \infty)$. Tačiau apie atsitiktinį dydį galime turėti ir mažiau informacijos. Sakysime, galime žinoti, tik kad jis yra diskretusis arba tolydusis. Tada Θ teks laikyti daug sudėtingesne aibe.

Atsitiktinio dydžio n stebėjimų bus n-matis atsitiktinis vektorius $X = (X_1, ..., X_n)$. Kadangi stebėjimai yra nepriklausomi, tai $X_1, ..., X_n$ bus nepriklausomi atsitiktiniai dydžiai. X pasiskirstymą nusakys jo indukuota tikimybinė erdvė

$$\{R^n, \mathcal{B}^n, \mathcal{P}_\theta\} = \{R^n, \mathcal{B}^n, P_\theta^n\} = \{R, \mathcal{B}, P_\theta\}^n$$

(žr. V.10 skyrelį). Atsitiktinis dydis X_k atitinka k-ąjį stebėjimą, o vektorius $X = (X_1, ..., X_n)$ yra *atsitiktinė*, arba *matematinė, imtis*. Konkreti to vektoriaus reikšmė $x = (x_1, ..., x_n)$ yra konkreti imtis arba atsitiktinės imties $X = (X_1, ..., X_n)$ *realizacija*. Erdvė $\{R, \mathcal{B}, P_\theta\}^n$ paprastai vadinama *imčių erdve*.

Kartais stebėjimų skaičius n yra labai didelis. Tada tikslinga kalbėti apie begalinę nepriklausomų stebėjimų seką. Jos matematinis modelis yra erdvė $\{R, \mathcal{B}, P_\theta\}^\infty$. Kiekvienam n erdvę $\{R, \mathcal{B}, P_\theta\}^n$ galima traktuoti kaip begalinės erdvių sandaugos poerdvį.

Didžioji dalis klausimų, kuriuos tenka spręsti matematinėje statistikoje, yra dviejų tipų.

1. Į v e r t i n i m ų t e o r i j a. Jos tikslas – nurodyti metodus, kuriais galima būtų įvertinti stebimojo atsitiktinio dydžio pasiskirstymo funkciją arba kitas pasiskirstymo charakteristikas: vidurkį, dispersiją ir pan. Dažnai, remiantis kokiais nors teoriniais samprotavimais ar praktine patirtimi, galima teigti, kad pasiskirstymo funkcija yra žinomo analizinio pavidalo, bet priklauso nuo vieno arba kelių nežinomų parametrų θ. Reikia tuos parametrus įvertinti.

Panagrinėsime pavyzdį. Metame monetą. Realios monetos nėra simetriškos. Iš stebėjimo rezultatų reikia įvertinti herbo atsivertimo tikimybę p. Su monetos metimu galime susieti atsitiktinį dydį, įgyjantį reikšmę 1, kai atsiverčia herbas, ir 0, kai atsiverčia kita monetos pusė. To dydžio pasiskirstymo

funkcija

$$(1-p)\varepsilon(y)+p\varepsilon(y-1)$$

priklauso nuo nežinomo parametro p; čia $\varepsilon(y)$ yra III.7.2 teoremos įrodyme nusakyta vienetinė pasiskirstymo funkcija.

Suprantama, iš stebėjimo rezultatų negalime tiksliai nusakyti nežinomų pasiskirstymo charakteristikų, galime tik apytiksliai jas įvertinti.

2. H i p o t e z i ų t i k r i n i m a s. Bet kokią prielaidą apie stebimojo atsitiktinio dydžio pasiskirstymo dėsnį vadiname *statistine hipoteze*. Reikia patikrinti, ar stebėjimo duomenys neprieštarauja tai prielaidai. Matematiškai tą uždavinį galime formuluoti šitaip. Tarkime, kad stebimojo atsitiktinio dydžio pasiskirstymas P_θ priklauso klasei $\{P_\theta, \theta\in\Theta\}$. Statistine hipoteze vadinsime prielaidą, kad $\theta\in\Theta_0$; čia Θ_0 yra aibės Θ tikrinis poaibis. Rašoma $H:\theta\in\Theta_0$. Hipotezė vadinama *paprastąja*, kai Θ_0 yra sudaryta tik iš vieno elemento, ir *sudėtingąja*, kai aibėje Θ_0 yra daugiau elementų. Tikrinamoji hipotezė, kad $\theta\in\Theta_0$, dar vadinama *nuline hipoteze* H_0, o hipotezė $H_1:\theta\in\Theta\backslash\Theta_0$ vadinama *alternatyviąja hipoteze*, arba tiesiog *alternatyva*.

Grįžkime prie jau minėto pavyzdžio apie monetą. Hipotezė $p=1/2$ (moneta simetriška) yra paprastoji; jos alternatyva yra $p\neq 1/2$. Hipotezė $p>1/2$ yra sudėtingoji.

Čia paminėjome tik porą pagrindinių matematinės statistikos uždavinių – su kitais susipažinsime kituose skyreliuose. Tačiau ir ten pateiksime tik mažą dalį uždavinių, kuriuos nagrinėja matematinė statistika. Šis skyrius yra tik trumpas įvadas į matematinės statistikos idėjas ir metodus.

2. ATSITIKTINIO DYDŽIO EMPIRINĖS CHARAKTERISTIKOS

Statistinėms išvadoms daryti iš stebėjimo rezultatų naudojamos įvairios imties funkcijos, iš tradicijos vadinamos statistikomis. Su jomis tenka atlikinėti algebrines bei analizines operacijas. Todėl natūralu reikalauti, kad tos funkcijos būtų mačios. Apibrėšime jas tiksliai.

Tarkime, kad, be erdvės $\{R,\mathcal{B},P_\theta\}^n$, tiriame dar ir mačią erdvę $\{\Gamma,\mathcal{E}\}$, ir funkcija $T:R^n\to\Gamma$ yra $(\mathcal{B}^n,\mathcal{E})$ mati, t. y. $T^{-1}(A)\in\mathcal{B}^n$ kiekvienai $A\in\mathcal{E}$. Tada imties funkcija $T(X)=T(X_1,...,X_n)$ yra vadinama *statistika*. Paprastai erdvė $\{\Gamma,\mathcal{E}\}$ yra $\{R^s,\mathcal{B}^s\}$, atskiru atveju $\{R,\mathcal{B}\}$. Todėl statistika $T(X)$ yra daugiamatis, arba atskiru atveju vienamatis, atsitiktinis dydis. Kai $x=(x_1,...,x_n)$ yra konkreti imtis, $T(x)=T(x_1,...,x_n)$ yra konkreti statistikos reikšmė, jos *realizacija*.

Viena iš pagrindinių statistikų yra vadinamoji *empirinė pasiskirstymo funkcija* $\mathcal{F}_n(y)$. Ji apibrėžiama šitaip:

$$\mathcal{F}_n(y)=\frac{1}{n}\sum_{X_k<y}1,$$

kitaip tariant, tai yra mažesnių už y imties elementų X_k skaičius, padalytas iš n. Tą funkciją galime ir kitaip apibrėžti. Surašykime imties elementus didėjančia tvarka. Gausime vadinamąją *variacinę seką*

$$X_1^* \le X_2^* \le ... \le X_n^*.$$

Tarkime, kad $Y_1 < Y_2 < ... < Y_r$ yra skirtingi imties elementai, pasikartojantys atitinkamai $N_1, N_2, ..., N_r$ kartų. Tada empirinę pasiskirstymo funkciją galime užrašyti pavidalu

$$\mathcal{F}_n(y) = \sum_{Y_k<y} \frac{N_k}{n}.$$

Atkreipsime dėmesį, kad šioje formulėje N_k ir r yra atsitiktiniai dydžiai.

Pasiskirstymo funkcija kiekvienam $y \in R$, aišku, yra atsitiktinis dydis. Jei $(x_1, ..., x_n)$ yra konkreti imtis, tai $\mathcal{F}_n(y)$ realizacija bus pasiskirstymo funkcija

$$F_n(y) = \frac{1}{n} \sum_{x_k<y} 1 = \sum_{y_m<y} \frac{n_m}{n};$$

čia $y_1, ..., y_s$ – skirtingi variacinės sekos realizacijos elementai, pasikartojantys atitinkamai $n_1, ..., n_r$ kartų.

Apibrėšime *empirinius momentus*. Analogiškai II.9 skyreliui (pradiniu) *empiriniu l-uoju momentu* vadinsime

$$A_l = \int_{-\infty}^{\infty} y^l d\mathcal{F}_n(y) = \frac{1}{n} \sum_{k=1}^{n} X_k^l \quad (l = 1, 2, ...).$$

Visi jie, aišku, egzistuoja. Pirmasis empirinis momentas, arba *empirinis vidurkis*, paprastai žymimas $\bar{X}$:

$$\bar{X} = \frac{X_1 + ... + X_n}{n}.$$

l-uoju empiriniu centriniu momentu vadinsime

$$M_l = \int_{-\infty}^{\infty} (y - \bar{X})^l d\mathcal{F}_n(y) = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X})^l \quad (l = 1, 2, ...).$$

Antrąjį empirinį centrinį momentą, arba *empirinę dispersiją*, žymėsime

$$S^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - X)^2.$$

Iš dispersijos savybių (žr. II.9 skyrelį) išplaukia

$$S^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}^2.$$

Dydis

$$S = \Big(\frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2\Big)^{1/2}$$

vadinamas *empiriniu vidutiniu kvadratiniu*, arba *empiriniu standartiniu, nuokrypiu*. Visi tie dydžiai yra atsitiktiniai. Konkrečioms imtims $(x_1, ..., x_n)$ gauname tų dydžių realizacijas

$$\begin{aligned}
a_l &= \frac{1}{n}\sum_{k=1}^{n} x_k^l \quad (l = 1, 2, ...),\\
\bar{x} &= \frac{x_1 + ... + x_n}{n},\\
m_l &= \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^l \quad (l = 1, 2, ...),\\
s^2 &= \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2,\\
s &= \Big(\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\Big)^{1/2}.
\end{aligned}$$

Tarkime, kad $0 < p < 1$. *Empiriniu p-kvantiliu* laikome atsitiktinį dydį

$$\hat{X}_p = X^*_{[np]+1};$$

čia $[np]$ yra skaičiaus np sveikoji dalis, o $X^*_{[np]+1}$ – variacinės sekos narys su numeriu $[np] + 1$. Kai $p = 1/2$, turime *empirinę medianą*; kai $p = 1/4, 3/4$, turime *apatinį* ir *viršutinį empirinius kvartilius*. Kartais dar vartojami empiriniai deciliai, procentiliai ir pan.

Paminėsime dar vieną empirinę charakteristiką – *imties plotį*

$$\max_{1\le k\le n} X_k - \min_{1\le k\le n} X_k = X^*_n - X^*_1.$$

Vartojamos ir kitokios empirinės charakteristikos.

3. STEBĖJIMO DUOMENŲ GRUPAVIMAS

Stebėjimo duomenys retai būna "gražūs" skaičiai. Tokie jie būna tik vidurinių mokyklų matematikos uždavinynuose, o praktiniuose uždaviniuose – paprastai labai "negražūs". Todėl, kai stebėjimo duomenų daug, juos apdoroti gana sunku. Skaičiavimams palengvinti stebėjimo duomenys paprastai apvalinami ir grupuojami.

Intervalas, kuriame telpa stebėjimo duomenys $x_1, ..., x_n$, paprastai skaidomas į intervalus $[\tau_l - h/2, \tau_l + h/2)$; čia $\tau_l = \tau_0 + hl (l = 0, \pm 1, ...)$, h – atitinkamai parinktas skaičius. Duomenys, patekę į intervalą $[\tau_l - h/2, \tau_l + h/2)$, pakeičiami skaičiais τ_l. Taip elgiantis, galima gerokai suprastinti skaičiavimus, jei tik skaičiai τ_0 ir h tinkamai parinkti. Skaičius τ_l reikia parinkti kuo "gražesnius". Kuo skaičius h bus didesnis, tuo paprastesni skaičiavimai, bet tuo didesnė bus padaryta paklaida. Ir atvirkščiai, kuo mažesnis h, tuo skaičiavimai sudėtingesni, bet paklaida mažesnė.

Panagrinėsime, kaip apskaičiuojami empiriniai momentai, naudojantis sugrupuotais duomenimis. Pradinius momentus, gautus iš sugrupuotų duomenų, žymėsime a'_j, centrinius momentus – m'_j. Tarkime, kad intervale $[\tau_l - h/2, \tau_l + h/2)$ yra n_l stebėjimo duomenų. Tada

$$a'_j = \frac{1}{n} \sum_l n_l \tau_l^j;$$

sumuojame pagal visus intervalus, kuriuose yra stebėjimo duomenų. Analogiškai

$$m'_j = \frac{1}{n} \sum_l n_l (\tau_l - a'_1)^j.$$

Tos formulės rodo, kad skaičiavimai su sugrupuotais duomenimis yra paprastesni. Tačiau, kaip minėjome, padarome paklaidas. Šepardas[1] pasiūlė apytiksles formules

$$m'_j \approx \frac{1}{j+1} \sum_{r=0}^{[j/2]} \binom{j+1}{2r+1} \left(\frac{h}{2}\right)^{2r} m_{j-2r},$$

kuriomis naudojantis, dažniausiai pasitaikančiais atvejais gaunamos mažesnės paklaidos. Iš čia, pasinaudoję ryšiu tarp centrinių ir pradinių momentų, galime gauti ir apytiksles formules pradiniams momentams. Tų formulių pagrindimą galima rasti [6] knygoje. Iš jų gauname

[1] W. F. Sheppard – anglų statistikas.

$$
\begin{aligned}
a_1 &\approx a_1', \\
a_2 &\approx a_2' - \frac{1}{12}h^2, \\
a_3 &\approx a_3' - \frac{1}{4}a_1'h^2, \\
a_4 &\approx a_4' - \frac{1}{2}a_2'h^2 + \frac{7}{240}h^4, \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
m_2 &\approx m_2' - \frac{1}{12}h^2, \\
m_3 &\approx m_3', \\
m_4 &\approx m_4' - \frac{1}{2}m_2'h^2 + \frac{7}{240}h^4, \\
&\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}
$$

Matome, kad a_j nuo a_j' bei m_j nuo m_j' skiriasi papildomais dėmenimis – Šepardo pataisomis.

Yra ir kitokių pataisų.

Stebėjimo duomenys dažnai vaizduojami grafiškai įvairiais būdais. Paminėsime tik vadinamąsias *histogramas*. Kiekvienam intervalui $[\tau_l - h/2, \tau_l + h/2)$ brėžiame stačiakampį, kurio pagrindas yra tas intervalas, o plotas lygus n_l/n (kitaip tariant, aukštinė lygi $n_l/(nh)$; žr. 33 pav.).

P a v y z d y s. Automatinės staklės gamina rutuliukus. Iš vienos dienos produkcijos parinkome 60 rutuliukų (kiekvieną kartą grąžindami rutuliuką atgal) ir išmatavome jų skersmenis. Matavimų duomenys šitokie (milimetrais):

7,38	7,29	7,43	7,40	7,36	7,41	7,35	7,31	7,26	7,37
7,28	7,37	7,36	7,35	7,24	7,33	7,42	7,36	7,39	7,35
7,45	7,36	7,42	7,40	7,28	7,38	7,25	7,34	7,33	7,32
7,33	7,30	7,32	7,30	7,39	7,34	7,38	7,39	7,27	7,35
7,35	7,32	7,35	7,27	7,34	7,32	7,38	7,41	7,36	7,44
7,32	7,37	7,31	7,46	7,35	7,35	7,29	7,34	7,30	7,40

Suskaičiuosime pirmuosius keturis imties momentus. Po gana varginančių skaičiavimų galime gauti (autorius naudojosi kišeniniu programuojamu kalkuliatoriumi)

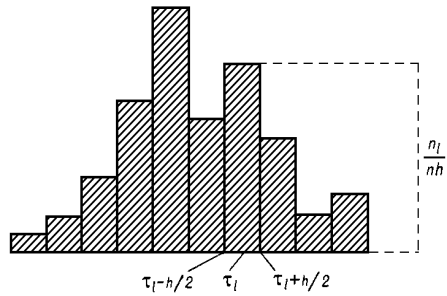
$$
\begin{aligned}
&a_1 = \bar{x} = 7,3490;\ a_2 \approx 54,10267;\ a_3 \approx 396,957690;\ a_4 \approx 2917,641561; \\
&m_2 = s^2 \approx 0,002466;\ m_3 \approx -0,000001;\ m_4 \approx 0,000016.
\end{aligned}
$$

Skaičiavimus palengvinsime, sugrupavę duomenis. Grupavimo intervalo ilgį imsime lygų 0,04. Pasirinktus skaičius τ_l ir n_l surašysime lentelėje:

τ_l	7,26	7,30	7,34	7,38	7,42	7,46
n_l	5	9	20	14	9	3

Vėl apskaičiuosime pirmuosius keturis momentus. Skaičiavimai bus trumpesni. Gausime

$$a_1' \approx 7,354667;\ a_2' \approx 54,093733;\ a_3' \approx 397,879800;\ a_4' \approx 2926,697274;$$
$$m_2' \approx 0,002612;\ m_3' \approx 0,000008;\ m_4' \approx 0,000017.$$



33 pav.

Kaip matome, gautos reikšmės nedaug skiriasi nuo anksčiau apskaičiuotųjų. Su Šepardo pataisomis turėtume

$$a_2'' \approx 54,093600;\ a_3'' \approx 397,876858;\ a_4'' \approx 2926,653999;$$
$$m_2'' \approx 0,002478;\ m_4'' \approx 0,000015.$$

Tos pataisos šiuo atveju nedaug padeda. Praktiniuose skaičiavimuose nereikėtų imti tiek daug skaitmenų po kablelio. Čia norėjome tik pademonstruoti metodo tikslumą.

4. PAKANKAMOSIOS STATISTIKOS

Iš visų galimų statistikų išskirsime labai svarbią jų klasę – pakankamąsias statistikas. Šią sąvoką įvedė R. Fišeris.

Pamėginsime pakankamosios statistikos sąvoką paaiškinti paprastu pavyzdžiu. Sakykime, turime n nepriklausomų eksperimentų. Po kiekvieno eksperimento gali įvykti kuris nors įvykis su nežinoma tikimybe p (Bernulio eksperimentai). Tarkime, kad $X_k = 1$, kai tas įvykis įvyko po k-ojo eksperimento, ir $X_k = 0$, kai jis neįvyko. Imtis $(X_1, ..., X_n)$ rodo skaičių $T = X_1 + ... + X_n$ ir numerius eksperimentų, kai stebimasis įvykis įvyko. Intuityviai aišku, kad tų numerių žinojimas neduoda jokios papildomos informacijos apie p reikšmę. Tai galima paaiškinti ir šitaip. Imkime tokius skaičius

x_k $(k = 1, ..., n)$, lygius 0 arba 1, kad būtų $x_1 + ... + x_n = t$. Sąlyginis $(X_1, ..., X_n)$ pasiskirstymas, kai $T = t$,

$$\mathcal{P}(X_1 = x_1, ..., X_n = x_n \mid T = t) = 1 \Big/ \binom{n}{t}$$

nepriklauso nuo p. Galime laikyti statistiką $T = X_1 + ... + X_n$ pakankama parametrui p įvertinti.

Pateiksime griežtus apibrėžimus.

Nagrinėsime tikimybinį-statistinį modelį $\{R^n, \mathcal{B}^n, \mathcal{P}_\theta\}$, $\theta \in \Theta \subset R^s$, aprašytą 1 skyrelyje. Tarkime, kad $T : R^n \to R^s$ yra $(\mathcal{B}^n, \mathcal{B}^s)$ mati funkcija. Sakysime, kad statistika $T(X)$ yra *pakankama* pasiskirstymų klasei $\{\mathcal{P}_\theta,\ \theta \in \in \Theta\}$, jei egzistuoja sąlyginės tikimybės variantas $\mathcal{P}_\theta\big(B|T(X)\big)$, nepriklausantis nuo θ. Taip pat sakoma, kad statistika $T(X)$ yra pakankama parametrui θ, jei aišku apie kokią Θ kalbama.

Teorijai suprastinti laikysime, kad tikimybinių pasiskirstymų klasė $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ tenkina reikalavimą: egzistuoja σ baigtinis matas μ, apibrėžtas mačioje erdvėje $\{R^n, \mathcal{B}^n\}$, ir μ integruojama funkcija $\mathbf{p}_\theta(x)$, apibrėžta erdvėje R^n, su sąlyga

$$\mathcal{P}_\theta(B) = \int_B \mathbf{p}_\theta(x)\mu(dx)$$

visiems $\theta \in \Theta$ ir visoms $B \in \mathcal{B}^n$; kitaip tariant, matai $\mathcal{P}_\theta$ yra absoliučiai tolydūs mato μ atžvilgiu (žr. V.9 skyrelį). Funkciją $\mathbf{p}_\theta$ galime vadinti *tankiu mato μ atžvilgiu*. Sakome, kad matas μ *dominuoja* pasiskirstymų klasę $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$, o ta klasė yra *dominuota*.

Šios sąvokos vartojamos ir tada, kai $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ yra bet kuri tikimybinių pasiskirstymų klasė, nesusijusi su minėtuoju tikimybiniu-statistiniu modeliu.

Nors dominavimo reikalavimas šiek tiek susiaurina nagrinėjamų pasiskirstymų klasę, bet jį tenkina visi praktiniuose uždaviniuose pasitaikantys pasiskirstymai.

Panagrinėsime keletą pavyzdžių.

1 p a v y z d y s. Tarkime, kad stebimasis atsitiktinis dydis yra pasiskirstęs pagal normalųjį dėsnį $N(a, \sigma^2)$ su nežinomais parametrais $a \in R$, $\sigma > 0$. Tikimybinį matą $\mathcal{P}_{(a,\sigma^2)}$, remiantis Lebego integralu, galima užrašyti šitaip:

$$\mathcal{P}_{(a,\sigma^2)}(B) = \frac{1}{(\sigma\sqrt{2\pi})^n} \int_B \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - a)^2 \Big\} dx.$$

Taigi klasę $\{\mathcal{P}_{(a,\sigma^2)}, a \in R, \sigma > 0\}$ dominuoja Lebego matas.

2 p a v y z d y s. Nesunku suvokti, kad ir bendresniu atveju, kai stebimasis atsitiktinis dydis yra absoliučiai tolydus ir jo tankis yra $p_\theta(u)$,

$$\mathcal{P}_\theta(B) = \int_B p_\theta(x_1)...p_\theta(x_n)dx.$$

Ir šiuo atveju Lebego matas dominuos klasę $\{\mathcal{P}_\theta, \theta \in \Theta\}$.

3 p a v y z d y s. Tarkime, kad stebimasis atsitiktinis dydis yra pasiskirstęs pagal Puasono dėsnį su parametru $\lambda > 0$. Tada atsitiktinė imtis $(X_1, ..., X_n)$ įgyja reikšmes $(x_1, ..., x_n)$ su sveikomis neneigiamomis koordinatėmis ir tikimybėmis

$$\frac{\lambda^{x_1+...+x_n}}{x_1!...x_n!}e^{-\lambda n}.$$

Imkime funkciją

$$\mathbf{p}_\lambda(x_1, ..., x_n) = \frac{\lambda^{x_1+...+x_n}}{x_1!...x_n!}e^{-\lambda n},$$

kai $x_1, ..., x_n$ yra sveikieji neneigiami skaičiai, ir lygią 0 kitais atvejais (arba bet kaip apibrėžtą, kad tik $\mathbf{p}_\lambda$ būtų Borelio funkcija). Visoms erdvės R^n Borelio aibėms pažymėkime $\mu(B)$ skaičių taškų su sveikomis neneigiamomis koordinatėmis aibėje B. Ši aibės funkcija yra σ baigtinis matas mačioje erdvėje $\{R^n, \mathcal{B}^n\}$. Klasės $\{\mathcal{P}_\lambda,\ \lambda > 0\}$ matus galime parašyti šitaip:

$$\mathcal{P}_\lambda(B) = \int_B \mathbf{p}_\lambda(x_1, ..., x_n)\mu(dx).$$

Juos dominuoja matas μ.

4 p a v y z d y s. Nagrinėkime bendresnį atvejį, kai stebimasis atsitiktinis dydis yra diskretusis ir įgyja reikšmes iš baigtinės arba skaičios aibės A su tikimybėmis $p_\theta(u)$, $u \in A$, $\theta \in \Theta$. Kiekvienai $B \in \mathcal{B}^n$ pažymėkime $\mu(B)$ skaičių taškų $(x_1, ..., x_n)$, kurių koordinatės $x_k \in A$ $(k = 1, ..., n)$. Imkime funkciją $\mathbf{p}_\theta(x_1, ..., x_n) = p_\theta(x_1)...p_\theta(x_n)$, kai $x_k \in A$ $(k = 1, ..., n)$, ir laikykime ją lygia 0 kitiems $(x_1, ..., x_n)$. Tada

$$\mathcal{P}_\theta(B) = \int_B \mathbf{p}_\theta(x_1, ..., x_n)\mu(dx).$$

Vėl turime dominuojamą pasiskirstymų klasę $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$.

Pakankamąsias statistikas apibrėžėme, remdamiesi sąlyginėmis tikimybėmis. Jomis ne visada patogu operuoti. Tačiau dominuojamiems pasiskirstymams galima rasti gana paprastą vadinamąjį faktorizavimo kriterijų, nusakytą šitokios teoremos.

Teorema. *Tarkime, kad $T : R^n \to R^s$ yra $(\mathcal{B}^n, \mathcal{B}^s)$ mati funkcija. $T(X)$ yra pakankama pasiskirstymų klasei $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$, dominuojamai mato μ su tankiu $\mathbf{p}_\theta$, tada ir tik tada, kai visiems $\theta \in \Theta$ egzistuoja tokia neneigiama $(\mathcal{B}^s, \mathcal{B}^1)$ mati funkcija $g_\theta : R^s \to R^1$ ir tokia Borelio funkcija $h : R^n \to R^n$, nepriklausanti nuo θ, kad visiems $\theta \in \Theta$*

$$\mu\Big\{\mathbf{p}_\theta(x) \neq g_\theta\big(T(x)\big)h(x)\Big\} = 0,$$

kitaip sakant,

$$\mathbf{p}_\theta(x) = g_\theta\big(T(x)\big)h(x) \tag{1}$$

beveik visur mato μ atžvilgiu.

Į r o d y m a s. Teoremą įrodinėsime tik diskretiesiems pasiskirstymams. (Bendrą jos įrodymą galima rasti, pvz., [20] knygoje.) Šiuo atveju, norint nustatyti pakankamąją statistiką, užtenka reikalauti, kad lygybės

$$\mathcal{P}_\theta(x) = g_\theta\big(T(x)\big)h(x) \tag{2}$$

su pakankamosios statistikos apibrėžime nurodytomis funkcijomis g_θ ir h būtų teisingos tiems $x \in R^n$, kuriems $\mathcal{P}_\theta(x) > 0$.

Pirmiausia įrodysime pakankamumą. Tarkime, kad (2) lygybė yra teisinga. Fiksuokime x ir t. Laikydami $\mathcal{P}_\theta\big(T(X) = t\big) > 0$, iš lygybės

$$\mathcal{P}_\theta\big(X = x|T(X) = t\big) = \frac{\mathcal{P}_\theta\big(X = x, T(X) = t\big)}{\mathcal{P}_\theta\big(T(X) = t\big)}$$

gauname

$$\mathcal{P}_\theta\big(X = x|T(X) = t\big) = 0,$$

kai $T(x) \neq t$, ir

$$\mathcal{P}_\theta\big(X = x|T(X) = t\big) = \frac{\mathcal{P}_\theta(X = x)}{\mathcal{P}_\theta\big(T(X) = t\big)} =$$
$$= \frac{g_\theta(t)h(x)}{g_\theta(t)\sum\limits_{\{y:T(y)=t\}} h(y)} = \frac{h(x)}{\sum\limits_{\{y:T(y)=t\}} h(y)},$$

kai $T(x) = t$. Matome, kad sąlyginė tikimybė $\mathcal{P}_\theta\big(X = x|T(X) = t\big)$ nepriklauso nuo θ. Nuo θ nepriklauso ir $\mathcal{P}_\theta\big(B|T(X) = t\big)$.

Dabar įrodysime sąlygos būtinumą. Tarkime, kad sąlyginė tikimybė $\mathcal{P}_\theta\big(X = x|T(X) = t\big)$ nepriklauso nuo θ ir lygi, sakysime, $w(x, t)$. Jei $T(x) = t$, tai

$$\mathcal{P}_\theta(X = x) = \mathcal{P}_\theta\big(X = x,\ T(X) = t\big) = \mathcal{P}_\theta\big(X = x, |T(X) = t\big)\times$$
$$\times\, \mathcal{P}_\theta\big(T(X) = t\big) = w(x, t)\mathcal{P}_\theta\big(T(X) = t\big).$$

Vadinasi, $\mathcal{P}_\theta(X = x)$ tenkina (2) lygybę. □

(1) formulės dešinės pusės antrasis dauginamasis nuo θ nepriklauso, o norint apskaičiuoti pirmąjį dauginamąjį, priklausantį nuo θ, reikia žinoti tik statistikos T reikšmę ir visai nereikia konkrečių imties x reikšmių. Vaizdžiai kalbant, pakankamoji statistika turi tą pačią informaciją apie parametrą θ, kaip ir visi stebėjimo duomenys.

5 p a v y z d y s. Stebimasis atsitiktinis dydis įgyja reikšmę 1 su nežinoma tikimybe $\alpha, 0 < \alpha < 1$, ir reikšmę 0 su tikimybe $1 - \alpha$. Pagal 4 pavyzdį visiems x_k $(k = 1, ..., n)$, lygiems 0 arba 1,

$$\mathbf{p}_\alpha(x_1, ..., x_n) = \alpha^{x_1+...+x_n}(1-\alpha)^{n-(x_1+...+x_n)} = \\ = (1-\alpha)^n\Big(\frac{\alpha}{1-\alpha}\Big)^{x_1+...+x_n}.$$

Statistika $X_1 + ... + X_n$ yra pakankama parametrui α.

6 p a v y z d y s. Kai turime Puasono dėsnį su nežinomu parametru $\lambda > 0$, visiems sveikiesiems neneigiamiems x_k $(k = 1, ..., n)$ (žr. 3 pvz.)

$$\mathbf{p}_\lambda(x_1, ..., x_n) = e^{-\lambda n}\lambda^{x_1+...+x_n}(x_1!...x_n!)^{-1}.$$

Ir čia statistika $X_1 + ... + X_n$ yra pakankama parametrui λ.

7 p a v y z d y s. Imkime normalųjį dėsnį $N(a, \sigma^2)$. Atsitiktinė imtis $(X_1, ..., X_n)$ turi tankį

$$\frac{1}{(\sigma\sqrt{2\pi})^n}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-a)^2\Big\}. \tag{3}$$

Jei a žinomas, o $\sigma > 0$ – nežinomas, tai statistika

$$\sum_{k=1}^{n}(X_k - a)^2$$

yra pakankama parametrui σ^2. Kai σ žinomas, o $a \in R$ – nežinomas, užrašę (3) pavidalu

$$\frac{1}{(\sigma\sqrt{2\pi})^n}\exp\Big\{-\frac{na^2}{2\sigma^2}+\frac{a}{\sigma^2}\sum_{k=1}^{n}x_k\Big\}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}x_k^2\Big\}, \tag{4}$$

matome, kad statistika $X_1 + ... + X_n$ yra pakankama parametrui a. Jei abu parametrai $a \in R$ ir $\sigma > 0$ yra nežinomi, tai iš (4) išplaukia, kad vektorinė statistika (T_1, T_2), kur

$$T_1 = X_1 + ... + X_n,\ T_2 = X_1^2 + ... + X_n^2,$$

yra pakankama parametrams (a, σ^2).

5. SUDERINTIEJI IR NEPASLINKTIEJI ĮVERČIAI

Kaip minėjome 1 skyrelyje, vienas iš pagrindinių matematinės statistikos uždavinių yra nežinomų pasiskirstymo parametrų įvertinimas. Dažnai, remdamiesi kokiais nors samprotavimais, galime pasirinkti nežinomos pasiskirstymo funkcijos pavidalą, bet į jį įeina nežinomi parametrai. Reikia įvertinti

tuos parametrus. Suprantama, remiamės stebėjimo rezultatais – tinkamai parinkta rezultatų funkcija.

Sakykime, turime tikimybinį-statistinį modelį $\{R^n, \mathcal{B}^n, \mathcal{P}_\theta\}$, $\theta \in \Theta$. Tirsime atvejį, kai Θ yra realiųjų skaičių aibė. Nežinomą parametrą θ reikia įvertinti, remiantis imtimi $X = (X_1, ..., X_n)$. Kiekvieną realiąją statistiką $\theta^*(X) = \theta^*(X_1, ..., X_n)$ vadinsime parametro θ *įverčio funkcija*, arba tiesiog *įverčiu*. Žinoma, ne kiekviena statistika tiks tam reikalui. Ji turi būti kokia nors prasme artima nežinomojo parametro reikšmei. Išskirsime dvi R. Fišerio įvestas įverčių klases: suderintuosius ir nepaslinktuosius įverčius.

Tarkime, kad $\theta_n^*(X_1, ..., X_n)$ $(n = 1, 2, ...)$ yra įverčių seka. Sakysime, jog ji yra suderinta, jei kiekvienam $\varepsilon > 0$

$$\mathcal{P}_\theta\big(|\theta_n^*(X_1, ..., X_n) - \theta| \geq \varepsilon\big) \to 0,$$

kai $n \to \infty$, visiems $\theta \in \Theta$, t. y. θ_n^* konverguoja į θ pagal tikimybę. Tai reiškia, kad su tikimybe, kiek norima artima 1, $\theta_n^*(X)$ kiek norima mažai skirsis nuo θ, jei tik n bus pakankamai didelis. Paprastai (nors tai ir nėra tikslu) ne tik seka, bet ir kiekvienas θ_n^* vadinamas *suderintu*. Šiame apibrėžime tenka operuoti erdve $\{R, \mathcal{B}, P_\theta\}^\infty$.

Galima kalbėti ne tik apie nežinomo parametro θ, bet ir apie jo funkcijos $\psi(\theta)$ suderintąjį įvertį. Tinka analogiškas apibrėžimas.

1 p a v y z d y s. Sakykime, stebime atsitiktinį dydį su nežinomu vidurkiu $a \in R$. Parodysime, kad $\bar{X}$ yra suderintasis parametro a įvertis. Reikia įrodyti, kad $\bar{X}$ konverguoja pagal tikimybę į a, kai $n \to \infty$, t. y. kiekvienam $\varepsilon > 0$

$$\mathcal{P}_a\left\{\left|\frac{1}{n}(X_1 + ... + X_n) - a\right| > \varepsilon\right\} \to 0,$$

kai $n \to \infty$. Tačiau šis teiginys išplaukia iš Chinčino III.3.3. (žr. taip pat III.10.3 pavyzdį) teoremos.

2 p a v y z d y s. Jei stebimasis atsitiktinis dydis turi žinomą vidurkį a, bet nežinomą dispersiją $\sigma^2 > 0$, tai statistika

$$S_0^2 = \frac{1}{n}\sum_{k=0}^{n}(X_k - a)^2$$

yra suderintas σ^2 įvertis. Iš tikrųjų, pritaikę Chinčino teoremą atsitiktiniams dydžiams $(X_k - a)^2$, gauname, kad kiekvienam $\varepsilon > 0$

$$\mathcal{P}_{\sigma^2}\left\{\left|\frac{1}{n}\sum_{k=1}^{n}(X_k - a)^2 - \sigma^2\right| \geq \varepsilon\right\} \to 0,$$

kai $n \to \infty$.

3 p a v y z d y s. Jei stebimasis atsitiktinis dydis turi nežinomą dispersiją $\sigma^2 > 0$ ir jo vidurkis $a \in R$ yra taip pat nežinomas, tai

$$S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

yra suderintas parametro σ^2 įvertis. Įrodysime tą teiginį.

Kadangi

$$S^2 = \frac{1}{n}\sum_{k=1}^{n}\big((X_k - a) - (\bar{X} - a)\big)^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - a)^2 - (\bar{X} - a)^2 =$$
$$= S_0^2 - (\bar{X} - a)^2,$$

tai

$$\mathcal{P}_{(a,\sigma^2)}\Big\{\Big|S^2 - \sigma^2\Big| \geq \varepsilon\Big\} \leq \mathcal{P}_{(a,\sigma^2)}\Big\{\Big|S_0^2 - \sigma^2\Big| \geq \frac{\varepsilon}{2}\Big\} +$$
$$+ \mathcal{P}_{(a,\sigma^2)}\Big\{\Big|\bar{X} - a\Big| \geq \sqrt{\frac{\varepsilon}{2}}\Big\}.$$

Pasinaudoję 1 ir 2 pavyzdžių rezultatais, gauname reikiamą teiginį.

Sakykime, turime realiąją integruojamą statistiką $\theta^* = \theta^*(X_1, ..., X_n)$, kuria norime įvertinti nežinomą parametrą $\theta \in \Theta$. Dydį $M_\theta\theta^*(X_1, ..., X_n) - \theta$ natūralu vadinti *įverčio poslinkiu*, arba *sistemine paklaida*. Čia indeksas θ prie vidurkio ženklo M rodo, kad vidurkis imamas mato $\mathcal{P}_\theta$ atžvilgiu. Jei visiems $\theta \in \Theta$

$$M_\theta\theta^*(X_1, ..., X_n) = \theta,$$

tai sakome, kad įvertis θ^* yra *nepaslinktas*.

Panašiai galime apibūdinti ir parametro funkcijos $\psi(\theta)$ įverčius.

4 p a v y z d y s. $\bar{X}$ yra nepaslinktasis nežinomo vidurkio a įvertis, nes

$$M_a\bar{X} = \frac{1}{n}\sum_{k=1}^{n} M_a X_k = a.$$

5 p a v y z d y s . Jei atsitiktinis dydis turi žinomą vidurkį a, bet nežinomą dispersiją $\sigma^2 > 0$, tai statistika S_0^2 yra nepaslinktasis σ^2 įvertis. Iš tikrųjų

$$M_{\sigma^2}S_0^2 = \frac{1}{n}\sum_{k=1}^{n} M_{\sigma^2}(X_k - a)^2 = \sigma^2.$$

6 p a v y z d y s. Tarkime, kad stebimasis atsitiktinis dydis turi nežinomą vidurkį $a \in R$ ir nežinomą dispersiją $\sigma^2 > 0$. Panagrinėsime statistiką

$$S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - a)^2 - \Big(\frac{1}{n}\sum_{k=1}^{n}(X_k - a)\Big)^2 =$$

$$= \frac{1}{n}\Big(1 - \frac{1}{n}\Big)\sum_{k=1}^{n}(X_k - a)^2 - \frac{1}{n^2}\sum_{1\le j<k\le n}(X_j - a)(X_k - a).$$

Iš čia

$$M_{(a,\sigma^2)}S^2 = \Big(1 - \frac{1}{n}\Big)\sigma^2.$$

Vadinasi, S^2 nėra nepaslinktasis σ^2 įvertis. Kuo mažesnis n, tuo didesnis poslinkis (kai $n = 2$, $M_{(a,\sigma^2)}S^2 = \sigma^2/2$). Dideliems n tas poslinkis yra nedidelis: jis konverguoja į nulį, kai $n \to \infty$. Todėl sakoma, kad įvertis S^2 yra *asimptotiškai nepaslinktas*. Pastebėsime, kad

$$S_1^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$$

yra nepaslinktasis σ^2 įvertis.

Tam pačiam nežinomam parametrui įvertinti galime sudaryti daug nepaslinktųjų įverčių. Matėme, kad $\bar{X}$ yra nepaslinktasis vidurkio a įvertis. Imkime kokius nors pastovius skaičius $q_1, ..., q_n$, tenkinančius sąlygą $q_1 + ... + +q_n = 1$, ir sudarykime statistiką

$$V = q_1X_1 + ... + q_nX_n.$$

Kadangi

$$M_aV = \sum_{k=1}^{n} q_kM_aX_k = a\sum_{k=1}^{n} q_k = a,$$

tai V yra taip pat nepaslinktasis a įvertis. $\bar{X}$ yra specialus jo atvejis, kai $q_1 = ... = q_n = 1/n$. Įvairiai parinkdami skaičius $q_1, ..., q_n$, galime gauti be galo daug nežinomo parametro a įverčių. Kuris iš jų geriausias?

Jei turime du nežinomo parametro θ įverčius, tai natūralu laikyti geresniu tą, kurio reikšmės yra mažiau išsibarsčiusios apie parametrą θ. Išsibarstymo matą galime parinkti įvairiai. Patogus ir dažniausiai vartojamas yra antrasis momentas $M_\theta(\theta^* - \theta)^2$, jei jis egzistuoja. Jei įvertis θ^* yra nepaslinktasis, tai tas dydis yra įverčio dispersija.

Tarkime, kad $\{\mathcal{P}_\theta, \theta \in \Theta\}$ yra netuščia tikimybinių matų sistema erdvėje $\{R^n, \mathcal{B}^n\}$ ir $\psi : \Theta \to R$. Kiekvienam $\theta_0 \in \Theta$ pažymėkime H_{θ_0} klasę visų nepaslinktųjų $\psi(\theta)$ įverčių $T = T(X)$, turinčių dispersiją $D_{\theta_0}T$. Sakome, kad $\psi(\theta)$ įvertis $T_0 \in H_{\theta_0}$ turi *lokaliai mažiausią dispersiją* taške $\theta = \theta_0$, jei visiems $T \in H_{\theta_0}$ teisingos nelygybės

$$M_{\theta_0}\big(T_0 - \psi(\theta_0)\big)^2 \le M_{\theta_0}\big(T - \psi(\theta_0)\big)^2.$$

Jei H yra klasė visų nepaslinktųjų $\psi(\theta)$ įverčių $T = T(X)$, visiems $\theta \in \Theta$ turinčių dispersiją $D_\theta T$, tai sakome, kad $\psi(\theta)$ įvertis $T_0 \in H$ turi *tolygiai mažiausią dispersiją*, kai visiems $\theta \in \Theta$ ir visiems $T \in H$ teisingos nelygybės

$$M_\theta\big(T_0 - \psi(\theta)\big)^2 \le M_\theta\big(T - \psi(\theta)\big)^2.$$

1 teorema. *Pažymėkime K klasę visų realiųjų statistikų W, turinčių savybes $M_\theta W = 0$, $M_\theta W^2 < \infty$. Tarkime, kad įverčių klasė H nėra tuščia. $\psi(\theta)$ įvertis $T_0 \in H$ turi tolygiai mažiausią dispersiją tada ir tik tada, kai kiekvienai statistikai $W \in K$ ir visiems $\theta \in \Theta$*

$$M_\theta W T_0 = 0.$$

Atkreipsime dėmesį, jog iš Koši nelygybės išplaukia, kad $M_\theta W T_0$ egzistuoja.

Į r o d y m a s. B ū t i n u m a s. Tarkime, kad T_0 turi tolygiai mažiausią dispersiją ir kuriam nors $\theta_0 \in \Theta$ bei kuriai nors $W_0 \in K$ teisinga nelygybė $M_{\theta_0} W_0 T_0 \ne 0$. Kiekvienam realiajam λ įvertis $T_0 + \lambda W_0 \in H$. Vidurkis $M_{\theta_0} W_0^2$ negali būti lygus 0, nes tada ir $M_{\theta_0} W_0 T_0$ būtų lygus 0. Imkime $\lambda_0 = -M_{\theta_0} W_0 T_0 / M_{\theta_0} W_0^2$. Tada gautume

$$D_{\theta_0}(T_0 + \lambda_0 W_0) = D_{\theta_0} T_0 + 2\lambda_0 M_{\theta_0} W_0 T_0 + \lambda_0^2 M_{\theta_0} W_0^2 =$$
$$= D_{\theta_0} T_0 - \frac{M_{\theta_0}^2 W_0 T_0}{M_{\theta_0} W_0^2} < D_{\theta_0} T_0,$$

bet tai prieštarautų prielaidai, kad T_0 turi tolygiai mažiausią dispersiją.

P a k a n k a m u m a s. Tarkime, kad kuriam nors įverčiui $T_0 \in H$ ir visiems $\theta \in \Theta$ bei visoms statistikoms $W \in K$ teisinga lygybė

$$M_\theta W T_0 = 0.$$

Kiekvienam įverčiui $T \in H$ turime $T_0 - T \in K$. Vadinasi, visiems $\theta \in \Theta$

$$M_\theta (T_0 - T) T_0 = 0,$$

t. y.

$$M_\theta T_0^2 = M_\theta T_0 T.$$

Iš Koši nelygybės išplaukia

$$M_\theta T_0^2 \le (M_\theta T_0^2 \cdot M_\theta T^2)^{1/2}.$$

Jei $M_\theta T_0^2 = 0$, tai nieko nereikia įrodinėti. Jei $M_\theta T_0^2 > 0$, tai

$$M_\theta T_0^2 \le M_\theta T^2.$$

Tai ir reikėjo įrodyti, nes $M_\theta T_0 = M_\theta T = \psi(\theta)$. □

Įverčiams su lokaliai mažiausia dispersija teisinga analogiška teorema.

2 teorema. *Pažymėkime K_{θ_0} klasę visų realiųjų statistikų W, turinčių savybes $M_{\theta_0} W = 0$, $M_{\theta_0} W^2 < \infty$. Tarkime, kad įverčių klasė H_{θ_0} nėra tuščia. $\psi(\theta)$ įvertis $T_0 \in H_{\theta_0}$ turi lokaliai mažiausią dispersiją taške $\theta = \theta_0$ tada ir tik tada, kai kiekvienai statistikai $W \in K_{\theta_0}$*

$$M_{\theta_0} W T_0 = 0.$$

Į r o d y m a s analogiškas 1 teoremos įrodymui.

1 ir 2 teoremos nurodo būdą, kaip patikrinti, ar įvertis turi mažiausią dispersiją. Konstruojant tokius įverčius, praverčia dvi toliau įrodomos teoremos.

3 (Rao[1]–Blekvelo[2]) teorema. *Tarkime, kad pasiskirstymų sistema $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ turi pakankamąją statistiką T ir įverčių klasė H nėra tuščia. Tada sąlyginio vidurkio variantas $M_\theta(V|T)$, $V \in H$, pagal pakankamosios statistikos apibrėžimą nepriklausantis nuo θ, yra $\psi(\theta)$ nepaslinktasis įvertis. Be to, visiems $\theta \in \Theta$*

$$M_\theta\big(M_\theta(V|T) - \psi(\theta)\big)^2 \le M_\theta\big(V - \psi(\theta)\big)^2. \tag{1}$$

Ši formulė virsta lygybe visiems $\theta \in \Theta$ tada ir tik tada, kai $V = M_\theta(V|T)$ beveik visur matų $\mathcal{P}_\theta$ atžvilgiu.

Į r o d y m a s. Pagal II.10.4 teoremą

$$M_\theta\big(M_\theta(V|T)\big) = M_\theta V.$$

Todėl $M(V|T)$ yra nepaslinktasis $\psi(\theta)$ įvertis. Vadinasi, pakanka visiems $\theta \in \Theta$ įrodyti nelygybę

$$M_\theta\big(M_\theta(V|T)\big)^2 \le M_\theta V^2.$$

Kadangi $M_\theta V^2 = M_\theta\big(M_\theta(V^2|T)\big)$, tai pakanka įrodyti, kad beveik visur mato $\mathcal{P}_\theta$ atžvilgiu

$$\big(M_\theta(V|T)\big)^2 \le M_\theta(V^2|T). \tag{2}$$

Nesunku suvokti, kad atsitiktiniam vidurkiui $M_\theta(V|T)$ beveik visur mato $\mathcal{P}_\theta$ atžvilgiu galima taikyti Koši nelygybę. Gausime, kad beveik visur mato $\mathcal{P}_\theta$ atžvilgiu

---

[1] Calyampudi Radhakrishna Rao (g. 1920 m.) – indų matematikas.

[2] David Blackwell (g. 1919 m.) – amerikiečių matematikas.

$$M_\theta^2(V|T) \le M_\theta(V^2|T) \cdot M_\theta(1|T),$$

o tai ir yra (2) nelygybė.

Jei (1) formulėje kuriam nors θ turime lygybės ženklą, tai teisinga lygybė

$$M_\theta\big(M_\theta(V|T)\big)^2 = M_\theta V^2. \tag{3}$$

Tačiau pagal II.10.8 teoremą

$$M_\theta\big(VM_\theta(V|T)\big) = M_\theta\big(M_\theta\big(VM_\theta(V|T)T\big)\big) = \\ = M_\theta\big(M_\theta(V|T)M_\theta(V|T)\big).$$

Iš šios ir (3) formulės

$$M_\theta\big(V - M_\theta(V|T)\big)^2 = M_\theta V^2 - 2M_\theta\big(VM_\theta(V|T)\big) + \\ + M_\theta\big(M_\theta(V|T)\big)^2 = M_\theta V^2 - M_\theta\big(M_\theta(V|T)\big)^2 = 0.$$

Iš čia išplaukia, kad $V = M_\theta(V|T)$ beveik visur mato $\mathcal{P}_\theta$ atžvilgiu. □

Kita teorema pagrįsta pilnosios statistikos sąvoka.

Sakoma, kad pasiskirstymų sistema $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$, nusakyta erdvėje $\{R^k, \mathcal{B}^k\}$, yra *pilna*, jei kiekviena Borelio funkcija $\varphi : R^k \to R$ su sąlyga

$$\int_{R^k} \varphi(x)\mathcal{P}_\theta(dx) = 0,\ \theta \in \Theta,$$

yra beveik visur lygi 0 visų matų $\mathcal{P}_\theta,\ \theta \in \Theta$, prasme.

7 p a v y z d y s. Nagrinėkime binominių pasiskirstymų sistemą $\{P_\theta,\ 0 < \theta < 1\}$, nusakytą erdvėje $\{R, \mathcal{B}\}$ lygybėmis

$$P_\theta(y) = \begin{cases} \binom{s}{y}\theta^y(1-\theta)^{s-y}, & \text{kai } y = 0, 1, ..., s, \\ 0 & \text{visiems kitiems } y. \end{cases}$$

Parodysime, kad ši sistema yra pilna. Tarkime, kad kokiai nors Borelio funkcijai φ

$$\int_R \varphi(x)P_\theta(dx) = \sum_{k=0}^{s} \varphi(k)\binom{s}{k}\theta^k(1-\theta)^{s-k} = 0,\ 0 < \theta < 1. \tag{4}$$

Pažymėkime $z = \theta/(1-\theta)$. Tada (4) lygybė virsta šitokia:

$$\sum_{k=0}^{s} \varphi(k)\binom{s}{k}z^k = 0,\ 0 < z < \infty.$$

Kadangi s-ojo laipsnio polinomas turi ne daugiau kaip s skirtingų šaknų, o šis polinomas virsta nuliu visiems teigiamiems z, tai visi jo koeficientai turi būti lygūs 0. Vadinasi, $\varphi(k) = 0$ $(k = 0, 1, ..., s)$.

Statistika T yra vadinama *pilnąja* pasiskirstymų sistemos $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ *statistika*, jei jos indukuotų pasiskirstymų sistema $\{\mathcal{P}_\theta^T,\ \theta \in \Theta\}$ yra pilna.

8 p a v y z d y s. Sakykime, atliekame n Bernulio eksperimentų, kuriuose stebime kokį nors įvykį. To įvykio tikimybė $p \in (0,1)$ yra nežinoma, bet ta pati visuose eksperimentuose. Imkime statistiką $T(X) = X_1 + ... + X_n$ (įvykių skaičių, atlikus n eksperimentų). Jos generuota pasiskirstymų sistema nusakyta 7 pavyzdyje. Vadinasi, statistika T yra pilna šiai pasiskirstymų sistemai.

Dažnai statistikos pilnumą padeda nustatyti tolesnė teorema.

4 teorema. *Tarkime, kad pasiskirstymų sistema* $\{\mathcal{P}_\theta,\ \theta \in \Theta\},\ \Theta \subset R$, *nusakyta erdvėje* $\{R, \mathcal{B}\}$, *yra dominuojama kokio nors* σ *baigtinio mato su tankio funkcija*

$$p_\theta(y) = h(y)\exp\{\theta \cdot U(y) + V(\theta)\};$$

čia h, U *yra Borelio funkcijos. Tarkime, kad aibėje* Θ *telpa neišsigimęs intervalas. Tada*

$$T(X) = \sum_{k=1}^{n} U(X_k)$$

yra pilna ir pakankama pasiskirstymų sistemos $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ *statistika.*

Šios teoremos įrodymą galima rasti, pavyzdžiui, [39], p. 98.

5 (Lemano[1]–Šefė[2]) teorema. *Tarkime, kad pasiskirstymų sistema* $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ *turi pilną pakankamąją statistiką* T *ir įverčių klasė* H *nėra tuščia. Tada kiekvienam* $V \in H$ *sąlyginis vidurkis* $M_\theta(V|T)$ *yra nepaslinktasis* $\psi(\theta)$ *įvertis su tolygiai mažiausia dispersija.*

Į r o d y m a s. Pirmiausia įrodysime, kad bet kuriems $V_1 \in H,\ V_2 \in H$ beveik visur matų $\mathcal{P}_\theta$ atžvilgiu

$$M_\theta(V_1|T) = M_\theta(V_2|T). \tag{5}$$

Pažymėkime $\mathcal{P}_\theta^T$ matą, indukuotą statistikos T. Sakykime, $Q = T(R^n)$. Kadangi $M_\theta(V_1|T)$ ir $M_\theta(V_2|T)$ yra nepaslinktieji $\psi(\theta)$ įverčiai, tai

$$\int_Q \big(M_\theta(V_1|T) - M_\theta(V_2|T)\big)d\mathcal{P}_\theta^T$$

visiems $\theta \in \Theta$. Iš matų $\mathcal{P}_\theta^T$ sistemos pilnumo išplaukia, kad (5) teisinga beveik visur matų $\mathcal{P}_\theta^T$ atžvilgiu. Taigi visi $M_\theta(V|T),\ V \in H$, sutampa vienas su kitu beveik visur matų $\mathcal{P}_\theta$ atžvilgiu. Pagal 1 teoremą įvertis $M_\theta(V|T)$ turi tolygiai mažiausią dispersiją. □

---

[1] Erich Leo Lehmann (g. 1917 m.) – amerikiečių matematikas.

[2] Henry Scheffé (g. 1907 m.) – amerikiečių matematikas.

6. ĮVERČIŲ SUDARYMO METODAI

Įverčius galime sudaryti įvairiais būdais. Tačiau ne visi bet kaip sudaryti įverčiai bus tinkami. Todėl naudinga žinoti bendrus metodus, kuriuos taikant, galima gauti gerus įverčius. Tokių metodų yra nemažai. Nagrinėsime tik du iš jų.

Istoriškai pirmasis iš tokių metodų yra vadinamasis *momentų metodas*, pasiūlytas K. Pirsono. Tarkime, kad pasiskirstymų klasė $\{\mathcal{P}_\theta,\ \theta \in \Theta\},\ \Theta \subset \subset R^s$, priklauso nuo vektorinio parametro $\theta = (\theta_1, ..., \theta_s)$ ir tie pasiskirstymai turi s pirmųjų momentų $\alpha_r(\theta_1, ..., \theta_s)$ $(r = 1, ..., s)$. Imame pirmuosius s empirinių momentų A_r $(r = 1, ..., s)$ ir prilyginame juos atitinkamiems teoriniams momentams. Gauname s lygčių su s nežinomųjų sistemą

$$A_r = \alpha_r(\theta_1, ..., \theta_s) \quad (r = 1, ..., s).$$

Spręsime ją $\theta_1, ..., \theta_s$ atžvilgiu. Jei sprendiniai

$$\theta_r^* = \theta_r^*(X_1, ..., X_n) \quad (r = 1, ..., s)$$

egzistuoja, tai juos ir laikysime parametrų $\theta_1, ..., \theta_s$ įverčiais.

Žinoma, užuot ėmę iš eilės pirmuosius s pradinių momentų, galime imti s kitų momentų, nebūtinai pirmųjų ir nebūtinai pradinių, jei tik tie momentai egzistuoja. Tačiau iš pirmųjų momentų gauname paprastesnius ir tikslesnius įverčius.

Paaiškinsime šį metodą pavyzdžiu.

1 p a v y z d y s. Tarkime, kad stebimasis atsitiktinis dydis yra pasiskirstęs pagal normalųjį dėsnį $N(a, \sigma^2)$ su nežinomais parametrais $a \in R$ ir $\sigma > 0$. Imkime pirmąjį pradinį ir antrąjį centrinį momentus, lygindami teorinius ir empirinius momentus

$$\begin{cases} \bar{X} = a, \\ S^2 = \sigma^2. \end{cases}$$

Iš karto gauname a ir σ^2 įverčius $\bar{X}$ ir S^2.

Momentų metodas yra labai paprastas, bet jį taikant, ne visada galima gauti gerus rezultatus. Geresni įverčiai gaunami *didžiausio tikėtinumo metodu*, pasiūlytu R. Fišerio 1912 m.

Tirsime pasiskirstymus $\{\mathcal{P}_\theta,\ \theta \in \Theta\},\ \Theta \subset R^s$, priklausančius nuo kelių realių parametrų $\theta = (\theta_1, ..., \theta_s)$ ir dominuojamus mato μ su tankio funkcija

$$\mathbf{p}_\theta(x) = p_\theta(x_1)...p_\theta(x_n)$$

(žr. 4 skyrelį). Parametro $\theta \in \Theta$ reikšmėms ir taškams $x = (x_1, ..., x_n) \in R^n$ su sąlyga $\mathbf{p}_\theta(x) > 0$ apibrėžkime funkcijas

$$h(\theta, x) = p_\theta(x_1)...p_\theta(x_n)$$

ir

$$l(\theta, x) = \ln h(\theta, x) = \ln p_\theta(x_1) + ... + \ln p_\theta(x_n).$$

Funkcija

$$H(\theta) = H(\theta, X) = p_\theta(X_1)...p_\theta(X_n)$$

vadinama *tikėtinumo funkcija*. Nagrinėjamas ir jos logaritmas

$$L(\theta) = L(\theta, X) = \ln p_\theta(X_1) + ... + \ln p_\theta(X_n).$$

$h(\theta, x)$ ir $l(\theta, x)$ yra jų realizacijos. Jei egzistuoja statistika $\theta^* = \theta^*(X_1, ..., X_n)$, tenkinanti sąlygą

$$l(\theta^*) = \sup_{\theta \in \Theta} l(\theta),$$

tai ji vadinama parametro θ *didžiausio tikėtinumo įverčiu*.

Jei Θ yra s-matis intervalas, $l(\theta)$ turi maksimumą jo viduje ir yra diferencijuojama $\theta_1, ..., \theta_s$ atžvilgiu, tai jos maksimumo taške

$$\frac{\partial H(\theta)}{\partial \theta_r} = 0 \quad (r = 1, ..., s)$$

arba

$$\frac{\partial L(\theta)}{\partial \theta_r} = 0 \quad (r = 1, ..., s).$$

Vadinasi, didžiausio tikėtinumo įvertis yra šių, vadinamųjų didžiausio tikėtinumo, lygčių sprendinys. Dažnai tos lygtys turi tik vieną sprendinį.

2 p a v y z d y s. Tirkime Puasono pasiskirstymą su nežinomu parametru $\lambda > 0$. Pagal 4.3 pavyzdį tikėtinumo funkcija

$$H(\lambda) = \frac{\lambda^{X_1 + ... + X_n}}{X_1!...X_n!} e^{-\lambda n}.$$

Iš čia

$$L(\lambda) = -\lambda n + \sum_{k=1}^{n} (X_k \ln \lambda - \ln X_k!).$$

Diferencijuodami pagal λ, gauname didžiausio tikėtinumo lygtį

$$\frac{\partial L}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{k=1}^{n} X_k = 0.$$

Iš jos randame didžiausio tikėtinumo įvertį

$$\lambda^* = \bar{X}.$$

3 p a v y z d y s. Panagrinėkime normalųjį dėsnį $N(a, \sigma^2)$ su žinomu σ^2, bet nežinomu $a \in R$. Pagal 4.1 pavyzdį tikėtinumo funkcija

$$H(a) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (X_k - a)^2 \Big\}$$

ir jos logaritmas

$$L(a) = -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (X_k - a)^2.$$

Didžiausio tikėtinumo lygtis yra

$$\frac{\partial L}{\partial a} = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - a) = 0.$$

Iš jos gauname didžiausio tikėtinumo įvertį

$$a^* = \bar{X}.$$

4 p a v y z d y s. Jei turime normalųjį dėsnį $N(a, \sigma^2)$ su ž i n o m u a, bet n e ž i n o m a dispersija $\sigma^2 > 0$, tai, diferencijuodami tikėtinumo funkcijos logaritmą

$$L(\sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (X_k - a)^2$$

pagal σ^2, gauname didžiausio tikėtinumo lygtį

$$\frac{\partial L}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} (X_k - a)^2.$$

Iš jos gauname didžiausio tikėtinumo įvertį

$$(\sigma^2)^* = S_0^2.$$

5 p a v y z d y s. Jei turime normalųjį pasiskirstymą su abiem nežinomais parametrais $a \in R$ ir $\sigma^2 > 0$, tai, kaip ir 3 bei 4 pavyzdžiuose, išdiferencijavę tikėtinumo funkcijos logaritmą pagal a ir σ^2, gauname lygčių sistemą

$$\frac{\partial L}{\partial a} = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - a) = 0,$$

$$\frac{\partial L}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} (X_k - a)^2 = 0.$$

Iš pirmosios lygties

$$a^* = \bar{X}.$$

Įrašę šį sprendinį į antrąją lygtį, gauname

$$-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{k=1}^{n}(X_k-\bar{X})^2=0.$$

Iš jos

$$(\sigma^2)^*=S^2.$$

Vadinasi, $a^*=\bar{X}$ ir $(\sigma^2)^*=S^2$ yra parametrų a ir σ^2 didžiausio tikėtinumo įverčiai.

Paminėsime be įrodymo (žr. [6], 33.3 skyrelį): jei yra išpildytos gana bendros sąlygos, tai didžiausio tikėtinumo įverčiai yra suderinti ir asimptotiškai nepaslinki.

7. EFEKTYVIEJI ĮVERČIAI

Jei pasiskirstymų klasė ir įverčiai tenkina gana bendras sąlygas, tai galima rasti tokių įverčių dispersijų apatinį rėžį.

(Rao–Kramero) teorema. *Tarkime, kad pasiskirstymų klasė* $\{\mathcal{P}_\theta,\ \theta\in\Theta\}$, *dominuojama mato* μ *su tankio funkcija* $\mathbf{p}_\theta(x)$, *priklauso nuo vieno realaus parametro* θ, Θ *yra visa realiųjų skaičių tiesė arba intervalas, o* $T(X)$ *– integruojama statistika. Pareikalaukime, kad aibė* $A=\{x:\mathbf{p}_\theta(x)>0\}$ *nepriklausytų nuo* θ *ir visiems* $\theta\in\Theta$ *būtų*

$$\frac{\partial}{\partial\theta}\int_A \mathbf{p}_\theta(x)\mu(dx)=\int_A\frac{\partial\mathbf{p}_\theta(x)}{\partial\theta}\mu(dx),$$
$$\frac{\partial}{\partial\theta}\int_A T(x)\mathbf{p}_\theta(x)\mu(dx)=\int_A T(x)\frac{\partial\mathbf{p}_\theta(x)}{\partial\theta}\mu(dx),$$
$$I(\theta)=\int_A\Big(\frac{\partial\ln\mathbf{p}_\theta(x)}{\partial\theta}\Big)^2\mathbf{p}_\theta(x)\mu(dx)>0;$$

sakykime, kad čia nurodyti integralai ir išvestinės egzistuoja. Jei $M_\theta T(X)=\theta+b(\theta)$ *ir funkcija* $b(\theta)$ *yra diferencijuojama, tai*

$$M_\theta\big(T(X)-\theta\big)^2\ge\frac{\big(1+b'(\theta)\big)^2}{I(\theta)};$$

atskiru atveju, kai $T(X)$ *yra nepaslinktasis* θ *įvertis,*

$$D_\theta T(X)\ge\frac{I}{I(\theta)}. \tag{1}$$

Teoremos sąlygose minimų integralų integravimo sritis yra aibė A. Tačiau galima integruoti ir visoje erdvėje R^n, jei susitarsime laikyti pointegralines funkcijas lygias 0 tuose taškuose, kuriuose jos nėra apibrėžtos (juk ten $\mathbf{p}_\theta(x) = 0$).

Į r o d y m a s. Imkime lygybes

$$\mathcal{P}_\theta(R^n) = \int_A \mathbf{p}_\theta(x)\mu((dx),$$
$$M_\theta T(X) = \int_A T(x)\mathbf{p}_\theta(x)\mu(dx),$$

arba

$$\int_A \mathbf{p}_\theta(x)\mu(dx) = 1,$$
$$\int_A T(x)\mathbf{p}_\theta(x)\mu(dx) = \theta + b(\theta).$$

Diferencijuokime jas pagal θ. Teoremos sąlygos leidžia sukeisti diferencijavimo ir integravimo tvarką. Gausime

$$\int_A \frac{\partial \mathbf{p}_\theta(x)}{\partial\theta}\mu(dx) = 0, \tag{2}$$

$$\int_A T(x)\frac{\partial \mathbf{p}_\theta(x)}{\partial\theta}\mu(dx) = 1 + b'(\theta).$$

Atėmę iš antrosios lygybės pirmąją, padaugintą iš θ, turime

$$\int_A \big(T(x) - \theta\big)\frac{\partial \mathbf{p}_\theta(x)}{\partial\theta}\mu(dx) = 1 + b'(\theta),$$

arba

$$\int_A \big(T(x) - \theta\big)\frac{\partial \ln \mathbf{p}_\theta(x)}{\partial\theta}\mathcal{P}_\theta(dx) = 1 + b'(\theta).$$

Taikome Koši nelygybę (V.9.13 teoremą) ir gauname

$$\big(1 + b'(\theta)\big)^2 \le \int_A \big(T(x) - \theta\big)^2\mathcal{P}_\theta(dx)\int_A \Big(\frac{\partial \ln \mathbf{p}_\theta(x)}{\partial\theta}\Big)^2\mathcal{P}_\theta(dx). \tag{3}$$

Kadangi

$$\int_A \big(T(x) - \theta\big)^2\mathcal{P}_\theta(dx) = M_\theta\big(T(X) - \theta\big)^2,$$

tai (3) nelygybė ir yra ieškomoji. □

(1) nelygybė yra vadinama *Rao-Kramero nelygybe*.

Nepaslinktuosius įverčius $T(X)$, tenkinančius teoremos sąlygas, galima būtų vadinti reguliariaisiais. Iš Rao-Kramero nelygybės gaunamas tokių įverčių dispersijų įvertis iš apačios. Reguliariuosius įverčius $T(X)$, kuriems teisinga lygybė

$$D_\theta T(X) = \frac{1}{I(\theta)},$$

vadinsime *efektyviais*.

Panagrinėsime, kada Rao-Kramero nelygybė virsta lygybe. Kaip matyti iš teoremos įrodymo, tai įvyksta tada ir tik tada, kai (3) nelygybė virsta lygybe, o (3), kaip žinome iš V.9.13 teoremos, virsta lygybe tada ir tik tada, kai egzistuoja konstantos c_1 ir c_2, iš kurių bent viena nelygi 0, tenkinančios sąlygą

$$c_1 \frac{\partial \ln \mathbf{p}_\theta(x)}{\partial \theta} + c_2\big(T(x) - \theta\big) = 0$$

beveik visur mato $\mathcal{P}_\theta$ atžvilgiu.

Jei $c_1 = 0$, tai tada beveik visur $T(x) = \theta$. Tačiau šis atvejis netinka, nes įvertis $T(X)$ turi nepriklausyti nuo θ. Jei $c_2 = 0$, tai beveik visur mato $\mathcal{P}_\theta$ atžvilgiu

$$\frac{\partial \ln \mathbf{p}_\theta(x)}{\partial \theta} = 0;$$

tada būtų $I(\theta) = 0$, bet tai prieštarautų teoremos sąlygoms.

Todėl lieka atvejis, kai $c_1 \neq 0,\ c_2 \neq 0$. Tada beveik visur

$$\frac{\partial \ln \mathbf{p}_\theta(x)}{\partial \theta} = \kappa\big(T(x) - \theta\big); \tag{4}$$

čia $\kappa = \kappa(\theta)$ gali priklausyti nuo θ, bet ne nuo x. Jei $\theta^* = T(X)$ yra efektyvusis θ įvertis, tai iš (4) išplaukia, kad jį galima gauti didžiausio tikėtinumo metodu, nes tokie įverčiai yra lygties

$$\frac{\partial \ln \mathbf{p}_\theta(x)}{\partial \theta} = 0$$

sprendiniai. Didžiausio tikėtinumo lygties sprendiniai $\theta^* = \text{const}$ paprastai atmetami, nes jie atitinka atvejį $\kappa(\theta) = 0$.

Iš (4) išplaukia: jei $T(X)$ yra efektyvusis įvertis, tai jis yra pasiskirstymų klasės $\{\mathcal{P}_\theta,\ \theta \in \Theta\}$ pakankamoji statistika. Iš tikrųjų, suintegravę (4), gausime

$$\mathbf{p}_\theta(x) = h(x) \exp\{u(\theta)T(x) + v(\theta)\};$$

čia $h(x)$ priklauso tik nuo x, o $u(\theta)$ ir $v(\theta)$ tik nuo θ, be to, $u'(\theta) \neq 0$.

Išnagrinėsime keletą pavyzdžių. Skaičiavimams pravartu šiek tiek pertvarkyti reiškinį $I(\theta)$. Pastebėsime, kad

$$I(\theta) = M_\theta\Big(\frac{\partial \ln \mathbf{p}_\theta(X)}{\partial \theta}\Big)^2.$$

Iš (2) išplaukia, kad

$$M_\theta \frac{\partial \ln \mathbf{p}_\theta(X)}{\partial \theta} = \int_A \frac{\partial \ln \mathbf{p}_\theta(x)}{\partial \theta} \mathbf{p}_\theta(x)\mu(dx) = \int_A \frac{\partial \mathbf{p}_\theta(x)}{\partial \theta}\mu(dx) = 0.$$

Todėl

$$I(\theta) = D_\theta \frac{\partial \ln \mathbf{p}_\theta(X)}{\partial \theta}.$$

Kadangi $\mathbf{p}_\theta(X) = p_\theta(X_1)...p_\theta(X_n)$, tai

$$I(\theta) = nD_\theta \frac{\partial \ln p_\theta(X_1)}{\partial \theta}.$$

Reiškinį $I(\theta)$ galima užrašyti ir kitaip, jei be teoremos sąlygų dar teisinga ir sąlyga

$$\int \left| \frac{\partial^2 \ln \mathbf{p}_\theta(x)}{\partial \theta^2} \right| \mathbf{p}_\theta(x)\mu(dx) < \infty.$$

Tada egzistuoja vidurkis

$$M_\theta \frac{\partial^2 \ln \mathbf{p}_\theta(X)}{\partial \theta^2} = \int \frac{\frac{\partial^2 \mathbf{p}_\theta(x)}{\partial \theta^2} \mathbf{p}_\theta(x) - \left(\frac{\partial \mathbf{p}_\theta(x)}{\partial \theta}\right)^2}{\mathbf{p}_\theta^2(x)} \mathbf{p}_\theta(x)\mu(dx) =$$

$$= \int \frac{\partial^2 \mathbf{p}_\theta(x)}{\partial \theta^2}\mu(dx) - \int \left(\frac{\partial \ln \mathbf{p}_\theta(x)}{\partial \theta}\right)^2 \mathbf{p}_\theta(x)\mu(dx).$$

Kadangi antrasis dešinės pusės narys yra $I(\theta)$ ir

$$\int \frac{\partial^2 \mathbf{p}_\theta(x)}{\partial \theta^2}\mu(dx) = \frac{\partial^2}{\partial \theta^2} \int \mathbf{p}_\theta(x)\mu(dx) = 0,$$

tai

$$I(\theta) = -M_\theta \frac{\partial^2 \ln \mathbf{p}_\theta(X)}{\partial \theta^2} = -nM_\theta \frac{\partial^2 \ln p_\theta(X_1)}{\partial \theta^2}.$$

1 p a v y z d y s. Nagrinėsime atsitiktinį dydį, pasiskirsčiusį pagal binominį dėsnį. Tarkime, kad jis įgyja reikšmes $0, 1, ..., N$, reikšmę x_1 įgyja su tikimybe

$$\binom{N}{x_1} \alpha^{x_1}(1-\alpha)^{N-x_1};$$

čia $\alpha \in (0,1)$ yra nežinomas parametras. Nesunku patikrinti, kad šiuo atveju pasiskirstymų klasė tenkina Rao–Kramero teoremos sąlygas, be to,

$$I(\alpha) = nD_\alpha\Big(\frac{X_1}{\alpha(1-\alpha)} - \frac{N}{1-\alpha}\Big) = \frac{Nn}{\alpha(1-\alpha)}.$$

Įvertis $\bar{X}/N$ taip pat tenkina teoremos sąlygas. Kadangi

$$M(\bar{X}/N) = \frac{1}{nN}\sum_{k=1}^{n} M_\alpha X_k = \alpha,$$

$$D_\alpha(\bar{X}/N) = \frac{1}{n^2N^2}\sum_{k=1}^{n} D_\alpha X_k = \frac{\alpha(1-\alpha)}{Nn},$$

tai $\bar{X}/N$ yra efektyvusis α įvertis.

2 p a v y z d y s. Tirsime atsitiktinį dydį, pasiskirsčiusį pagal Puasono dėsnį su nežinomu parametru $\lambda > 0$. Šiuo atveju (žr. 4.6 ir 6.2 pavyzdžius) sveikiesiems neneigiamiems x_1

$$p_\lambda(x_1) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda},$$

$$\frac{\partial^2}{\partial\lambda^2}\ln p_\lambda(x_1) = -\frac{x_1}{\lambda^2},$$

$$I(\lambda) = -nM_\lambda\Big(-\frac{X_1}{\lambda^2}\Big) = \frac{n}{\lambda^2}M_\lambda X_1 = \frac{n}{\lambda}.$$

Kadangi

$$M\bar{X} = \lambda,\ D_\lambda\bar{X} = \lambda/n,$$

tai $\bar{X}$ yra efektyvusis λ įvertis.

5 skyrelyje matėme, kad ir $V = q_1X_1 + ... + q_nX_n$ su konstantomis $q_1, ..., q_n$, tenkinančiomis sąlygą $q_1 + ... + q_n = 1$, yra nepaslinktasis vidurkio įvertis. Jo dispersija nagrinėjamu atveju

$$D_\lambda V = \sum_{k=1}^{n} D_\lambda(q_kX_k) = \lambda\sum_{k=1}^{n} q_k^2 = \lambda\sum_{k=1}^{n}\Big(q_k - \frac{1}{n}\Big)^2 + \frac{\lambda}{n}.$$

Matome, kad V yra efektyvusis λ įvertis tada ir tik tada, kai $q_k = 1/n$.

3 p a v y z d y s. Nagrinėsime atsitiktinį dydį, pasiskirsčiusį pagal normalųjį dėsnį $N(a, \sigma^2)$ su ž i n o m u σ^2, bet n e ž i n o m u $a \in R$. Parodysime, kad ir čia $\bar{X}$ yra efektyvusis a įvertis. Visiems realiesiems x_1 (žr. 6.3 pavyzdį)

$$p_a(x_1) = \frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{(x_1-a)^2}{2\sigma^2}\Big),$$

$$\frac{\partial^2}{\partial a^2}\ln p_a(x_1) = -\frac{1}{\sigma^2}.$$

Po paprastų skaičiavimų

$$I(a) = \frac{n}{\sigma^2},\ D_a\bar{X} = \frac{\sigma^2}{n}.$$

4 p a v y z d y s. Vėl tirsime atsitiktinį dydį, pasiskirsčiusį pagal normalųjį dėsnį, bet dabar laikysime a ž i n o m u, o $\sigma^2 > 0$ – n e ž i n o m u. Parodysime, kad S_0^2 yra efektyvusis σ^2 įvertis. Visiems realiesiems x_1

$$\frac{\partial}{\partial(\sigma^2)} \ln p_{\sigma^2}(x_1) = \frac{(x_1 - a)^2}{2\sigma^4} - \frac{1}{2\sigma^2}.$$

Apskaičiuosime $I(\sigma^2)$ ir $D_{\sigma^2} S_0^2$:

$$I(\sigma^2) = \frac{n}{4\sigma^8} D_{\sigma^2}(X_1 - a)^2,$$
$$D_{\sigma^2} S_0^2 = \frac{1}{n} D_{\sigma^2}(X_1 - a)^2.$$

Abiejuose reiškiniuose turime to paties atsitiktinio dydžio dispersiją. Apskaičiuosime ją. Iš lygybės

$$D_{\sigma^2}(X_1 - a)^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (u-a)^4 \exp\Big(-\frac{(u-a)^2}{2\sigma^2}\Big) du -$$
$$- \left(\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (u-a)^2 \exp\Big(-\frac{(u-a)^2}{2\sigma^2}\Big) du\right)^2$$

po pakeitimo $u - a = \sigma y$ gauname

$$D_{\sigma^2}(X_1 - a)^2 = \frac{\sigma^4}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^4 e^{-y^2/2} dy - \left(\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy\right)^2.$$

Pirmąjį integralą integruojame dalimis

$$\int_{-\infty}^{\infty} y^4 e^{-y^2/2} dy = \int_{-\infty}^{\infty} (-y^3) de^{-y^2/2} = 3 \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy.$$

Visai taip pat

$$\int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy = \int_{-\infty}^{\infty} (-y) de^{-y^2/2} = \int_{-\infty}^{\infty} e^{-y^2/2} dy = \sqrt{2\pi}.$$

Čia galėjome naudotis ir II.9.1 pavyzdžio rezultatais. Todėl

$$D_{\sigma^2}(X_1 - a)^2 = 2\sigma^4$$

ir

$$I(\sigma^2) = \frac{n}{2\sigma^4}, \; D_{\sigma^2} S_0^2 = \frac{2\sigma^4}{n}.$$

Matome, kad S_0^2 yra efektyvusis σ^2 įvertis.

Jei vietoj S_0^2 imtume S_1^2, kuris, kaip matėme 5.6 pavyzdyje, yra nepaslinktasis σ^2 įvertis, tai gautume (apskaičiuokite!)

$$D_{\sigma^2} S_1^2 = \frac{2\sigma^4}{n-1}.$$

Vadinasi, S_1^2 nėra efektyvusis σ^2 įvertis. Tačiau

$$\frac{1/I(\sigma^2)}{D_{\sigma^2} S_1^2} = 1 - \frac{1}{n}$$

konverguoja į 1, kai $n \to \infty$. Tokį įvertį galima būtų pavadinti *asimptotiškai efektyviu*.

Pastebėsime (žr. [6], 33.3 skyrelį), kad didžiausio tikėtinumo įverčiai, kai patenkintos gana bendros sąlygos, yra asimptotiškai efektyvūs.

Jei θ^* yra bet kuris nepaslinktasis parametro θ įvertis, turintis dispersiją, tai

$$\frac{1}{I(\theta) D_\theta \theta^*}$$

paprastai vadinamas įverčio θ^* *efektyvumu*.

Tenka pastebėti, kad ne visada egzistuoja efektyvūs įverčiai čia nurodyta prasme. Kaip matėme, kai pasiskirstymas yra normalusis $N(a, \sigma^2)$ su nežinomais a, σ^2, nepaslinktasis parametro σ^2 įvertis S_1^2 turi dispersiją $2\sigma^4/(n-1)$. Galima būtų įrodyti, kad ji yra mažiausia galima. Tuo tarpu $1/I(\sigma^2) = 2\sigma^4/n$. Vadinasi, minimali pasiekiama reguliariųjų įverčių dispersija gali būti didesnė už Rao–Kramero nelygybėje nurodytą apatinį rėžį.

Šią teoriją galima apibendrinti ir tuo atveju, kai pasiskirstymų klasė priklauso nuo kelių realiųjų parametrų.

8. PASIKLIAUTINIEJI INTERVALAI

Jei nežinomam parametrui θ turime "gerą" įvertį $\theta^*(X)$ ir $x = (x_1, ..., x_n)$ yra imties realizacija, tai galime laikyti $\theta \approx \theta^*(x_1, ..., x_n)$. Tačiau toks parametro įvertinimo būdas turi didelių trūkumų – juk $\theta^*(X_1, ..., X_n)$ yra atsitiktinis dydis. Kad ir koks "geras" būtų įvertis θ^*, jo reikšmės yra išsibarsčiusios apie θ. Jei tas dydis yra tolydus, tai tikimybė jam įgyti konkrečią reikšmę lygi 0. Todėl, remdamiesi įverčiais, galime rasti ne pačią nežinomo parametro reikšmę, o tik sritį, kuriai su tam tikra tikimybe priklauso vertinamasis parametras. Jei ta tikimybė artima 1, tai praktiškai galime laikyti, jog parametras yra toje srityje.

Sakykime, reikia įvertinti nežinomą parametrą θ. Imkime dvi statistikas $\theta_1^*(X_1, ..., X_n) < \theta_2^*(X_1, ..., X_n)$. Pažymėkime

$$\alpha = \mathcal{P}_\theta\{\theta_1^*(X_1, ..., X_n) < \theta < \theta_2^*(X_1, ..., X_n)\}.$$

Jei α mažai skiriasi nuo 1, tai galime laikyti, kad praktiškai

$$\theta_1^* < \theta < \theta_2^*.$$

Intervalas (θ_1^*, θ_2^*) yra vadinamas parametro θ *pasikliautinuoju intervalu*, o α – *pasikliovimo tikimybe*, arba *pasikliovimo lygmeniu*. $1-\alpha$ yra klaidos tikimybė. Pasikliovimo tikimybė paprastai imama 0,9; 0,95; 0,99 ir pan. Šias sąvokas pasiūlė Dž. Neimanas[1].

Statistikas θ_1^* ir θ_2^* galima parinkti įvairiausiais būdais. Reikia stengtis, kad tai pačiai pasikliovimo tikimybei pasikliautinasis intervalas būtų kuo trumpesnis.

Sudarysime normaliojo pasiskirstymo $N(a, \sigma^2)$ parametrų pasikliautinuosius intervalus. Teks skirti atvejus, kai vienas iš parametrų a, σ^2 yra žinomas, o kitas – nežinomas, ir kai abu parametrai nežinomi.

1. Tarkime, kad ž i n o m a s σ^2, o n e ž i n o m a s $a \in R$. Tiesa, toks atvejis turi tik teorinę reikšmę: paprastai nežinome abiejų parametrų arba žinome a, bet nežinome σ^2. Matėme, kad $\bar{X}$ yra efektyvusis a įvertis. Juo ir remsimės, sudarydami pasikliautinuosius intervalus. Imkime intervalą $(\bar{X} - -\delta,\ \bar{X}+\delta)$; čia δ – teigiamas skaičius, kurį vėliau parinksime. Apskaičiuosime pasikliovimo tikimybę

$$\alpha = \mathcal{P}_a(\bar{X} - \delta < a < \bar{X} + \delta) = \mathcal{P}_a\Big(- \frac{\delta\sqrt{n}}{\sigma} < \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - a) < \frac{\delta\sqrt{n}}{\sigma}\Big).$$

Kadangi X_k yra nepriklausomi ir kiekvienas iš jų pasiskirstęs pagal $N(a, \sigma^2)$, tai suma

$$\frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - a) = (\bar{X} - a)\frac{\sqrt{n}}{\sigma}$$

yra pasiskirsčiusi pagal dėsnį $N(0,1)$. Parinkę $\delta = \sigma u/\sqrt{n}$, gauname

$$\alpha = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1 = \sqrt{\frac{2}{\pi}} \int_0^u e^{-y^2/2} dy. \tag{1}$$

Pasikliovimo tikimybę α parenkame, vadovaudamiesi praktiniais sumetimais. Dažnai tai bus ekonominiai samprotavimai. Turėdami α, iš (1) galime rasti $u = u(\alpha)$. Paprastai tam praverčia normaliojo pasiskirstymo lentelės. Žinodami u, randame pasikliautinąjį intervalą

$$\Big(\bar{X} - \frac{\sigma u}{\sqrt{n}},\ \bar{X} + \frac{\sigma u}{\sqrt{n}}\Big).$$

Kuo didesnis α, tuo didesnis ir $u(\alpha)$, ir atvirkščiai. Sakysime (žr., pvz., [17], 1 lentelę),

[1] Jerzy Neyman (1895–1981) – amerikiečių matematikas.

$$u(0,9) \approx 1,64; u(0,95) \approx 1,96; u(0,99) \approx 2,58; u(0,999) \approx 3,29.$$

Paėmę didesnį α, turėsime mažesnę klaidos tikimybę $1-\alpha$, tačiau tada ir $u(\alpha)$ bus didesnis bei platesnis pasikliautinasis intervalas, vadinasi, a įvertinsime ne taip tiksliai. Pasikliautinąjį intervalą galime susiaurinti, padidindami n (jei tai leidžia eksperimento sąlygos). Jei iš anksto parenkame α ir u, tai galime rasti n.

2. Tarkime, kad a yra ž i n o m a s, o $\sigma^2 > 0$ – n e ž i n o m a s. 7.4 pavyzdyje parodėme, kad S_0 yra efektyvusis σ^2 įvertis. Nagrinėkime intervalą $(v_1 S_0^2, v_2 S_0^2)$, kai $0 < v_1 < v_2$. Jį atitinka pasikliovimo tikimybė

$$\alpha = \mathcal{P}_{\sigma^2}(v_1 S_0^2 < \sigma^2 < v_2 S_0^2) = \mathcal{P}_{\sigma^2}\left(\frac{n}{v_2} < \sum_{k=1}^{n}\left(\frac{X_k - a}{\sigma}\right)^2 < \frac{n}{v_1}\right).$$

Iš III.9.3 pavyzdžio žinome, kad suma

$$\sum_{k=1}^{n}\left(\frac{X_k - a}{\sigma}\right)^2$$

yra pasiskirsčiusi pagal χ^2 su n laisvės laipsnių dėsnį. Todėl, parinkę $v_1 = n/u_2$, $v_2 = n/u_1$, kai $0 < u_1 < u_2$, turime

$$\alpha = P(u_1 < \chi_n^2 < u_2) = \int_{u_1}^{u_2} p_{\chi_n^2}(y)dy.$$

Jei pasikliovimo tikimybė α yra duota, tai u_1 ir u_2 reikia taip parinkti, kad jie tenkintų tą lygybę. Aišku, tai galime padaryti be galo daug būdų. Paprastai imama

$$\begin{aligned} \int_0^{u_1} p_{\chi_n^2}(y)dy &= \frac{1-\alpha}{2}, \\ \int_{u_2}^{\infty} p_{\chi_n^2}(y)dy &= \frac{1-\alpha}{2}. \end{aligned} \tag{2}$$

Radę $u_1 = u_1(\alpha)$, $u_2 = u_2(\alpha)$, gauname σ^2 pasikliautinąjį intervalą

$$\left(nu_2^{-1}(\alpha)S_0^2,\ nu_1^{-1}(\alpha)S_0^2\right).$$

Spręsdami (2) lygtis, naudojamės χ_n^2 pasiskirstymo lentelėmis. Dideliems n tokių lentelių nėra. Jų ir neverta sudarinėti, nes atitinkamai normuotas χ_n^2 pasiskirstymas konverguoja į normalųjį pasiskirstymą, kai n neaprėžtai didėja. Įrodysime šį faktą. Kadangi χ_n^2 yra n normaliųjų dydžių, pasiskirsčiusių pagal $N(0,1)$, kvadratų suma, tai jo vidurkis yra n, o dispersija $2n$ (įrodykite!). Iš centrinės ribinės teoremos (III.11.2 teoremos 2 išvados) išplaukia, kad

$$\frac{\chi_n^2 - n}{\sqrt{2n}}$$

yra asimptotiškai pasiskirstęs pagal $N(0,1)$. Pasirodo, kad jau gana nedideliems n liekamasis narys yra mažas. Yra sudarytos χ_n^2 kvantilių ir jų aproksimacijų skirtumų lentelės (žr. [17], III$^{\text{b}}$ lentelę).

3. Tirsime atvejį, kai abu parametrai $a \in R$ ir $\sigma^2 > 0$ yra n e ž i n o m i. Sudarydami pasikliautinuosius intervalus, jau negalime naudoti statistikų

$$\frac{\sqrt{n}(\bar{X} - a)}{\sigma},\ S_0^2,$$

nes a ir σ^2 yra nežinomi. Imsime statistikas

$$\frac{\sqrt{n}(\bar{X} - a)}{S_1},\ (n-1)\left(\frac{S_1}{\sigma}\right)^2 = n\left(\frac{S}{\sigma}\right)^2.$$

Laikysime, kad $n > 1$. Rasime jų pasiskirstymus. Mums reikės daugiamačių normaliųjų atsitiktinių dydžių savybių.

Lema. *Jei stebimasis atsitiktinis dydis yra pasiskirstęs pagal normalųjį dėsnį $N(a, \sigma^2)$, tai statistikos $\bar{X}$ ir S^2 yra nepriklausomos, be to, statistika*

$$(\bar{X} - a)\frac{\sqrt{n}}{\sigma} = \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - a)$$

yra pasiskirsčiusi pagal $N(0,1)$, statistika

$$(n-1)\left(\frac{S_1}{\sigma}\right)^2 = \sum_{k=1}^{n}\left(\frac{X_k - \bar{X}}{\sigma}\right)^2$$

– pagal χ^2 su $n-1$ laisvės laipsnių, o statistika

$$\frac{(\bar{X} - a)\sqrt{n}}{S_1}$$

– pagal Stjudento dėsnį su $n-1$ laisvės laipsnių.

Į r o d y m a s. Atsitiktinio vektoriaus $X = (X_1, ..., X_n)$ charakteristinė funkcija pagal III.12 skyrelį

$$f_X(t) = \exp\{i(a, ..., a)t' - \frac{1}{2}\sigma^2 tt'\};$$

čia $t = (t_1, ..., t_n)$, o brūkšnelis reiškia transponavimą. Imkime ortogonalią matricą

$$C = \left\| \begin{matrix} c_{11} & \dots & c_{1n} \\ \dots & \dots & \dots \\ c_{n1} & \dots & c_{nn} \end{matrix} \right\|,$$

kurioje $c_{11} = ... = c_{n1} = n^{-1/2}$. Kadangi matrica yra ortogonali, tai

$$\sum_{k=1}^{n} c_{kj}c_{kl} = \begin{cases} 1, & \text{kai } j = l, \\ 0, & \text{kai } j \neq l. \end{cases}$$

Iš čia

$$\sum_{k=1}^{n} c_{kj} = 0 \quad (j = 2, ..., n).$$

Nagrinėsime vektorių $Y = (Y_1, ..., Y_n) = XC$. Aišku,

$$X_1^2 + ... + X_n^2 = Y_1^2 + ... + Y_n^2. \tag{3}$$

Parodysime, kad atsitiktiniai dydžiai $Y_1, ..., Y_n$ yra nepriklausomi. Vektoriaus Y charakteristinė funkcija

$$f_Y(t) = f_X(tC') = \exp\left\{ i(a, ..., a)Ct' - \frac{1}{2}\sigma^2 tC'Ct \right\} =$$
$$= \exp\left\{ iat_1\sqrt{n} - \frac{1}{2}\sigma^2 tt' \right\}.$$

Matome, kad atsitiktiniai dydžiai $Y_1, ..., Y_n$ yra normalieji ir kas du nekoreliuoti. Pagal III.13 skyrelio teoremą jie yra nepriklausomi. Be to,

$$Y_1 = (X_1 + ... + X_n)n^{-1/2} = \bar{X}\sqrt{n},$$
$$MY_1 = a\sqrt{n}, \ MY_k = 0 \ \ (k = 2, ..., n),$$
$$DY_k = \sigma^2 \ \ (k = 1, ..., n).$$

Iš (3) išplaukia, kad dydžiai

$$\sum_{k=1}^{n}(X_k - \bar{X})^2 = \sum_{k=1}^{n} X_k^2 - n\bar{X}^2 = \sum_{k=2}^{n} Y_k^2$$

ir Y_1 yra nepriklausomi. Lieka remtis III.9.3 pavyzdžiu.

Iš III.9.4 pavyzdžio išplaukia, kad statistika

$$\frac{(\bar{X} - a)\sqrt{n}}{S_1}$$

yra pasiskirsčiusi pagal Stjudento dėsnį su $n - 1$ laisvės laipsnių. □

Dabar jau galime sudaryti pasikliautinuosius intervalus nežinomiems parametrams a ir σ^2. Juos imame pavidalo

$$\Big(\bar X - \frac{u_1 S_1}{\sqrt{n}},\quad \bar X - \frac{u_2 S_1}{\sqrt{n}}\Big),\quad u_1 > u_2,$$
$$\Big(\frac{nS_1^2}{v_1},\quad \frac{nS_1^2}{v_2}\Big),\quad v_1 > v_2 > 0.$$

Atitinkamos pasikliovimo tikimybės lygios

$$\alpha' = \mathcal{P}_{(a,\sigma)}\left(u_2 < \frac{(\bar X - a)\sqrt{n}}{S_1} < u_1\right) = \int_{u_2}^{u_1} p_{St_{n-1}}(y)dy,$$
$$\alpha'' = \mathcal{P}_{(a,\sigma)}\left(v_2 < n\Big(\frac{S_1}{\sigma}\Big)^2 < v_1\right) = \int_{v_2}^{v_1} p_{\chi^2_{n-1}}(y)dy;$$

čia $p_{St_{n-1}}(y)$ yra Stjudento pasiskirstymo su $n-1$ laisvės laipsnių tankis. Pasirinkę α' ir α'', skaičius u_1, u_2, v_1, v_2 paprastai randame iš lygčių

$$u_2 = -u_1,\quad \frac{1-\alpha'}{2} = \int_{u_1}^{\infty} p_{St_{n-1}}(y)dy,$$
$$\frac{1-\alpha''}{2} = \int_0^{v_2} p_{\chi^2_{n-1}}(y)dy,$$
$$\frac{1-\alpha''}{2} = \int_{v_1}^{\infty} p_{\chi^2_{n-1}}(y)dy.$$

Palyginę antrąjį pasikliautinąjį intervalą dispersijai įvertinti, kai vidurkis a yra nežinomas, su pasikliautinuoju intervalu, kai a yra žinomas (2 atvejis), matome, kad abiem atvejais vartojamas χ^2 pasiskirstymas, tik antruoju atveju su n laisvės laipsnių, o pirmuoju – su $n-1$ laisvės laipsnių.

Ieškant pasikliautinųjų intervalų, tenka naudotis χ^2 ir Stjudento pasiskirstymų lentelėmis (žr. [17], III, IV, V lenteles), kurios yra sudarytos tik nedideliems n. Jau matėme, kad χ^2 pasiskirstymą dideliems n galima aproksimuoti normaliuoju pasiskirstymu. Taip yra ir su Stjudento pasiskirstymu. Šio dėsnio su n laisvės laipsnių tankio funkcija (žr. III.9.4 pavyzdį) yra

$$p_{St_n}(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\sqrt{2\pi}}\Big(1 + \frac{y^2}{n}\Big)^{-(n+1)/2}.$$

Pagal Stirlingo formulę pirmasis dauginamasis konverguoja į 1, kai $n \to \infty$. Bet kuriam fiksuotam y

$$-\frac{n+1}{2}\ln\Big(1+\frac{y^2}{n}\Big)\to-\frac{y^2}{2},$$

vadinasi,

$$p_{St_n}(y)\to\frac{1}{\sqrt{2\pi}}e^{-y^2/2}.$$

Toliau

$$(1+y^2/n)^{(n+1)/2}\ge(1+y^2/n)^{[(n+1)/2]}\ge1+[(n+1)/2]y^2/n\ge1+y^2/2.$$

Todėl funkcija $p_{St_n}(y)$ yra mažoruojama funkcijos

$$C(1+y^2/2)^{-1}.$$

Pasirėmę V.9.16 teorema, gauname

$$\int_{-\infty}^{u}p_{St_n}(y)dy\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-y^2/2}dy=\Phi(u),$$

kai $n\to\infty$.

Šį teiginį galima ir kitaip įrodyti. III.9.4 pavyzdyje atsitiktinis dydis, pasiskirstęs pagal Stjudento dėsnį, buvo nusakytas kaip dviejų nepriklausomų atsitiktinių dydžių santykis. Skaitiklyje buvo atsitiktinis dydis, pasiskirstęs pagal $N(0,1)$, o vardiklyje – kvadratinė šaknis iš vienodai pasiskirsčiusių nepriklausomų atsitiktinių dydžių aritmetinio vidurkio. Iš Chinčino teoremos išplaukia, kad vardiklyje esančio atsitiktinio dydžio pasiskirstymas konverguoja į vienetinį pasiskirstymą $\varepsilon(y)$. Iš čia išplaukia, kad minėtųjų atsitiktinių dydžių santykis yra pasiskirstęs pagal $N(0,1)$ (įrodykite!).

4. Iki šiol nagrinėjome normaliojo pasiskirstymo parametrų įvertinimą. Analogiškas teorijas galima sukurti ir kai kuriems kitiems pasiskirstymams. Tačiau, kai eksperimento sąlygos leidžia turėti daug stebėjimo duomenų, dažnai jų pasiskirstymas mažai skiriasi nuo normaliojo.

Tarkime, kad stebimasis atsitiktinis dydis turi nežinomą vidurkį $a\in R$ ir žinomą dispersiją σ^2. Iš centrinės ribinės teoremos (III.11.2 teoremos 2 išvada) išplaukia, kad normuotos sumos

$$(\bar X-a)\sqrt{n}/\sigma$$

pasiskirstymas konverguoja į standartinį normalųjį pasiskirstymą. Vadinasi,

$$\alpha=\mathcal{P}_a\big(-u<(\bar X-a)\sqrt{n}/\sigma<u\big)\approx\sqrt{\frac{2}{\pi}}\int_0^u e^{-y^2/2}dy,$$

kai n yra pakankamai didelis. Pasirinkę pasikliovimo tikimybę α, iš tos apytikslės lygybės apskaičiuojame apytikslę $u=u(\alpha)$ reikšmę ir gauname nežinomo vidurkio pasikliautinąjį intervalą

$$\Big(\bar X - \frac{\sigma u}{\sqrt{n}},\ \bar X + \frac{\sigma u}{\sqrt{n}}\Big).$$

Jei dispersija $\sigma^2 > 0$ yra nežinoma, tai šiuose įvertinimuose σ^2 reikia pakeisti S_1^2 arba S^2. Galima įrodyti, kad dideliems n statistika $(\bar X - a)\sqrt{n}/S_1$ yra apytiksliai pasiskirsčiusi pagal $N(0,1)$ (priminsime, kad ir Stjudento pasiskirstymas, kai laisvės laipsnių skaičius yra didelis, mažai skiriasi nuo $N(0,1)$). Pasirinkę α, randame $u = u(\alpha)$ iš apytikslės lygybės

$$\alpha = \mathcal{P}_{(a,\sigma)}\big(-u < (\bar X - a)\sqrt{n}/S_1 < u\big) \approx \sqrt{\frac{2}{\pi}} \int_0^u e^{-y^2/2} dy$$

ir sudarome nežinomo vidurkio pasikliautinąjį intervalą

$$\Big(\bar X - \frac{S_1 u}{\sqrt{n}},\ \bar X + \frac{S_1 u}{\sqrt{n}}\Big).$$

5. Baigdami paminėsime dar vieną praktiškai svarbų pasikliautinųjų intervalų sudarymo metodą. Tarkime, kad nežinomam parametrui θ įvertinti naudojamės statistika $T(X)$, gauta didžiausio tikėtinumo metodu. Kai tenkinamos gana bendros sąlygos, atsitiktinis dydis $\big(T(X) - \theta\big)/D_\theta^{1/2} T(X)$ yra asimptotiškai pasiskirstęs pagal normalųjį dėsnį $N(0,1)$. Paėmę α, iš apytikslės lygties

$$\alpha = \mathcal{P}_\theta\big(-u < \frac{T(X) - \theta}{\sqrt{D_\theta T(X)}} < u\big) \approx \sqrt{\frac{2}{n}} \int_0^u e^{-y^2/2} dy$$

randame $u = u(\alpha)$ ir sudarome parametro θ pasikliautinąjį intervalą

$$\big(T(X) - u(\alpha)\sqrt{D_\theta T(X)},\ T(X) + u(\alpha)\sqrt{D_\theta T(X)}\big).$$

P a v y z d y s. Fabriko cechas gamina kokias nors detales. Iš jo vienos dienos produkcijos atsitiktinai parenkame detalę, ją pasveriame ir grąžiname atgal. Po 24 svėrimų gavome šitokius rezultatus (sakysime, gramais):

2,1	2,3	1,9	2,0	2,1	2,2	1,8	1,7	2,1	2,2	1,9	2,0
1,9	2,1	1,8	2,0	1,9	2,2	2,1	2,3	1,9	2,1	2,2	2,0

Laikydami, kad detalių svoris pasiskirstęs pagal normalųjį dėsnį, įvertinsime jo vidurkį ir dispersiją.

Turime

$$\bar x \approx 2,0333,\ s^2 \approx 0,0247,\ s_1^2 \approx 0,0258.$$

Paėmę $\alpha' = \alpha'' = 0,98$, iš Stjudento pasiskirstymo lentelių (su 23 laisvės laipsniais) randame

$$u_1 = -u_2 \approx 2,4999,$$

o iš χ^2 lentelių (su 23 laisvės laipsniais)

$$v_1 \approx 41,638,\ v_2 \approx 10,196.$$

Vidurkio pasikliautinojo intervalo galai yra

$$\bar{x} - \frac{u_1 s_1}{\sqrt{23}} \approx 1,9496,\quad \bar{x} + \frac{u_1 s_1}{\sqrt{23}} \approx 2,1170,$$

o dispersijos

$$\frac{24s^2}{v_1} \approx 0,0142,\quad \frac{24s^2}{v_2} \approx 0,0581.$$

9. STATISTINIŲ HIPOTEZIŲ TIKRINIMAS

Šį uždavinį jau formulavome 1 skyrelyje. Priminsime, kad statistinėmis vadiname hipotezes apie stebimojo atsitiktinio dydžio arba kelių dydžių pasiskirstymą. Statistinės bus, pavyzdžiui, hipotezės: stebimasis atsitiktinis dydis yra pasiskirstęs pagal normalųjį dėsnį, dviejų atsitiktinių dydžių vidurkiai yra lygūs, vieno atsitiktinio dydžio dispersija yra didesnė negu kito.

Su tokiais uždaviniais dažnai susiduriame praktikoje. Sakysime, kokiai nors ligai gydyti yra vartojami žinomi vaistai. Rasti nauji vaistai. Ar jie bus efektyvesni už senuosius? Fabrikas gamina elektros lemputes. Siūloma pakeisti technologinį procesą. Ar nauju būdu pagamintos lemputės degs ilgiau?

Sakykime, turime pasiskirstymų klasę $\{P_\theta,\ \theta \in \Theta\}$ ir Θ_0 yra aibės Θ poaibis. Statistinė hipotezė bus teiginys, kad stebimojo atsitiktinio dydžio pasiskirstymas P_θ priklauso klasei $\{P_\theta,\ \theta \in \Theta_0\}$, kitaip tariant $\theta \in \Theta_0$. Tikrinamoji hipotezė paprastai vadinama *pagrindine*, arba *nuline*, ir žymima H_0. Kartu nagrinėjama ir priešinga jai hipotezė H_1. Ji vadinama *konkuruojančia*, arba *alternatyviąja*, hipoteze, arba tiesiog *alternatyva*. Tai bus teiginys, kad $\theta \in \Theta_1 = \Theta \backslash \Theta_0$. Jei aibė Θ_0 yra sudaryta iš vieno taško, tai hipotezė vadinama *paprastąja*. Priešingu atveju ji vadinama *sudėtingąja*.

Jei, pavyzdžiui, turime klasę normaliųjų pasiskirstymų su žinomomis dispersijomis, bet nežinomais vidurkiais $a \in R$, tai teiginys, kad stebimojo dydžio vidurkis $a = 5$, yra paprastoji hipotezė, o jos alternatyva yra teiginys $a \neq 5$. Teiginys $a > 5$ yra sudėtingoji hipotezė.

Matematinė statistika nagrinėja kriterijus arba testus, kurie leistų spręsti, ar stebėjimo duomenys suderinami su nuline hipoteze.

Kriterijų sudarymo idėja šitokia. Erdvę R^n suskaidome į dvi sritis – Borelio aibes: R_0 ir $R_1 = R^n \backslash R_0$. Sritį R_1 parenkame taip, kad tikimybės imčiai $X = (X_1, ..., X_n)$ patekti į tą sritį, kai hipotezė H_0 yra teisinga,

$$\mathcal{P}_\theta(X \in R_1),\ \theta \in \Theta_0,$$

būtų mažos. Susitariame, kad hipotezė H_0 nesuderinama su stebėjimo duomenimis ir todėl yra atmestina, kai $x \in R_1$. Ši sritis paprastai vadinama *kritine*.

Jei $x \in R_0$, tai laikome hipotezę H_0 suderinama su stebėjimo duomenimis ir todėl priimtina.

Ši procedūra negarantuoja, kad padarytoji išvada bus visada teisinga. Mes galime tik teigti, kad ji bus teisinga su tam tikra tikimybe. Jei tikimybės $\mathcal{P}_\theta(X \in R_1),\ \theta \in \Theta_0$, yra mažos, tai praktiškai labai retai atmesime hipotezę H_0, kai ji teisinga. Sakysime, jei tos tikimybės ne didesnės kaip 0,01, tai vidutiniškai ne daugiau kaip vieną kartą iš 100 atmesime teisingą hipotezę.

Naudodamiesi statistiniu kriterijumi, galime padaryti dviejų rūšių klaidas: atmesti, kaip jau minėjome, hipotezę H_0, kai ji yra teisinga (*pirmosios rūšies klaida*), arba priimti H_0, nors ji klaidinga (*antrosios rūšies klaida*). Galimi atvejai nurodyti lentelėje.

	Išvada	
	Priimta H_0	Atmesta H_0
H_0 teisinga	Teisingai	I rūšies klaida
H_0 klaidinga	II rūšies klaida	Teisingai

Geriausia būtų kriterijus sudaryti taip, kad abiejų rūšių klaidų tikimybės būtų minimalios. Deja, kai stebėjimų skaičius yra ribotas, to padaryti negalima. Praplėtus kritinę sritį, antrosios rūšies klaidos tikimybė sumažės, bet padidės pirmosios rūšies klaidos tikimybė, ir atvirkščiai. Todėl paprastai parenkamas rėžis, kurio neturėtų viršyti pirmosios rūšies klaidos tikimybė, ir stengiamasi minimizuoti antrosios rūšies klaidos tikimybę. Kitaip tariant, parenkamas nedidelis skaičius $\alpha \in (0,1)$, vadinamas *reikšmingumo lygmeniu*, ir reikalaujama, kad pirmosios rūšies klaidos tikimybė būtų ne didesnė už tą skaičių

$$\sup_{\theta \in \Theta_0} \mathcal{P}_\theta(X \in R_1) \le \alpha,$$

o antrosios rūšies klaidos tikimybė

$$\mathcal{P}_\theta(X \in R_0),\ \theta \in \Theta_1,$$

būtų kiek galima mažesnė, t. y. tikimybė

$$\beta(\theta) = \mathcal{P}_\theta(X \in R_1) = 1 - \mathcal{P}_\theta(X \in R_0),\ \theta \in \Theta_1,$$

vadinama *kriterijaus galia*, kiek galima didesnė. Ši tikimybė, traktuojama kaip θ funkcija visiems $\theta \in \Theta$, yra vadinama *kriterijaus galios funkcija*. $\beta(\theta)$ reiškia tikimybę atmesti hipotezę H_0, kai tikroji parametro reikšmė yra θ.

Tarkime, kad turime du kriterijus su kritinėmis sritimis R_1 ir R_1', turinčius tą patį reikšmingumo lygmenį. Kriterijus su kritine sritimi R_1 vadinamas *tolygiai galingesniu* už kriterijų su kritine sritimi R_1', jei

$$\mathcal{P}_\theta(X \in R_1) \le \mathcal{P}_\theta(X \in R_1'), \quad \theta \in \Theta_0,$$
$$\mathcal{P}_\theta(X \in R_1) \ge \mathcal{P}_\theta(X \in R_1'), \quad \theta \in \Theta_1.$$

Kriterijų su kritine sritimi R_1 ir reikšmingumo lygmeniu α vadiname *tolygiai galingiausiu*, jei

$$\sup_{R_1'} \mathcal{P}_\theta(X \in R_1') = \mathcal{P}_\theta(X \in R_1), \quad \theta \in \Theta_1;$$

čia supremumas imamas pagal visas kritines sritis R_1' su reikšmingumo lygmeniu α. Kai alternatyva yra paprastoji (Θ_1 turi tik vieną elementą), tolygiai galingiausią kriterijų vadiname tiesiog galingiausiu.

10. FUNDAMENTALIOJI NEIMANO–PIRSONO LEMA

Praeitame skyrelyje nagrinėtus statistinius kriterijus galime aprašyti šitokiu būdu. Kiekvienam $x \in R^n$ apibrėžiame funkciją $\psi(x)$, įgyjančią dvi reikšmes: 0 ir 1, ir susitariame tikrinamąją hipotezę atmesti, kai imčiai x ta funkcija įgyja reikšmę 1, arba priimti, kai ji įgyja reikšmę 0. Jei $\psi(x)$ yra Borelio funkcija, tai ji nusako erdvėje R^n dvi sritis: kritinę sritį $R_1 = \{x : \psi(x) = 1\}$ ir jos papildomąją sritį $R_0 = \{x : \psi(x) = 0\}$.

Praplėsime funkcijų ψ klasę, nagrinėdami Borelio funkcijas, apibrėžtas erdvėje R^n ir galinčias įgyti reikšmes iš intervalo $[0, 1]$. Paėmę tokią funkciją, sudarysime naujo tipo kriterijų. Kai imčiai x turime $\psi(x) = 1$ arba $\psi(x) = 0$, tai, kaip ir anksčiau, hipotezę H_0 atmetame arba priimame. Jei $0 < \psi(x) < 1$, tai hipotezę H_0 su tikimybe $\psi(x)$ atmetame ir su tikimybe $1-\psi(x)$ priimame. Tą procedūrą galima interpretuoti šitaip: kiekvienai imčiai x atliekame atsitiktinį eksperimentą su dviem galimais rezultatais, kurių tikimybės yra $\psi(x)$ ir $1 - \psi(x)$. Gavę pirmąjį rezultatą, hipotezę atmetame, gavę antrąjį – priimame.

Tokius kriterijus vadiname *randomizuotais* (nuo anglų kalbos žodžio *random* – atsitiktinis). 9 skyrelyje nagrinėti kriterijai vadinami *nerandomizuotais* arba *determinuotais*.

Funkciją

$$\beta(\theta) = \beta(\theta, \psi) = M_\theta \psi(X), \quad \theta \in \Theta,$$

vadinsime randomizuoto kriterijaus *galios funkcija*. Uždavinys yra rasti tokias funkcijas ψ, kad galia būtų didžiausia visiems $\theta \in \Theta_1$, kai $M_\theta \psi(X) \le \alpha$ visiems $\theta \in \Theta_0$.

Toliau nagrinėsime tik atvejį, kai hipotezė ir jos alternatyva yra paprastosios, t. y. $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. Pažymėkime K_α klasę funkcijų ψ, kurioms pirmosios rūšies klaidos tikimybė yra ne didesnė už α:

$$K_\alpha = \{\psi : M_{\theta_0}\psi(X) \le \alpha\}.$$

Kriterijų su funkcija $\psi^* \in K_\alpha$ vadinsime *galingiausiu*, jei

$$\beta(\theta_1, \psi^*) = \sup_{\psi \in K_\alpha} \beta(\theta_1, \psi),$$

kitaip tariant, jei antrosios rūšies klaidos tikimybė tenkina sąlygą

$$M_{\theta_1}\big(1 - \psi^*(X)\big) = \inf_{\psi \in K_\alpha} M_{\theta_1}\big(1 - \psi(X)\big).$$

Laikysime, kad pasiskirstymus $\mathcal{P}_{\theta_k}(k = 0, 1)$ dominuoja matas μ su tankio funkcijomis $\mathbf{p}_{\theta_k}(x) = p_{\theta_k}(x_1)\ ...\ p_{\theta_k}(x_n)$ $(k = 0, 1)$.

Parodysime, kad galingiausias kriterijus egzistuoja ir rasime jo pavidalą.

Teorema (fundamentalioji Neimano–Pirsono lema). *Kiekvienam $\alpha \in (0, 1)$ galima rasti tokias konstantas $c = c_\alpha$ ir $h = h_\alpha$, kad funkcijai*

$$\psi^*(x) = \begin{cases} 1, & \text{kai } \mathbf{p}_{\theta_1}(x) > h\mathbf{p}_{\theta_0}(x), \\ c, & \text{kai } \mathbf{p}_{\theta_1}(x) = h\mathbf{p}_{\theta_0}(x), \\ 0, & \text{kai } \mathbf{p}_{\theta_1}(x) < h\mathbf{p}_{\theta_0}(x), \end{cases}$$

būtų teisinga lygybė

$$M_{\theta_0}\psi^*(X) = \alpha$$

ir ja pagrįstas kriterijus būtų galingiausias tarp klasės K_α funkcijas atitinkančių kriterijų.

Į r o d y m a s. 1. Nagrinėkime funkciją

$$w(h) = \mathcal{P}_{\theta_0}\big(\mathbf{p}_{\theta_1}(X) > h\mathbf{p}_{\theta_0}(X)\big) = \int_{\{x:\mathbf{p}_{\theta_1}(x) > h\mathbf{p}_{\theta_0}(x)\}} \mathbf{p}_{\theta_0}(x)\mu(dx).$$

Ši funkcija yra nedidėjanti, tolydi iš dešinės, $w(h) = 1$, kai $h < 0$, ir $w(h) \to 0$, kai $h \to \infty$.

Kai $\alpha \in (0, 1)$, pažymėkime h_α mažiausią h, kuriam

$$w(h) \le \alpha \le w(h - 0).$$

Jei $w(h_\alpha - 0) - w(h_\alpha) > 0$, imkime

$$c_\alpha = \frac{\alpha - w(h_\alpha)}{w(h_\alpha - 0) - w(h_\alpha)}.$$

Jei $w(h_\alpha - 0) - w(h_\alpha) = 0$, tai imkime $c_\alpha = 1$. Apskaičiuosime

$$M_{\theta_0}\psi^*(X) = \int_{R^n} \psi^*(x)\mathbf{p}_{\theta_0}(x)\mu(dx).$$

Integravimo sritį galime suskaidyti į tris viena kitos nedengiančias sritis $\{x : \mathbf{p}_{\theta_1}(x) > h_\alpha \mathbf{p}_{\theta_0}(x)\}$, $\{x : \mathbf{p}_{\theta_1}(x) = h_\alpha \mathbf{p}_{\theta_0}(x)\}$, $\{x : \mathbf{p}_{\theta_1}(x) < h_\alpha \mathbf{p}_{\theta_0}(x)\}$. Pirmojoje srityje integralas lygus $w(h_\alpha)$, antrojoje $c_\alpha\big(w(h_\alpha - 0) - w(h_\alpha)\big)$, trečiojoje 0. Taigi

$$M_{\theta_0}\psi^*(X) = w(h_\alpha) + c_\alpha\big(w(h_\alpha - 0) - w(h_\alpha)\big) = \alpha. \tag{1}$$

2. Tarkime, kad ψ yra bet kuri klasės K_α funkcija. Parodysime, kad

$$M_{\theta_1}\psi^*(X) - M_{\theta_1}\psi(X) = \int_{R^n} \big(\psi^*(x) - \psi(x)\big)\mathbf{p}_{\theta_1}(x)\mu(dx) \geq 0.$$

Imame integralą

$$\begin{aligned}&\int_{R^n} \big(\psi^*(x) - \psi(x)\big)\big(\mathbf{p}_{\theta_1}(x) - h_\alpha \mathbf{p}_{\theta_0}(x)\big)\mu(dx) = \\ &= \int_{\{x:\psi^*(x)>\psi(x)\}} + \int_{\{x:\psi^*(x)<\psi(x)\}}.\end{aligned}$$

Pirmajame integrale funkcija $\psi^*(x) > 0$, todėl pagal jos apibrėžimą $\mathbf{p}_{\theta_1}(x) - -h_\alpha \mathbf{p}_{\theta_0}(x) \geq 0$. Iš čia išplaukia, kad tas integralas yra neneigiamas. Visai taip pat antrajame integrale funkcija $\mathbf{p}_{\theta_1}(x) - h_\alpha \mathbf{p}_{\theta_0}(x) \leq 0$; todėl ir tas integralas yra neneigiamas. Iš čia

$$\begin{aligned}&\int_{R^n} \big(\psi^*(x) - \psi(x)\big)\mathbf{p}_{\theta_1}(x)\mu(dx) \geq \\ &\geq h_\alpha \int_{R^n} \big(\psi^*(x) - \psi(x)\big)\mathbf{p}_{\theta_0}(x)\mu(dx) = \\ &= h_\alpha\big(M_{\theta_0}\psi^*(X) - M_{\theta_0}\psi(X)\big).\end{aligned}$$

Remiantis klasės K_α apibrėžimu ir (1) lygybe, šis reiškinys yra neneigiamas. □

Įrodinėdami kriterijaus su funkcija ψ^* galingumą klasėje K_α, rėmėmės savybe, kad to kriterijaus pirmosios rūšies klaidos tikimybė yra α. Tarp randomizuotų kriterijų toks kriterijus visada egzistuoja. Jei apsiribotume nerandomizuotais kriterijais, tai ne kiekvienam α būtų galima rasti tokį h, kad funkciją

$$\hat{\psi}(x) = \begin{cases} 1, & \text{kai } \mathbf{p}_{\theta_1}(x) \geq h\mathbf{p}_{\theta_0}(x), \\ 0, & \text{kai } \mathbf{p}_{\theta_1}(x) < h\mathbf{p}_{\theta_0}(x), \end{cases} \tag{2}$$

atitinkančio kriterijaus pirmosios rūšies klaidos tikimybė būtų lygi α. Jei kuriam nors α tokį h galima rasti, tai kriterijus su funkcija $\hat{\psi}$ bus galingiausias klasėje K_α.

Jei funkcija $w(h)$ yra tolydi, tai galingiausias kriterijus bus nerandomizuotas visiems $\alpha \in (0,1)$; jį atitinkanti funkcija bus (2) pavidalo.

Jei kuriam nors α funkcija $w(h)$ taške h_α yra tolydi: $w(h_\alpha - 0) = w(h_\alpha)$, tai galingiausias kriterijus bus taip pat nerandomizuotas. Vadinasi, jei duotas kuris nors reikšmingumo lygmuo α, tai, pakeitę jį šiek tiek mažesniu (sakysime, skaičiumi $w(h_\alpha) - \varepsilon$), gausime nerandomizuotą kriterijų. Toks pakeitimas sumažina pirmosios rūšies klaidos tikimybę, bet padidina antrosios rūšies klaidos tikimybę. Tačiau taip keisti apsimoka, nes nerandomizuoti kriterijai yra paprastesni už randomizuotus.

Pastebėsime, kad galingiausias kriterijus yra pagrįstas funkcija

$$\frac{\mathbf{p}_{\theta_1}(x)}{\mathbf{p}_{\theta_0}(x)} = \frac{p_{\theta_1}(x_1)...p_{\theta_1}(x_n)}{p_{\theta_0}(x_1)...p_{\theta_0}(x_n)},$$

t. y. tikėtinumo funkcijų realizacijų santykiu. Dažnai patogu imti jos logaritmą

$$z = z(x) = \sum_{k=1}^{n} \ln \frac{p_{\theta_1}(x_k)}{p_{\theta_0}(x_k)}$$

ir funkciją ψ^* užrašyti šitaip:

$$\psi^*(x) = \begin{cases} 1, & \text{kai } z(x) > d_\alpha, \\ c_\alpha, & \text{kai } z(x) = d_\alpha, \\ 0, & \text{kai } z(x) < d_\alpha. \end{cases}$$

11. HIPOTEZIŲ APIE PASISKIRSTYMO PARAMETRUS TIKRINIMAS

Keliais pavyzdžiais parodysime, kaip sudaromi kriterijai statistinėms hipotezėms apie pasiskirstymo parametrus tikrinti. Remsimės 10 skyrelyje išdėstyta teorija.

1. Sakykime, turime normalųjį pasiskirstymą su n e ž i n o m u v i d u r k i u $a \in R$, bet ž i n o m a d i s p e r s i j a σ^2, ir reikia patikrinti hipotezę H_0, kad $a = a_0$, kai alternatyva yra $a = a_1 > a_0$ (nepainioti a_1 su empiriniu vidurkiu!). Tikėtinumo funkcijų realizacijų santykio logaritmo pagrindinis narys yra lygus

$$-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - a_1)^2 + \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - a_0)^2 =$$
$$= n(a_1 - a_0)\bar{x}/\sigma^2 + \frac{1}{2}n(a_1^2 - a_0^2)/\sigma^2.$$

Šiuo atveju funkcija $w(h)$ yra tolydi. Todėl galėsime sudaryti galingiausią nerandomizuotą kriterijų. Kritinė sritis bus pavidalo

$$(\bar{x} - a_0)\sqrt{n}/\sigma \geq u.$$

Jei hipotezė H_0 yra teisinga, tai statistika $(\bar{X} - a_0)\sqrt{n}/\sigma$, kaip žinome iš 8 skyrelio, yra pasiskirsčiusi pagal $N(0,1)$. Pirmosios rūšies klaidos tikimybė

$$(1) \qquad \alpha = \mathcal{P}_{a_0}\big((\bar{X} - a_0)\sqrt{n}/\sigma \geq u\big) = \frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-v^2/2} dv = 1 - \Phi(u).$$

Iš šios lygybės randame $u = u_\alpha$; jis yra standartinio normaliojo pasiskirstymo $(1-\alpha)$-kvantilis. Kriterijaus galia

$$\beta(a_1) = \mathcal{P}_{a_1}\big((\bar{X} - a_0)\sqrt{n}/\sigma \geq u_\alpha\big) = \mathcal{P}_{a_1}\big((\bar{X} - a_1)\sqrt{n}/\sigma \geq$$
$$\geq u_\alpha + (a_0 - a_1)\sqrt{n}/\sigma\big) = 1 - \Phi\big(u_\alpha + (a_0 - a_1)\sqrt{n}/\sigma\big).$$

Matome, kad tikimybė atmesti H_0 yra lygi kriterijaus reikšmingumo lygmeniui, kai $a = a_0$, ir monotoniškai artėja prie 1, kai $a_1 \to \infty$. Iš čia taip pat matome, kad, paėmę pakankamai didelį n, kriterijaus galią galime padaryti kiek norima artimą 1. Galios funkcijos grafikas pavaizduotas 34 paveiksle.

Kritinės srities pavidalas nepriklauso nuo alternatyvos $a = a_1$. Todėl gautasis kriterijus yra tolygiai galingiausias, kai alternatyva yra sudėtinga: $\Theta_1 = \{a : a > a_0\}$.

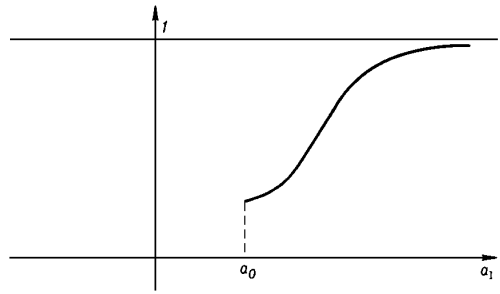
Jei alternatyva yra $a = a_1 < a_0$, tai hipotezę atmetame, kai konkreti imtis tenkina nelygybę

$$(\bar{x} - a_0)\sqrt{n}/\sigma \leq u.$$

Skaičių $u = u_\alpha$ randame iš lygties

$$(2) \qquad \alpha = \Phi(u);$$

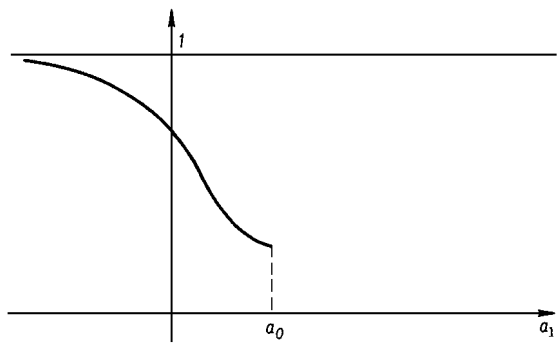
jis yra α-kvantilis. Kriterijaus galia



34 pav.

$$\beta(a_1) = \mathcal{P}_{a_1}\big((\bar{X} - a_0)\sqrt{n}/\sigma \le u\big) = \mathcal{P}_{a_1}\big((\bar{X} - a_1)\sqrt{n}/\sigma \le$$
$$\le u_\alpha + (a_0 - a_1)\sqrt{n}/\sigma\big) = \Phi\big(u_\alpha + (a_0 - a_1)\sqrt{n}/\sigma\big).$$

Galios funkcijos grafikas pavaizduotas 35 paveiksle.



35 pav.

Ir šis kriterijus yra tolygiai galingiausias, kai turime sudėtingą alternatyvą $\Theta_1 = \{a : a < a_0\}$.

Panagrinėkime alternatyvą $a = a_1 \neq a_0$. Hipotezę atmetame, kai konkreti imtis tenkina nelygybę

$$|\bar{x} - a_0|\sqrt{n}/\sigma \ge u.$$

Skaičių $u = u_\alpha$ rasime iš lygties

$$\alpha = \mathcal{P}_{a_0}\big(|\bar{X} - a_0|\sqrt{n}/\sigma \geq u\big) = \\ = \frac{1}{\sqrt{2\pi}} \int_{|y|\geq u} e^{-v^2/2} dv = \Phi(-u) + 1 - \Phi(u) = 2\big(1 - \Phi(u)\big); \tag{3}$$

jis yra $(1 - \alpha/2)$-kvantilis. Galios funkcija

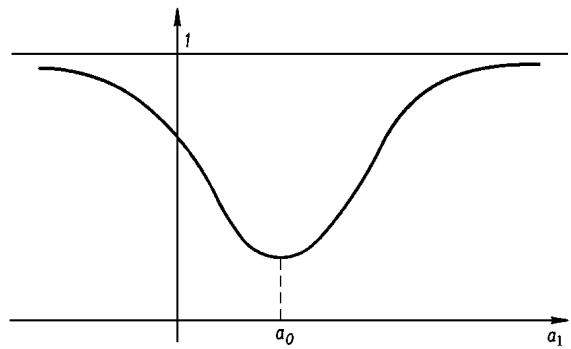
$$\beta(a_1) = \Phi\big(-u_\alpha + (a_0 - a_1)\sqrt{n}/\sigma\big) + 1 - \Phi\big(u_\alpha + (a_0 - a_1)\sqrt{n}/\sigma\big).$$

Jos grafikas pavaizduotas 36 paveiksle.

Šis kriterijus nėra tolygiai galingiausias sudėtingai alternatyvai $\Theta_1 = \{a : a \neq a_0\}$. Iš tikrųjų galingiausio paprastajai alternatyvai $a = a_1 \neq a_0$ kriterijaus kritinė sritis priklauso nuo a_1 (ji yra vienokia, kai $a_1 < a_0$, ir kitokia, kai $a_1 > a_0$).

Paskutinįjį iš nagrinėtų kriterijų galima būtų pavadinti dvipusiu, o pirmuosius du – vienpusiais: dešiniapusiu ir kairiapusiu.

Jei pasiskirstymas nėra normalusis, bet turi žinomą dispersiją σ^2, tai galime pasinaudoti statistikos $(\bar{X} - a)\sqrt{n}/\sigma$ asimptotiniu normalumu. Hipotezei $a = a_0$ su atitinkamomis alternatyvomis kriterijų kritinės sritys gali būti sudarytos taip pat, kaip ir turint normalųjį pasiskirstymą, tik (1), (2), (3) lygtis skaičiams u_α rasti reikėtų pakeisti apytikslėmis lygtimis, tuo tikslesnėmis, kuo didesnis n.



36 pav.

2. Tarkime, kad stebimasis atsitiktinis dydis įgyja dvi reikšmes: 1 ir 0 su tikimybėmis atitinkamai p ir $1 - p$. Tikimybė $p \in (0, 1)$ yra n e ž i-n o m a. Reikia patikrinti hipotezę $p = p_0$, kai alternatyva yra $p = p_1 < p_0$. Tikėtinumo funkcijų realizacijų santykis yra

$$\prod_{k=1}^{n} \Big(\frac{p_1}{p_0}\Big)^{x_k} \Big(\frac{1 - p_1}{1 - p_0}\Big)^{1 - x_k},$$

o jo logaritmas

$$\kappa_n \ln \frac{1/p_0 - 1}{1/p_1 - 1} + n \ln \frac{1 - p_1}{1 - p_0};$$

čia $\kappa_n = x_1 + ... + x_n$ yra sveikasis neneigiamas skaičius, o jo koeficientas – neigiamas. Todėl funkcija ψ^* yra šitokio pavidalo:

$$\psi^*(x) = \begin{cases} 1, & \text{kai } \kappa_n < m, \\ c, & \text{kai } \kappa_n = m, \\ 0, & \text{kai } \kappa_n > m. \end{cases}$$

Tarkime, kad reikšmingumo lygmuo yra α. Pirmosios rūšies klaidos tikimybė turi būti

$$\alpha = c\mathcal{P}_{p_0}(X_1 + ... + X_n = m) + \sum_{k<m} \mathcal{P}_{p_0}(X_1 + ... + X_n = k).$$

Iš čia reikia rasti $c = c_\alpha$ ir $m = m_\alpha$. Kadangi statistika $X_1 + ... + X_n$, kai teisinga nulinė hipotezė, yra pasiskirsčiusi pagal binominį dėsnį su $p = p_0$, tai

$$\mathcal{P}_{p_0}(X_1 + ... + X_n = k) = \binom{n}{k} p_0^k (1 - p_0)^{n-k}.$$

Jei egzistuoja toks sveikasis m, kad

$$\sum_{k=0}^{m} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = \alpha,$$

tai jį ir laikome m_α; tada c_α imame lygų 1. Gauname galingiausią nerandomizuotą kriterijų. Jei tokio sveikojo m nėra, tai parenkame sveiką m_α su sąlyga

$$\sum_{k=0}^{m_\alpha - 1} \binom{n}{k} p_0 (1 - p_0)^{n-k} < \alpha < \sum_{k=0}^{m_\alpha} \binom{n}{k} p_0 (1 - p_0)^{n-k}.$$

Tada

$$c_\alpha = \frac{\alpha - \sum_{k=0}^{m_\alpha - 1} \binom{n}{k} p_0^k (1 - p_0)^{n-k}}{\binom{n}{m_\alpha} p_0^{m_\alpha} (1 - p_0)^{n - m_\alpha}}.$$

Vėl gauname galingiausią, tačiau randomizuotą kriterijų.

Kritiškosios srities pavidalas nepriklauso nuo alternatyvos $p = p_1$. Todėl gautasis kriterijus yra tolygiai galingiausias, kai alternatyva yra sudėtinga: $\Theta_1 = \{p : p < p_0\}$.

Praktiškai abiem atvejais, kaip buvo rašyta 10 skyrelyje, vartojamas nerandomizuotas kriterijus su funkcija

$$\hat{\psi}(x) = \begin{cases} 1, & \text{kai } \kappa_n \le m, \\ 0, & \text{kai } \kappa_n > m. \end{cases}$$

Paliekame skaitytojui užrašyti galios funkciją bei išnagrinėti atvejus, kai alternatyva yra $p = p_1 > p_0$ arba $p = p_1 \neq p_0$.

1 p a v y z d y s. Laboratorijoje tiriami nauji vaistai kuriai nors ligai gydyti. Vaistų atradėjai teigia, kad jie efektyvūs 80% atvejų. Laboratorijoje vaistus tikrino 7 kartus, ir 4 kartus jie buvo efektyvūs. Patikrinsime, ar šie duomenys atitinka atradėjų teiginį, kai reikšmingumo lygmuo $\alpha = 0,1$.

Tikrinsime hipotezę $H_0 : p = 0,8$, kai alternatyva yra $H_1 : p = p_1 < 0,8$.

Nesunku suskaičiuoti, kad

$$\sum_{k=0}^{3} \binom{7}{k} 0,8^k \cdot 0,2^{7-k} \approx 0,033$$

ir

$$\binom{7}{4} 0,8^4 \cdot 0,2^3 \approx 0,115, \quad \sum_{k=0}^{4} \binom{7}{k} 0,8^k \cdot 0,2^{7-k} \approx 0,148.$$

Todėl $m_{0,1} = 4$ ir $c_{0,1} \approx 0,58$. Turime galingiausią randomizuotą kriterijų su funkcija

$$\psi^*(x) = \begin{cases} 1, & \text{kai } \kappa_7 < 4, \\ c_{0,1}, & \text{kai } \kappa_7 = 4, \\ 0, & \text{kai } \kappa_7 > 4, \end{cases}$$

ir nerandomizuotą kriterijų su funkcija

$$\hat{\psi}(x) = \begin{cases} 1, & \text{kai } \kappa_7 \le 4, \\ 0, & \text{kai } \kappa_7 > 4. \end{cases}$$

Jei pasirinksime nerandomizuotą kriterijų, tai hipotezę teks atmesti. Jei imsime randomizuotą kriterijų, tai hipotezę atmesime su tikimybe $c_{0,1} \approx 0,58$, o su tikimybe $1 - c_{0,1} \approx 0,42$ laikysime ją priimtina.

Skaičiavimai palengvės, pasinaudojus lygybe

$$\sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k} = 1 - I_p(m+1, n-m);$$

čia

$$I_p(m,n) = \frac{1}{B(m,n)} \int_0^p y^{m-1}(1-y)^{n-1} dy,$$

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

Yra sudarytos lentelės (žr. [17], XII lentelę; [2], 5.1, 5.2 lenteles). Vartojamos taip pat įvairios apytikslės formulės (Muavro–Laplaso, Puasono teoremos ir t. t.).

3. Stebime atsitiktinį dydį, pasiskirsčiusį pagal Puasono dėsnį su nežinomu parametru $\lambda > 0$. Reikia patikrinti hipotezę $\lambda = \lambda_0$, kai alternatyva yra $\lambda = \lambda_1 > \lambda_0$. Tikėtinumo funkcijų realizacijų santykio logaritmas yra

$$(x_1 + ... + x_n) \ln \frac{\lambda_1}{\lambda_0} - n(\lambda_1 - \lambda_0);$$

čia $x_1 + ... + x_n$ yra sveikasis neneigiamas skaičius, jo koeficientas yra teigiamas. Imame funkciją

$$\psi^*(x) = \begin{cases} 1, & \text{kai } x_1 + ... + x_n > m, \\ c, & \text{kai } x_1 + ... + x_n = m, \\ 0, & \text{kai } x_1 + ... + x_n < m. \end{cases}$$

Skaičius $c = c_\alpha$ ir $m = m_\alpha$ reikia rasti iš lygties

$$\alpha = c\mathcal{P}_{\lambda_0}(X_1 + ... + X_n = m) + \sum_{k>m} \mathcal{P}_{\lambda_0}(X_1 + ... + X_n = k).$$

Statistika $X_1 + ... + X_n$, kai teisinga nulinė hipotezė, yra pasiskirsčiusi pagal Puasono dėsnį su parametru $\lambda_0 n$. Todėl

$$\mathcal{P}_{\lambda_0}(X_1 + ... + X_n = k) = \frac{(\lambda_0 n)^k}{k!} e^{-\lambda_0 n}.$$

Jei egzistuoja sveikasis skaičius m, tenkinantis sąlygą

$$\sum_{k=m}^{\infty} \frac{(\lambda_0 n)^k}{k!} e^{-\lambda_0 n} = \alpha,$$

tai jį ir laikome m_α; tada $c_\alpha = 1$. Gauname galingiausią nerandomizuotą kriterijų. Jei tokio sveikojo m nėra, tai galime rasti sveikąjį m_α su sąlyga

$$\sum_{k=m_\alpha+1}^{\infty} \frac{(\lambda_0 n)^k}{k!} e^{-\lambda_0 n} < \alpha < \sum_{k=m_\alpha}^{\infty} \frac{(\lambda_0 n)^k}{k!} e^{-\lambda_0 n}.$$

Tada c_α imame lygų

$$c_\alpha = \frac{\alpha - \displaystyle\sum_{k=m_\alpha+1}^{\infty} \frac{(\lambda_0 n)^k}{k!} e^{-\lambda_0 n}}{\dfrac{(\lambda_0 n)^{m_\alpha}}{m_\alpha!} e^{-\lambda_0 n}}.$$

Gauname randomizuotą galingiausią kriterijų.

Iš čia kritinės srities pavidalas nepriklauso nuo alternatyvos $\lambda = \lambda_1 > \lambda_0$. Kriterijus yra tolygiai galingiausias, kai alternatyva yra sudėtinga: $\Theta_1 = \{\lambda : : \lambda > \lambda_0\}$.

Praktiškai abiem atvejais vartojamas nerandomizuotas kriterijus su funkcija

$$\hat{\psi}(x) = \begin{cases} 1, & \text{kai } x_1 + ... + x_n \geq m, \\ 0, & \text{kai } x_1 + ... + x_n < m. \end{cases}$$

Siūlome skaitytojui išnagrinėti galios funkciją bei sudaryti kriterijus, kai alternatyvos yra $\lambda = \lambda_1 < \lambda_0$ arba $\lambda = \lambda_1 \neq \lambda_0$.

Praktiniuose skaičiavimuose galima naudotis lygybe

$$\sum_{k=0}^{m} \frac{\lambda^k}{k!} e^{-\lambda} = P(\chi^2_{2m+2} > 2\lambda),$$

atitinkamomis lentelėmis (žr. [17], III lentelę; [2], 5.3 lentelę) bei įvairiomis apytikslėmis formulėmis.

4. Nagrinėsime normalųjį pasiskirstymą, kai a b u p a r a m e t r a i $a \in R$ ir $\sigma^2 > 0$ yra n e ž i n o m i. Tikrinsime hipotezę H_0, kad $a = a_0$. Statistika $(\bar{X} - a_0)\sqrt{n}/S_1$, kai teisinga H_0, yra pasiskirsčiusi pagal Stjudento dėsnį su $n-1$ laisvės laipsnių (žr. 8 skyrelio lemą).

Kai alternatyva yra $a = a_1 > a_0$, hipotezė atmetama, jei

$$(\bar{x} - a_0)\sqrt{n}/s_1 \geq u.$$

Tada

$$\alpha = \mathcal{P}_{a_0}\big((\bar{X} - a_0)\sqrt{n}/S_1 \geq u\big) = \int_u^\infty p_{S_{t_{n-1}}}(v)dv;$$

$u = u_\alpha$ yra Stjudento pasiskirstymo su $n-1$ laisvės laipsnių $(1-\alpha)$-kvantilis.

Jei alternatyva yra $a = a_1 < a_0$, tai hipotezę atmetame, kai

$$(\bar{x} - a_0)\sqrt{n}/s_1 \leq u;$$

$u = u_\alpha$ randame iš lygybės

$$\alpha = \int_{-\infty}^u p_{S_{t_{n-1}}}(v)dv;$$

u_α yra atitinkamo Stjudento pasiskirstymo α-kvantilis.

Pagaliau, kai alternatyva yra $a = a_1 \neq a_0$, hipotezė atmetama, jei

$$|\bar{x} - a_0|\sqrt{n}/s_1 \geq u,$$

ir $u = u_\alpha$ randamas iš lygties

$$\alpha = \int_{|v| \geq u} p_{S_{t_{n-1}}}(v)dv.$$

Jei pasiskirstymas yra bet koks, tačiau turi dispersiją σ^2, tai galima būtų parodyti, kad statistika $(\bar{X} - a_0)\sqrt{n}/S_1$ yra asimptotiškai pasiskirsčiusi pagal $N(0,1)$, kai hipotezė $H_0 : a = a_0$ yra teisinga. Tam pakanka pastebėti, kad statistika $(\bar{X} - a_0)\sqrt{n}/\sigma$, kai teisinga nulinė hipotezė, yra asimptotiškai pasiskirsčiusi pagal $N(0,1)$, o S_1 konverguoja pagal tikimybę į σ. Kritines sritis galime konstruoti kaip ir 1 uždavinyje. Skaičiai u randami iš apytikslių lygybių, analogiškų (1), (2), (3).

2 p a v y z d y s. Knygos pradžioje (I.2 skyrelyje) pasakojome, kad Biufonas metė monetą 4040 kartų, ir 2048 kartus atvirto herbas. Ar šie duomenys atitinka hipotezę, kad moneta yra simetriška, kai reikšmingumo lygmuo lygus 0,05?

Reikia patikrinti hipotezę, kad herbo atvirtimo tikimybė $p = 0,5$, kai alternatyva yra $p = p_1 \neq 0,5$. Statistika $(\bar{X} - 0,5)\sqrt{2n}$, kai teisinga nulinė hipotezė, yra asimptotiškai pasiskirsčiusi pagal $N(0,1)$. Kritinė sritis yra $|\bar{x} - 0,5|\sqrt{2n} \geq u$. Skaičių $u = u_{0,05}$ randame iš apytikslės lygties

$$0,05 \approx 2\big(1 - \Phi(u)\big).$$

Iš čia $\Phi(u) \approx 0,975$ ir $u_{0,05} \approx 1,960$. Vadinasi, kritinę sritį galime imti $|\bar{x} - -0,5| \geq 1,96(2n)^{-1/2} \approx 0,0218$. Iš Biufono duomenų $\bar{x} - 0,5 \approx 0,007$. Hipotezė priimtina.

5. Panagrinėkime hipotezių apie normaliojo pasiskirstymo dispersiją tikrinimą. Tarkime, kad v i d u r k i s a yra ž i n o m a s, o d i s p e r s i j a $\sigma^2 > 0$ – n e ž i n o m a, ir reikia patikrinti hipotezę $\sigma^2 = \sigma_0^2$, kai alternatyva yra $\sigma^2 = \sigma_1^2 > \sigma_0^2$. Remsimės Neimano–Pirsono fundamentaliąja lema. Tikėtinumo funkcijų realizacijų santykio logaritmo pagrindinis narys yra

$$-\frac{1}{2}\Big(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\Big)\sum_{k=1}^{n}(x_k - a)^2.$$

Todėl hipotezė atmetama, kai

$$ns_0^2/\sigma_0^2 \leq u.$$

Statistika nS_0^2/σ_0^2, kai teisinga nulinė hipotezė, yra pasiskirsčiusi pagal χ^2 su n laisvės laipsnių. Skaičių $u = u_\alpha$ randame iš lygties

$$\alpha = \int_0^u p_{\chi_n^2}(v)dv.$$

Jis yra to pasiskirstymo α-kvantilis.

Jei alternatyva yra $\sigma^2 = \sigma_1^2 < \sigma_0^2$, tai hipotezė atmetama, kai

$$ns_0^2/\sigma_0^2 \geq u$$

su $u = u_\alpha$, lygiu $(1 - \alpha)$-kvantiliui.

Alternatyvai $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$ nulinė hipotezė atmetama, kai

$$ns_0^2/\sigma_0^2 \leq u' \quad \text{arba} \quad ns_0^2/\sigma_0^2 \geq u'';$$

čia $u' = u'_\alpha$ yra $\alpha/2$-kvantilis, o $u'' = u''_\alpha$ yra $(1-\alpha/2)$-kvantilis.

Šių kriterijų galios atitinkamai lygios

$$1 - F_{\chi_n^2}(u_\alpha \sigma_0^2/\sigma_1^2), \quad F_{\chi_n^2}(u_\alpha \sigma_0^2/\sigma_1^2),$$
$$F_{\chi_n^2}(u'_\alpha \sigma_0^2/\sigma_1^2) + 1 - F_{\chi_n^2}(u''_\alpha \sigma_0^2/\sigma_1^2).$$

Visi trys yra galingiausi, kai atitinkamos alternatyvos paprastosios. Pirmieji du yra tolygiai galingiausi, kai alternatyvos sudėtingos.

Jei vidurkis $a \in R$ yra n e ž i n o m a s, tai vietoj statistikos nS_0^2/σ_0^2 imame statistiką $(n-1)S_1^2/\sigma_0^2$. Kai teisinga nulinė hipotezė, ši statistika yra pasiskirsčiusi pagal χ^2 su $n-1$ laisvės laipsnių. Kritinės sritys sudaromos analogiškai.

6. Dažnai tenka lyginti dviejų atsitiktinių dydžių nežinomus vidurkius. Tarkime, kad du stebimieji atsitiktiniai dydžiai yra pasiskirstę pagal $N(a_1, \sigma_1^2)$ ir $N(a_2, \sigma_2^2)$ su n e ž i n o m a i s v i d u r k i a i s $a_1 \in R, a_2 \in R$, bet ž i n o m o m i s d i s p e r s i j o m i s σ_1^2 ir σ_2^2. Stebime n kartų pirmąjį dydį ir m kartų – antrąjį. Gauname imtis $X = (X_1, ..., X_n)$ ir $Y = (Y_1, ..., Y_m)$. Imkime statistiką

$$\frac{\bar{X} - \bar{Y} - b_0}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}.$$

Jei teisinga hipotezė $a_1 - a_2 = b_0$, tai ta statistika pasiskirsčiusi pagal $N(0,1)$.

Tikriname hipotezę $a_1 - a_2 = b_0$ su alternatyva $a_1 - a_2 = b_1 > b_0$. Remdamiesi Neimano–Pirsono lema, gauname, kad hipotezę reikia atmesti, kai

$$\frac{\bar{x} - \bar{y} - b_0}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \geq u.$$

Jei alternatyva yra $a_1 - a_2 < b_0$, tai hipotezę atmetame, kai

$$\frac{\bar{x} - \bar{y} - b_0}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \leq u.$$

Pagaliau, jei alternatyva yra $a_1 - a_2 \neq b_0$, tai hipotezę atmetame, kai

$$\frac{|\bar{x} - \bar{y} - b_0|}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \geq u.$$

Lygtis skaičiams $u = u_\alpha$ rasti paliekame parašyti skaitytojui. Galima būtų įrodyti, kad pirmieji du kriterijai yra tolygiai galingiausi sudėtingoms dešiniapusei ir kairiapusei alternatyvoms.

Jei d i s p e r s i j o s σ_1^2 ir σ_2^2 n ė r a ž i n o m o s, bet žinomas jų santykis $\sigma_1^2/\sigma_2^2 = \lambda$, tai, sudarydami kritines sritis, galime naudotis statistika

$$\frac{\bar{X} - \bar{Y} - b_0}{\sqrt{(n-1)S_{11}^2 + \lambda(m-1)S_{12}^2}} \sqrt{\frac{\lambda nm(n+m-2)}{n+\lambda m}};$$

čia

$$S_{11}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2, \quad S_{12}^2 = \frac{1}{m-1}\sum_{k=1}^{m}(Y_k - \bar{Y})^2. \tag{4}$$

Jei teisinga hipotezė $a_1 - a_2 = b_0$, tai analogiškai 8 skyrelio lemai galima parodyti, kad ši statistika pasiskirsčiusi pagal Stjudento dėsnį su $n + m - 2$ laisvės laipsnių.

Jei ir dispersijų santykis yra nežinomas, tai galima imti statistiką

$$\frac{\bar{X} - \bar{Y} - b_0}{\sqrt{(n-1)S_{11}^2 + (m-1)S_{12}^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}.$$

Kai teisinga nulinė hipotezė, ta statistika yra apytiksliai pasiskirsčiusi pagal Stjudento dėsnį su $n + m - 2$ laisvės laipsnių. Kritines sritis galima sudaryti kaip ir anksčiau. Tačiau šiuo atveju statistika gali patekti į kritinę sritį ir dėl dispersijų skirtumo, nors nulinė hipotezė apie vidurkius ir teisinga. Esama būdų, leidžiančių sumažinti σ_1^2/σ_2^2 įtaką.

7. Tenka lyginti ir dviejų dydžių dispersijas. Sakykime, turime du normaliuosius atsitiktinius dydžius su n e ž i n o m o m i s d i s p e r s i j o m i s σ_1^2 ir σ_2^2. Reikia patikrinti hipotezę $\sigma_1^2 = \lambda_0\sigma_2^2$. Jei v i d u r k i a i a_1 ir a_2 yra ž i n o m i, tai imame statistiką $S_{01}^2/(\lambda_0 S_{02}^2)$; čia

$$S_{01}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - a_1)^2, \quad S_{02}^2 = \frac{1}{m}\sum_{k=1}^{m}(Y_k - a_2)^2.$$

Jei hipotezė teisinga, tai ta statistika pagal III.9.5 pavyzdį turi Fišerio pasiskirstymą su n ir m laisvės laipsnių.

Jei alternatyva yra $\sigma_1^2 > \lambda_0\sigma_2^2$, tai hipotezę atmetame, kai

$$s_{01}^2/(\lambda_0 s_{02}^2) \geq u;$$

čia $u = u_\alpha$ yra Fišerio pasiskirstymo su n ir m laisvės laipsnių $(1-\alpha)$-kvantilis. Jei alternatyva yra $\sigma_1^2 < \lambda_0^2\sigma_2^2$, tai hipotezė atmetama, kai

$$s_{01}^2/(\lambda_0 s_{02}^2) \leq u;$$

čia $u = u_\alpha$ yra α-kvantilis (žr. [17], VII lentelę).

Pagaliau, jei alternatyva yra $\sigma_1^2 \neq \lambda_0\sigma_2^2$, tai nulinę hipotezę atmetame, kai

$$s_{01}^2/(\lambda_0 s_{02}^2) \leq u' \quad \text{arba} \quad s_{01}^2/(\lambda_0 s_{02}^2) \geq u'';$$

skaičius $u' = u'_\alpha$ yra Fišerio pasiskirstymo su m ir n laisvės laipsnių $\alpha/2$-kvantilis, o $u'' = u''_\alpha$ – to pasiskirstymo $(1-\alpha/2)$-kvantilis.

Jei v i d u r k i a i a_1 ir a_2 yra n e ž i n o m i, tai nagrinėjame statistiką $S_{11}^2/(\lambda_0 S_{12}^2)$; dydžiai S_{11}^2 ir S_{12}^2 apibrėžti (4) formule. Ši statistika, kai teisinga nulinė hipotezė, pagal 8 skyrelio lemą ir III.9.5 pavyzdį turi Fišerio pasiskirstymą su $n-1$ ir $m-1$ laisvės laipsnių. Kritinės sritys sudaromos analogiškai.

Paliekame skaitytojui parašyti tų kriterijų galios funkcijas.

3 p a v y z d y s. Detalėms gaminti buvo pasiūlyti du būdai. Kadangi jos gaminamos iš reto metalo, tai buvo tiriama, kuriuo būdu gaminant detalę, reikia mažiau metalo. Pirmuoju būdu buvo pagamintos 6 detalės; joms prireikė 3,1; 3,8; 3,6; 4,0; 3,4; 3,7 gramų metalo. Antruoju būdu pagamino 5 detales; joms prireikė 4,1; 4,3; 4,7; 4,8; 4,6 gramų metalo.

Laikysime, kad abiem atvejais turime normalųjį pasiskirstymą. Pirmiausia patikrinsime hipotezę, ar su reikšmingumo lygmeniu 0,1 galime laikyti, kad abiem atvejais dispersijos yra tos pačios. Turime:

$$\bar{x} = 3,6, \quad \bar{y} = 4,5,$$
$$5s_{11}^2 = 0,5, \quad 4s_{12}^2 = 0,34.$$

Kai teisinga ši hipotezė, statistika S_{11}^2/S_{12}^2 yra pasiskirsčiusi pagal Fišerio dėsnį su 5 ir 4 laisvės laipsniais. Randame jo 0,05 ir 0,95-kvantilius. Jie apytiksliai lygūs 0,19 ir 6,26, o $s_{11}^2/s_{12}^2 \approx 1,18$. Hipotezė priimtina.

Laikydami, kad abi dispersijos yra lygios, patikrinsime, ar ir vidurkiai yra lygūs. Reikšmingumo lygmenį imsime 0,05. Rasime Stjudento pasiskirstymo su 9 laisvės laipsniais 0,975-kvantilį. Jis lygus 2,26. Hipotezę reikia atmesti, kai

$$\frac{|\bar{x}-\bar{y}|}{\sqrt{5s_{11}^2+4s_{12}^2}}\sqrt{\frac{6\cdot 5\cdot 9}{11}} \geq 2,26.$$

Kairioji šios lygybės pusė mūsų atveju yra apytiksliai lygi 4,87. Hipotezė atmestina.

12. χ^2 KRITERIJUS

Tarp daugybės kriterijų statistinėms hipotezėms tikrinti svarbią vietą užima vadinamasis χ^2 kriterijus, pasiūlytas K. Pirsono. Jis yra gana paprastas ir plačiai taikomas.

Tarkime, kad atliekame n nepriklausomų eksperimentų. Atlikus bet kurį eksperimentą, įvyksta vienas iš nesutaikomų įvykių $A_1, ..., A_r, A_1 \cup ... \cup A_r = \Omega$, su pastoviomis tikimybėmis $p_1, ..., p_r; p_1 + ... + p_r = 1$. Pažymėkime

įvykio A_k įvykimų skaičių, atlikus n eksperimentų, raide κ_k. Taigi $\kappa_1 + ... + +\kappa_r = n$. Statistiniai dažniai $\kappa_1/n, ..., \kappa_r/n$ yra tikimybių p_k įverčiai.

Įvesime dažnių $\kappa_1/n, ..., \kappa_r/n$ ir tikimybių $p_1, ..., p_r$ nukrypimo matą

$$\sum_{k=1}^{r} c_k \Big(\frac{\kappa_k}{n} - p_k \Big)^2;$$

čia c_k yra bet kokios teigiamos konstantos. Tinkamai jas parinkus, tas reiškinys turi paprastas asimptotines savybes. Taip, pavyzdžiui, yra, kai $c_k = n/p_k$. Pažymėkime

$$\Delta_n = \sum_{k=1}^{r} \frac{(\kappa_k - np_k)^2}{np_k} = \sum_{k=1}^{r} \frac{\kappa_k^2}{np_k} - n.$$

Lema. *Jei kompleksinis skaičius z tenkina nelygybę $|z| \le 1/2$, tai*

$$\Big| \ln(1+z) - z + \frac{z^2}{2} \Big| \le \frac{2}{3}|z|^3.$$

Į r o d y m a s analogiškas III.10 skyrelio lemos įrodymui. Turime

$$\Big| \ln(1+z) - z + \frac{z^2}{2} \Big| \le \sum_{k=3}^{\infty} \frac{|z|^k}{k} \le \frac{1}{3} \sum_{k=3}^{\infty} |z|^k = \frac{|z|^3}{3(1-|z|)} \le \frac{2}{3}|z|^3. \ \square$$

1 (Pirsono) teorema. Δ_n *yra asimptotiškai pasiskirstęs pagal* χ^2 *su* $r-1$ *laisvės laipsnių.*

Į r o d y m a s. Tikimybė, kad $\kappa_1 = m_1, ..., \kappa_r = m_r$, yra lygi

$$\frac{n!}{m_1!...m_r!} p_1^{m_1} ... p_r^{m_r}.$$

Todėl vektoriaus $(\kappa_1, ..., \kappa_r)$ charakteristinė funkcija

$$\begin{aligned} & f_{(\kappa_1,...,\kappa_r)}(t_1, ..., t_r) = \\ & = \sum_{\substack{m_1 \ge 0, ..., m_r \ge 0 \\ m_1 + ... + m_r = n}} e^{i(t_1 m_1 + ... + t_r m_r)} \frac{n!}{m_1!...m_r!} p_1^{m_1} ... p_r^{m_r} = \\ & = (p_1 e^{it_1} + ... + p_r e^{it_r})^n. \end{aligned}$$

Pažymėkime

$$Z_k = \frac{\kappa_k - np_k}{\sqrt{np_k}} \quad (k = 1, ..., r).$$

Tada

$$\Delta_n = \sum_{k=1}^{r} Z_k^2, \quad \sum_{k=1}^{r} Z_k \sqrt{p_k} = 0.$$

Vektoriaus $Z = (Z_1, ..., Z_r)$ charakteristinė funkcija

$$f_Z(t_1, ..., t_r) = \exp\Big(- i\sqrt{n} \sum_{k=1}^{r} t_k \sqrt{p_k}\Big) f\Big(\frac{t_1}{\sqrt{np_1}}, ..., \frac{t_r}{\sqrt{np_r}}\Big) =$$

$$= \Big(\sum_{k=1}^{r} p_k e^{it_k/\sqrt{np_k}}\Big)^n \exp\Big(- i\sqrt{n} \sum_{k=1}^{r} t_k \sqrt{p_k}\Big).$$

Pasinaudoję nelygybe (žr. III.8 skyrelio lemą)

$$\Big| e^{iv} - 1 - iv + \frac{v^2}{2} \Big| \le \frac{|v|^3}{6},$$

teisinga visiems realiesiems v, ir laikydami n pakankamai dideliu, gauname

$$\ln f_Z(t_1, ..., t_r) = n \ln \sum_{k=1}^{r} p_k e^{it_k/\sqrt{np_k}} - i\sqrt{n} \sum_{k=1}^{r} t_k \sqrt{p_k} =$$

$$= n \ln \sum_{k=1}^{r} p_k \Big(1 + \frac{it_k}{\sqrt{np_k}} - \frac{t_k^2}{2np_k} + \frac{B|t_k|^3}{(np_r)^{3/2}}\Big) - i\sqrt{n} \sum_{k=1}^{r} t_k \sqrt{p_k} =$$

$$= n \ln \Big(1 + \frac{i}{\sqrt{n}} \sum_{k=1}^{r} t_k \sqrt{p_k} - \frac{1}{2n} \sum_{k=1}^{r} t_k^2 + \frac{B}{n^{3/2}}\Big) - i\sqrt{n} \sum_{k=1}^{r} t_k \sqrt{p_k}.$$

Čia ir toliau B reiškia skaičių, ne visada tą patį, bet aprėžtą konstantos, nepriklausančios nuo n. Remsimės šio skyrelio lema. Pakankamai dideliems n

$$\ln f_Z(t_1, ..., t_r) = n\Big(\frac{i}{\sqrt{n}} \sum_{k=1}^{r} t_k \sqrt{p_k} - \frac{1}{2n} \sum_{k=1}^{r} t_k^2 + \frac{B}{n^{3/2}}\Big) -$$

$$- \frac{n}{2}\Big(\frac{i}{\sqrt{n}} \sum_{k=1}^{r} t_k \sqrt{p_k} + \frac{B}{n}\Big)^2 + \frac{B}{\sqrt{n}} - i\sqrt{n} \sum_{k=1}^{r} t_k \sqrt{p_k} =$$

$$= -\frac{1}{2} \sum_{k=1}^{r} t_k^2 + \frac{1}{2}\Big(\sum_{k=1}^{r} t_k \sqrt{p_k}\Big)^2 + \frac{B}{\sqrt{n}}.$$

Todėl, kai $n \to \infty$,

$$f_Z(t_1, ..., t_r) \to \exp\Big\{ - \frac{1}{2} Q(t_1, ..., t_r)\Big\};$$

čia

$$Q(t_1,...,t_r)=\sum_{k=1}^{r}t_k^2-\Big(\sum_{k=1}^{r}t_k\sqrt{p_k}\Big)^2$$

yra neneigiamai apibrėžta kvadratinė forma (įrodykite!).

Tarkime, kad $Y=(Y_1,...,Y_r)$ yra normalusis vektorius su charakteristine funkcija $\exp\Big\{-\frac{1}{2}Q(t_1,...,t_r)\Big\}$. Kvadratinę formą Q, panaudoję ortogonalią transformaciją $(t_1,...,t_r)=(u_1,...,u_r)C$ su sąlyga $u_r=t_1p_1^{1/2}+...+t_rp_r^{1/2}$, galime užrašyti pavidalu $Q(t_1,...,t_r)=u_1^2+...+u_{r-1}^2$. Tai bus vektoriaus $Y(C^{-1})'=(V_1,...,V_{r-1},0)$ (žr. III.13 skyrelį) charakteristinė funkcija. Atsitiktiniai dydžiai $V_1,...,V_{r-1}$ yra nepriklausomi ir pasiskirstę pagal $N(0,1)$.

Pasinaudoję teorema apie tolydžią atitiktį tarp pasiskirstymo funkcijų ir charakteristinių funkcijų (žr. III.13 skyrelį), gauname

$$f_{\Delta_n}(t)=\int...\int_{R^r}e^{it(v_1^2+...+v_r^2)}dF_Z(v_1,...,v_r)\to$$
$$\to\int...\int_{R^r}e^{it(v_1^2+...+v_r^2)}dF_Y(v_1,...,v_r)=f_{Y_1^2+...+Y_r^2}(t).$$

Kadangi matrica $(C^{-1})'$ yra ortogonali, tai $Y_1^2+...+Y_r^2=V_1^2+...+V_{r-1}^2$. Todėl

$$f_{\Delta_n}(t)\to f_{V_1^2+...+V_{r-1}^2}(t),$$

kai $n\to\infty$. Vadinasi, Δ_n yra asimptotiškai pasiskirstęs pagal χ^2 su $r-1$ laisvės laipsnių. □

Remdamiesi šia teorema, galime sudaryti kriterijų hipotezei $H_0: p_1=$ $=p_1^0,...,p_r=p_r^0$ tikrinti. Jei įvykių $A_1,...,A_r$ tikimybės žymiai skiriasi nuo $p_1^0,...,p_r^0$, tai skirtumai $\kappa_k/n-p_k^0$ dažniau įgis didesnes reikšmes, negu jas įgytų tuo atveju, kai hipotezė teisinga.

Paėmę reikšmingumo lygmenį α, kritinę kriterijaus sritį apibrėšime nelygybe $\Delta_n\geq u$; skaičių $u=u_\alpha$ randame iš lygties

$$\mathcal{P}(\Delta_n\geq u)=\alpha.$$

Statistikos Δ_n pasiskirstymo nežinome, bet žinome, kad ji asimptotiškai pasiskirsčiusi pagal χ^2 su $r-1$ laisvės laipsnių. Todėl tą lygtį, kai n yra pakankamai didelis, pakeičiame apytiksle lygtimi

$$\int_u^\infty p_{\chi^2_{r-1}}(v)dv\approx\alpha.$$

Taigi u_α laikome apytiksliai lygiu pasiskirstymo χ^2 su $r-1$ laisvės laipsnių $(1-\alpha)$-kvantiliu.

1 p a v y z d y s. Genetikos pradininkas G. Mendelis[1] darė bandymus su žirniais. Tarp jo užaugintų žirnių buvo 315 lygūs ir geltoni, 108 – lygūs ir žali, 101 – raukšlėti ir geltoni bei 32 – raukšlėti ir žali. Pagal paveldimumo teoriją tokių žirnių skaičių santykis turėtų būti 9:3:3:1. Ar Mendelio duomenys patvirtina jo teoriją su reikšmingumo lygmeniu $\alpha = 0,05$?

Šiuo atveju $p_1 = 9/16; p_2 = p_3 = 3/16; p_4 = 1/16; n = 315 + 108 + 101 + 32 = 556$. Pasiskirstymo χ^2 su 3 laisvės laipsniais 0,95-kvantilis yra $\approx 7,815$. Statistikos Δ_n reikšmė

$$\frac{315^2}{556 \cdot 9/16} + \frac{108^2}{556 \cdot 3/16} + \frac{101^2}{556 \cdot 3/16} + \frac{32^2}{556 \cdot 1/16} - 556 \approx 0,470.$$

Hipotezė atitinka stebėjimo duomenis.

Iš Pirsono teoremos įrodymo išplaukia, kad Δ_n pasiskirstymo konvergavimas į ribinį yra tuo lėtesnis, kuo mažesnės p_k. Rekomenduojama šį kriterijų taikyti tada, kai $np_k \geq 5$.

Remdamiesi Pirsono teorema, galime sudaryti kriterijų tikrinti hipotezei H_0, kad stebimasis atsitiktinis dydis turi kokią nors konkrečią pasiskirstymo funkciją $F(y)$. Ta funkcija turi būti visiškai apibrėžta ir neturi priklausyti nuo jokių nežinomų parametrų.

Padalykime atsitiktinio dydžio visų galimų reikšmių aibę W į poaibius $W_1, ..., W_r$, kurie kas du neturi bendrų elementų ir kurių sąjunga lygi aibei W. Pažymėkime A_k įvykį, kai atsitiktinis dydis įgyja reikšmę iš srities W_k. Jei hipotezė H_0 yra teisinga, tai galime apskaičiuoti įvykių A_k tikimybes p_k^0. Užuot tikrinę hipotezę H_0, tikriname hipotezę H_0', kad tikimybės stebimajam atsitiktiniam dydžiui įgyti reikšmes iš aibių W_k yra p_k^0 $(k = 1, ..., r)$.

Dažnai tenka tikrinti hipotezes, kai pasiskirstymas yra, pavyzdžiui, normalusis, Puasono ar pan. Tokie pasiskirstymai priklauso nuo parametrų, kurie paprastai yra nežinomi. Sakykime, reikia patikrinti hipotezę, kad stebimojo dydžio pasiskirstymas yra $P_{(\theta_1,...,\theta_s)}$ su kokiomis nors parametrų $\theta_1, ..., \theta_s$ reikšmėmis. Kaip ir anksčiau, suskaidę to dydžio galimų reikšmių aibę W į sritis $W_1, ..., W_r$ $(r > s)$, imame tikimybes $p_k(\theta_1, ..., \theta_s)$ $(k = 1, ..., r)$, kad stebimasis dydis įgis reikšmes iš poaibių W_k, ir sudarome atstumo matą

$$\Delta_n(\theta_1, ..., \theta_s) = \sum_{k-1}^{r} \frac{\big(\kappa_k - np_k(\theta_1, ..., \theta_s)\big)^2}{np_k(\theta_1, ..., \theta_s)}. \tag{1}$$

Jei parametrų $\theta_1, ..., \theta_s$ reikšmės būtų žinomos, tai turėtume jau išnagrinėtą atvejį. Tačiau jos nežinomos. Peršasi paprasta išeitis: reikia parametrus pakeisti jų įverčiais. Imkime kuriuos nors parametrų įverčius $\theta_1^*, ..., \theta_s^*$ ir pakeiskime jais (1) formulėje pačius parametrus $\theta_1, ..., \theta_s$. Deja, vėl iškils nauji sunkumai: $p_k(\theta_1^*, ..., \theta_s^*)$ bus ne konstantos, bet atsitiktiniai dydžiai. Todėl 1 teorema nebus pritaikoma.

[1] Gregor–Johann Mendel (1822–1884) – austrų botanikas.

Įverčius $\theta_1^*, ..., \theta_s^*$ galime parinkti daugybe būdų. Suprantama, statistikos Δ_n pasiskirstymas priklausys nuo tų įverčių parinkimo. Juos galima parinkti ir taip, kad po nedidelių pakeitimų tiktų anksčiau išdėstyta teorija. Natūralu stengtis įverčius $\theta_1^*, ..., \theta_s^*$ parinkti taip, kad statistika Δ_n būtų kuo mažesnė. Tai – vadinamasis χ^2 *minimumo metodas*. Tam reikalui (1) diferencijuojame parametrų $\theta_1, ..., \theta_s$ atžvilgiu (jei toks diferencijavimas yra galimas) ir gautas dalines išvestines prilyginame 0. Gauname lygčių sistemą

$$(2) \qquad -\frac{1}{2}\frac{\partial \Delta_n}{\partial \theta_j} = \sum_{k=1}^{r} \Big(\frac{\kappa_k - np_k}{p_k} + \frac{(\kappa_k - np_k)^2}{2np_k^2} \Big) \frac{\partial p_k}{\partial \theta_j} = 0 \quad (j = 1, ..., s).$$

Iš jų randame įverčius $\tilde{\theta}_j$ ir jais pakeičiame θ_j (1) formulėje. Kai patenkintos gana bendros sąlygos, $\Delta_n(\tilde{\theta}_1, ..., \tilde{\theta}_s)$ yra asimptotiškai pasiskirsčiusi pagal χ^2 su $r - s - 1$ laisvės laipsnių.

(2) lygtis net paprasčiausiais atvejais sunku išspręsti. Galima būtų parodyti, kad antruosius narius tose lygtyse, kai n yra didelis, galima atmesti; nuo to nesikeičia statistikos Δ_n asimptotinis pasiskirstymas. Turime paprastesnę lygčių sistemą

$$(3) \qquad \sum_{k=1}^{r} \frac{\kappa_k - np_k}{p_k} \frac{\partial p_k}{\partial \theta_j} = 0 \quad (j = 1, ..., s).$$

Ši sistema gaunama, prilyginus nuliui $\Delta_n(\theta_1, ..., \theta_s)$ išvestines pagal parametrus $\theta_1, ..., \theta_s$, bet, skaičiuojant išvestines, laikoma, kad (1) formulės vardikliai yra pastovūs. Kadangi

$$\sum_{k=1}^{r} p_k(\theta_1, ..., \theta_s) = 1,$$

tai (3) sistemą galima dar suprastinti:

$$(4) \qquad \sum_{k=1}^{r} \frac{\kappa_k}{p_k} \frac{\partial p_k}{\partial \theta_j} = 0 \quad (j = 1, ..., s).$$

Šis metodas yra vadinamas *modifikuotu* χ^2 *minimumo metodu*.

2 teorema. *Tarkime, kad funkcijos* $p_k(\theta_1, ..., \theta_s)$ $(k = 1, ..., r)$ *kuriame nors neišsigimusiame erdvės* R^s *intervale* Θ *tenkina sąlygas:*

$$p_k(\theta_1, ..., \theta_s) \geq c > 0 \quad (k = 1, ..., r),$$
$$\sum_{k=1}^{r} p_k(\theta_1, ..., \theta_s) = 1,$$

egzistuoja tolydžios išvestinės

$$\frac{\partial p_k}{\partial \theta_j},\ \frac{\partial^2 p_k}{\partial \theta_j \partial \theta_l} \quad (k=1,...,r; j=1,...,s; l=1,...,s)$$

ir matricos

$$\left\| \frac{\partial p_k}{\partial \theta_j} \right\| \quad (k=1,...,r; j=1,...,s)$$

rangas yra lygus s. Sakykime, atliekame šio skyrelio pradžioje aprašytus n nepriklausomų eksperimentų ir tikimybė, kad įvykis A_k įvyks, atlikus kurį nors iš tų eksperimentų, yra $p_k^0 = p_k(\theta_1^0,...,\theta_s^0)$; čia $(\theta_1^0,...,\theta_s^0)$ – vidinis intervalo Θ taškas. Tada (3) *lygtys turi vienintelį sprendinį $(\hat{\theta}_1,...,\hat{\theta}_s)$, konverguojantį pagal tikimybę į $(\theta_1^0,...,\theta_s^0)$. Statistika*

$$\Delta_n(\hat{\theta}_1,...,\hat{\theta}_s) = \sum_{k=1}^{r} \frac{\left(\kappa_k - np_k(\hat{\theta}_1,...,\hat{\theta}_s)\right)^2}{np_k(\hat{\theta}_1,...,\hat{\theta}_s)} \tag{5}$$

yra asimptotiškai pasiskirsčiusi pagal χ^2 su $r-s-1$ laisvės laipsnių.

Šios teoremos įrodymas yra gana ilgas ir sudėtingas. Jį galima rasti, pavyzdžiui, [6], 30.3 skyrelyje.

Ta teorema grindžiamas kriterijus konstruojamas jau mums žinomais būdais.

Sakykime, reikia patikrinti hipotezę H_0, jog stebimojo atsitiktinio dydžio pasiskirstymas priklauso klasei $\{P_\theta,\ \theta \in \Theta\}$, $\Theta_0 \subset R^s$. Suskirstykime to dydžio galimų reikšmių aibę į r aibių W_k $(k=1,...,r)$. Pažymėkime p_k tikimybę įvykio A_k, kad atsitiktinis dydis pateks į W_k. Ši tikimybė priklauso nuo parametrų $\theta = (\theta_1,...,\theta_s)$. Jei teisingos 2 teoremos sąlygos, hipotezę H_0 pakeičiame hipoteze H_0', kad įvykių A_k tikimybės būtų p_k, sprendžiame (4) lygčių sistemą ir sudarome statistiką $\Delta_n(\hat{\theta}_1,...,\hat{\theta}_s)$.

Pailiustruosime šią procedūrą dviem pavyzdžiais.

Sakykime, reikia patikrinti hipotezę H_0, kad stebimasis atsitiktinis dydis yra pasiskirstęs pagal Puasono dėsnį su nežinomu parametru $\lambda > 0$. Šis dydis įgyja sveikąsias neneigiamas reikšmes. Suskaidykime sveikųjų neneigiamų skaičių aibę į poaibius $W_1 = \{0,1,...,l\}$, $W_k = \{l+k-1\}$ $(k=2,...,r-1)$, $W_r = \{l+r-1, l+r,...\}$. Tada

$$p_1 = p_1(\lambda) = \sum_{m=0}^{l} \frac{\lambda^m}{m!} e^{-\lambda},$$

$$p_k = p_k(\lambda) = \frac{\lambda^{l+k-1}}{(l+k-1)!} e^{-\lambda} \ (k=2,...,r-1),$$

$$p_r = p_r(\lambda) = \sum_{m=l+r-1}^{\infty} \frac{\lambda^m}{m!} e^{-\lambda}.$$

Tikimybės p_k tenkina 2 teoremos sąlygas. (4) sistema yra sudaryta iš vienos lygties

$$\kappa_1 \frac{\sum\limits_{m=0}^{l}\Big(\frac{m}{\lambda}-1\Big)\frac{\lambda^m}{m!}}{\sum\limits_{m=0}^{l}\frac{\lambda^m}{m!}} + \sum_{k=2}^{r-1}\Big(\frac{l+k-1}{\lambda}-1\Big)\kappa_k + \\ + \kappa_r \frac{\sum\limits_{m=l+r-1}^{\infty}\Big(\frac{m}{\lambda}-1\Big)\frac{\lambda^m}{m!}}{\sum\limits_{m=l+r-1}^{\infty}\frac{\lambda^m}{m!}} = 0.$$

Iš čia

$$\lambda = \frac{1}{n}\left(\kappa_1 \frac{\sum\limits_{m=0}^{l} m\frac{\lambda^m}{m!}}{\sum\limits_{m=0}^{l}\frac{\lambda^m}{m!}} + \sum_{k=2}^{r-1}(l+k-1)\kappa_k + \kappa_r \frac{\sum\limits_{m=l+r-1}^{\infty} m\frac{\lambda^m}{m!}}{\sum\limits_{m=l+r-1}^{\infty}\frac{\lambda^m}{m!}}\right).$$

Vidurinė suma lygi

$$\sum_{l<x_k<l+r-1} X_k;$$

pirmoji ir trečioji sumos paprastai nedaug skiriasi nuo sumų

$$\sum_{x_k\le l} X_k, \quad \sum_{x_k\ge l+r-1} X_k.$$

Todėl sprendinys yra $\tilde{\lambda}\approx\bar{X}$. Patariama poaibius W_k parinkti taip, kad būtų $np_k\ge 5$.

2 p a v y z d y s. Buvo stebimos radioaktyviosios dalelės ir registruojama, kiek signalų gaunama kas valandą. Štai rezultatai: 0 signalų užregistruota 1924 kartus, 1 signalas – 541 kartą, 2 signalai – 103 kartus, 3 signalai – 17 kartų, 4 signalai – 1 kartą, 5 signalai – 1 kartą, 6 ir daugiau signalų – nė vieno karto. Su reikšmingumo lygmeniu 0,05 patikrinsime hipotezę, kad signalų skaičius pasiskirstęs pagal Puasono dėsnį.

Čia $n = 1924+541+103+17+1+1 = 2587$, empirinis vidurkis $\bar{x} = (0\cdot 1924 + +1\cdot 541 + 2\cdot 103 + 3\cdot 17 + 4\cdot 1 + 5\cdot 1 + 6\cdot 0)/n \approx 0,3119443$.

Sudarome poaibius $W_k = \{k-1\}$ $(k = 1, 2, 3)$, $W_4 = \{3, 4, ...\}$.

Skaičiavimai surašyti lentelėje.

k	κ_k	p_k	np_k	$(\kappa_k - np_k)^2/(np_k)$
1	1924	0,7320223	1893,74	0,48
2	541	0,2283502	590,74	4,19
3	103	0,0356163	92,14	1,28
4	19	0,0040112	10,38	7,17
Iš viso	2587	1,0000000	2587	13,12

χ^2 pasiskirstymo su 2 laisvės laipsniais 0,95-kvantilis yra apytiksliai lygus 5,991. Todėl hipotezė atmestina.

Remdamiesi nagrinėjamuoju kriterijumi, tikrinsime hipotezę, kad stebimasis atsitiktinis dydis yra pasiskirstęs pagal normalųjį dėsnį su nežinomais parametrais $a \in R$ ir $\sigma > 0$. Suskaidykime realiųjų skaičių tiesę į sritis $W_1 = (-\infty, \tau - h/2), W_k = [\tau_k - h/2, \tau_k + h/2)$ $(k = 2, ..., r-1), W_r = [\tau_{r-1} + h/2, \infty)$; čia $\tau_k = \tau + (k-1)h$. Jei hipotezė teisinga, tai

$$p_k = p_k(a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{W_k} \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du \quad (k = 1, ..., r).$$

Iš čia

$$\frac{\partial p_k}{\partial a} = \frac{1}{\sigma^3\sqrt{2\pi}} \int_{W_k} (u-a) \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du,$$

$$\frac{\partial p_k}{\partial \sigma^2} = \frac{1}{\sigma^4\sqrt{2\pi}} \int_{W_k} (u-a)^2 \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du -$$

$$- \frac{1}{\sigma^2\sqrt{2\pi}} \int_{W_k} \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du.$$

(4) lygčių sistemą galima užrašyti šitaip:

$$a = \frac{1}{n} \sum_{k=1}^{r} \kappa_k \frac{\displaystyle\int_{W_k} u \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du}{\displaystyle\int_{W_k} \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du},$$

$$\sigma^2 = \frac{1}{n} \sum_{k=1}^{r} \kappa_k \frac{\displaystyle\int_{W_k} (u-a)^2 \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du}{\displaystyle\int_{W_k} \exp\Big(- \frac{(u-a)^2}{2\sigma^2} \Big) du}.$$

Mažiems h teisingos apytikslės lygybės

$$\int_{W_k} u \exp\Big(-\frac{(u-a)^2}{2\sigma^2}\Big)du \approx \tau_k \int_{W_k} \exp\Big(-\frac{(u-a)^2}{2\sigma^2}\Big)du,$$
$$\int_{W_k} (u-a)^2 \exp\Big(-\frac{(u-a)^2}{2\sigma^2}\Big)du \approx (\tau_k - a)^2 \int_{W_k} \exp\Big(-\frac{(u-a)^2}{2\sigma^2}\Big)du,$$

kai $k = 2, ..., r-1$. Jei $\kappa_1 = \kappa_r = 0$, tai gauname apytikslius (4) sistemos sprendinius

$$\hat{a} = \frac{1}{n}\sum_k \kappa_k \tau_k,\ \hat{\sigma}^2 = \frac{1}{n}\sum_k \kappa_k (\tau_k - \hat{a})^2.$$

Išskleidę pointegralinius reiškinius Teiloro eilutėmis taškų τ_k aplinkose, gautume tikslesnius sprendinius:

$$\hat{a} = \frac{1}{n}\sum_k \kappa_k \tau_k + O(h^4),$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_k \kappa_k (\tau_k - \hat{a})^2 - \frac{h^2}{12} + O(h^4).$$

Atmetę narius $O(h^4)$, galime gauti apytikslius (4) sistemos sprendinius: a įvertis yra sugrupuotų stebėjimo duomenų vidurkis, o σ^2 įvertis – sugrupuotų duomenų dispersija su Šepardo pataisa.

Sprendiniai

$$\hat{a} = \frac{1}{n}\sum_{k=1}^{r} \kappa_k \tau_k,$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{r} \kappa_k (\tau_k - \hat{a})^2 - \frac{h^2}{12} \tag{6}$$

paprastai naudojami ir tada, kai h nėra labai mažas, o κ_1 ir κ_r nėra lygūs nuliui, nes dažnai jie gerai aproksimuoja (4) sistemos sprendinius. Rekomenduojama sekti, kad būtų $np_k \geq 5$. Kai į W_1 ir W_r patenka nedaug reikšmių κ_k, jas galima sujungti.

3 p a v y z d y s. Buvo stebimas atsitiktinis dydis. Stebėjimo duomenys sugrupuoti, paėmus $h = 2$. Į intervalus [4, 6), ..., [20,22) pateko atitinkamai 15, 20, 25, 30, 30, 27, 24, 16, 13 stebimojo dydžio reikšmių. Patikrinsime hipotezę, kad su reikšmingumo lygmeniu 0,05 atsitiktinis dydis pasiskirstęs pagal normalųjį dėsnį. Skaičiavimo rezultatai surašyti lentelėje.

k	κ_k	p_k	np_k	$(\kappa_k - np_k)^2/(np_k)$
1	15	0,0830	16,60	0,15
2	20	0,0901	18,02	0,22
3	25	0,1359	27,18	0,17
4	30	0,2129	42,58	3,72
5	30	0,1290	25,80	0,68
6	27	0,1461	29,22	0,17
7	24	0,1017	20,34	0,66
8	16	0,0583	11,66	1,62
9	13	0,0429	8,58	2,28
Iš viso	200	0,9999	199,98	9,67

Pagal (6) formules

$$\hat{a} = 12,25;\ \hat{\sigma}^2 \approx 4,512.$$

Sudarome sritis $W_1 = (-\infty, 6), W_2 = [6, 8), W_3 = [8, 10), ..., W_8 = [18, 20), W_9 = [20, \infty)$. Tikimybes p_k randame iš formulių

$$\begin{aligned}
p_1 &= \Phi\Big(\frac{6-\hat{a}}{\hat{\sigma}}\Big),\\
p_2 &= \Phi\Big(\frac{8-\hat{a}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{10-\hat{a}}{\hat{\sigma}}\Big),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
p_8 &= \Phi\Big(\frac{20-\hat{a}}{\hat{\sigma}}\Big) - \Phi\Big(\frac{18-\hat{a}}{\hat{\sigma}}\Big),\\
p_9 &= 1 - \Phi\Big(\frac{20-\hat{a}}{\hat{\sigma}}\Big).
\end{aligned}$$

Lentelėje visų p_k suma dėl apvalinimo paklaidų nėra lygi 1. Pasiskirstymo χ^2 su $9 - 2 - 1 = 6$ laisvės laipsniais 0,95-kvantilis yra apytiksliai lygus 12,592. Hipotezė priimtina.

χ^2 kriterijus labai plačiai taikomas. Paminėsime dar porą atvejų. Pirmasis yra *požymių nepriklausomumo tikrinimas.* Tarkime, kad generalinės aibės elementai turi du požymius A ir B. Pirmasis požymis turi r kategorijų $A_1, ..., A_r$, o antrasis – s kategorijų $B_1, ..., B_s$. Vadinasi, visus generalinės aibės elementus galime suskirstyti į rs klasių, priskirdami vienai klasei elementus, turinčius požymius A_j ir $B_k (j = 1, ..., r; k = 1, ..., s)$. Pažymėkime $p_{j\cdot}$ tikimybę, kad atsitiktinai parinktas generalinės aibės elementas turi požymį A_j, raide $p_{\cdot k}$ – tikimybę, kad jis turi požymį B_k, ir raide p_{jk} – tikimybę, kad jis turi požymius A_j ir B_k. Reikia patikrinti hipotezę, kad tie požymiai yra nepriklausomi, t. y. $p_{jk} = p_{j\cdot}p_{\cdot k}$ visiems $j = 1, ..., r; k = 1, ..., s$. Aišku,

$$\sum_{j=1}^{r} p_{j\cdot} = 1,$$

$$\sum_{k=1}^{s} p_{\cdot k} = 1.$$

Kai hipotezė teisinga, tikimybės p_{jk} yra $r + s - 2$ nežinomų parametrų $p_{1\cdot}, ..., p_{r-1\cdot}; p_{\cdot 1}, ..., p_{\cdot s-1}$ funkcijos.

Pažymėkime raide κ_{jk} skaičių imties elementų, turinčių požymius A_j ir B_k. (4) sistema bus pavidalo

$$\sum_{j=1}^{r} \Big(\frac{\kappa_{jk}}{p_{\cdot k}} - \frac{\kappa_{js}}{p_{\cdot s}} \Big) = 0 \quad (k = 1, ..., s-1),$$

$$\sum_{k=1}^{s} \Big(\frac{\kappa_{jk}}{p_{j\cdot}} - \frac{\kappa_{rk}}{p_{r\cdot}} \Big) = 0 \quad (j = 1, ..., r-1).$$

Gauname įverčius

$$\hat{p}_{j\cdot} = \frac{\kappa_{j\cdot}}{n} \quad (j = 1, ..., r),$$

$$\hat{p}_{\cdot k} = \frac{\kappa_{\cdot k}}{n} \quad (k = 1, ..., s);$$

čia

$$\kappa_{j\cdot} = \sum_{k=1}^{s} \kappa_{jk} \quad (j = 1, ..., r),$$

$$\kappa_{\cdot k} = \sum_{j=1}^{r} \kappa_{jk} \quad (k = 1, ..., s).$$

(5) statistika yra pavidalo

$$n \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\Big(\kappa_{jk} - \dfrac{\kappa_{j\cdot}\kappa_{\cdot k}}{n} \Big)^2}{\kappa_{j\cdot}\kappa_{\cdot k}} = n \Big(\sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\kappa_{jk}^2}{\kappa_{j\cdot}\kappa_{\cdot k}} - 1 \Big). \tag{7}$$

Ji yra asimptotiškai pasiskirsčiusi pagal χ^2 su $rs-(r+s-2)-1 = (r-1)(s-1)$ laisvės laipsnių.

4 p a v y z d y s. Dvi grupės ligonių A ir B, po 100 asmenų kiekviena, serga kokia nors liga. Grupės A ligoniai buvo gydomi serumu, o grupės B ligoniai to serumo negavo (B buvo kontrolinė grupė). Šiaip abi grupės buvo gydomos vienodai. Nustatyta, kad pirmojoje grupėje pagijo 75 ligoniai, o antrojoje 65. Su reikšmingumo lygmeniu 0,05 reikia patikrinti hipotezę, kad serumas nebuvo efektyvus.

Bandymo rezultatus surašome lentelėje.

	Pasveiko	Nepasveiko	Iš viso
Grupė A	75	25	100
Grupė B	65	35	100
Iš viso	140	60	

(7) statistikos reikšmė

$$200\Big(\frac{75^2}{140\cdot 100}+\frac{65^2}{140\cdot 100}+\frac{25^2}{60\cdot 100}+\frac{35^2}{60\cdot 100}-1\Big)\approx 2,38.$$

χ^2 pasiskirstymo su 1 laisvės laipsniu 0,95-kvantilis yra apytiksliai lygus 3,841. Todėl hipotezė atitinka stebėjimo duomenis.

Čia aprašytą procedūrą nesunku pritaikyti, remiantis anksčiau nurodytu būdu, dviejų atsitiktinių dydžių nepriklausomumui tikrinti. Paliekame tai padaryti skaitytojui.

Pritaikysime dar χ^2 kriterijų stebėjimų *homogeniškumui* tikrinti. Sakykime, turime s serijų bandymų; tose serijose yra atitinkamai $n_1,...,n_s$ pavienių bandymų. Atlikus bet kurį bandymą, gali įvykti vienas iš nesutaikomų įvykių $A_1,...,A_r$; visų tų įvykų sąjunga yra būtinas įvykis. Įvykių A_j skaičių k-ojoje bandymų serijoje pažymėkime $\kappa_{kj}(\kappa_{k1}+...+\kappa_{kr}=n_k)$, o įvykio A_j pasirodymo tikimybę, atlikus k-osios serijos bandymą, pažymėkime $p_{kj}(p_{k1}+$ $+...+p_{kr}=1)$. Remiantis stebėjimų duomenimis, reikia patikrinti hipotezę, kad kiekvienoje bandymų serijoje įvykio A_j pasirodymo tikimybė yra ta pati: $p_{kj}=p_j(j=1,...,r;k=1,...,s)$. Jei hipotezė teisinga, tai $p_1+...+p_r=1$.

Sprendžiant šį uždavinį, reikia atsižvelgti į tai, kiek iš tikimybių p_j yra žinomų.

Jei visos jos žinomos ir $s=1$, tai turime šio skyrelio pradžioje aprašytą atvejį. Atitinkama statistika Δ_n, remiantis Pirsono teorema, yra asimptotiškai pasiskirsčiusi pagal χ^2 dėsnį su $r-1$ laisvės laipsnių. Jei $s>1$, tai kiekvienai bandymų serijai sudarome statistiką Δ_{n_k} ir visas jas sudedame. Gautoji statistika

$$\sum_{k=1}^{s}\sum_{j=1}^{r}\frac{(\kappa_{kj}-n_kp_j)^2}{n_kp_j},$$

analogiškai 1 teoremai, bus asimptotiškai pasiskirsčiusi pagal χ^2 su $s(r-1)$ laisvės laipsnių.

Jei visos tikimybės $p_1,...,p_r$ yra nežinomos, tai p_{kj} yra nežinomų parametrų $p_1,...,p_{r-1}$ funkcijos. Tų parametrų įverčius randame iš (4) lygčių sistemos; jie lygūs

$$\hat{p}_j=\frac{\kappa_{\cdot j}}{n}\ (j=1,...,r);$$

čia $\kappa_{\cdot j} = \kappa_{1j} + ... + \kappa_{sj}$. Sudarome statistiką

$$\sum_{k=1}^{s}\sum_{j=1}^{r}\frac{(\kappa_{kj} - n_k\hat{p}_j)^2}{n_k\hat{p}_j} = n\Big(\sum_{k=1}^{s}\sum_{j=1}^{r}\frac{\kappa_{kj}^2}{n_k\kappa_{\cdot j}} - 1\Big). \tag{8}$$

Galima būtų parodyti, kad ji turi asimptotinį pasiskirstymą χ^2 su $(r-1)\times$ $\times(s-1)$ laisvės laipsnių, kai tikrinamoji hipotezė yra teisinga.

Jei tikimybės $p_1, ..., p_r$ yra parametrų $\theta_1, ..., \theta_l$ $(1 \le l < s)$ funkcijos, tai parametrų θ_m įverčius randame iš (4) sistemos ir sudarome statistiką, analogišką (5). Galima būtų parodyti, kad ji turi asimptotinį χ^2 pasiskirstymą su $s(r-1) - l$ laisvės laipsnių.

5 p a v y z d y s. Lentelėje pateikti įvairiais 1935 m. mėnesiais Švedijoje gimusių berniukų ir mergaičių skaičiai (pavyzdys paimtas iš [6] knygos).

Mėnuo	Berniukų skaičius	Mergaičių skaičius	Iš viso
1	3743	3537	7280
2	3550	3407	6957
3	4017	3866	7883
4	4173	3711	7884
5	4117	3775	7892
6	3944	3665	7609
7	3964	3621	7585
8	3797	3596	7393
9	3712	3491	7203
10	3512	3991	6903
11	3392	3160	6552
12	3761	3371	7132
Iš viso	45682	42591	88273

Su reikšmingumo lygmeniu 0,05 patikrinsime hipotezę, kad visais mėnesiais berniuko gimimo tikimybė yra ta pati.

Čia turime $r = 2, s = 12$. Nežinomas parametras yra berniuko gimimo tikimybė p. Jos įvertis $\hat{p} = \kappa_{\cdot 1}/n$.

(8) statistikos

$$\sum_{k=1}^{12}\Big(\frac{(\kappa_{k1} - n_k\hat{p})^2}{n_k\hat{p}} + \frac{\big(\kappa_{k2} - n_k(1-\hat{p})\big)^2}{n_k(1-\hat{p})}\Big) =$$

$$= \sum_{k=1}^{12}\frac{(\kappa_{k1} - n_k\hat{p})^2}{n_k\hat{p}(1-\hat{p})} = \frac{1}{1-\hat{p}}\Big(\frac{1}{\hat{p}}\sum_{k=1}^{12}\frac{\kappa_{k1}^2}{n_k} - \kappa_{\cdot 1}\Big)$$

reikšmė yra apytiksliai lygi 14,986. χ^2 su $(s-1)(r-1) = 11$ laisvės laipsnių 0,95-kvantilis yra 19,675. Galime laikyti, kad hipotezė neprieštarauja stebėjimo duomenims.

13. KRITERIJAI, PAGRĮSTI EMPIRINĖS IR TEORINĖS PASISKIRSTYMO FUNKCIJŲ SKIRTUMU

Nagrinėsime, kaip empirinė pasiskirstymo funkcija aproksimuoja teorinę pasiskirstymo funkciją.

Tarkime, kad stebimojo atsitiktinio dydžio su pasiskirstymo funkcija $F(y)$ atsitiktinė imtis yra $(X_1, ..., X_n)$. Jo empirinę pasiskirstymo funkciją apibrėžėme 2 skyrelyje. Kiekvienam fiksuotam y tai yra atsitiktinis dydis

$$\mathcal{F}_n(y) = \frac{1}{n} \sum_{X_k < y} 1.$$

Pažymėkime Y_{ky} atsitiktinį dydį, įgyjantį reikšmę 1 su tikimybe $\mathcal{P}(X_k < y) = F(y)$ ir reikšmę 0 su tikimybe $\mathcal{P}(X_k \geq y) = 1 - F(y)$. Dydžiai $Y_{1y}, ..., Y_{ny}$ yra nepriklausomi. Tada empirinę pasiskirstymo funkciją galime užrašyti šitaip:

$$\mathcal{F}_n(y) = \frac{1}{n} \sum_{k=1}^{n} Y_{ky}.$$

Pastebėsime, kad dydžio Y_{ky} vidurkis

$$MY_{ky} = F(y),$$

o dispersija

$$DY_{ky} = F(y)\big(1 - F(y)\big).$$

Iš vidurkio ir dispersijos savybių išplaukia, kad

$$M\mathcal{F}_n(y) = F(y), \; D\mathcal{F}_n(y) = \frac{1}{n} F(y)\big(1 - F(y)\big).$$

Iš stipriojo didžiųjų skaičių dėsnio (III.5.2 teoremos 2 išvados) išplaukia, kad kiekvienam $y \in R$ empirinė pasiskirstymo funkcija konverguoja su tikimybe 1 į teorinę pasiskirstymo funkciją

$$\mathcal{P}\{\mathcal{F}_n(y) \xrightarrow[n\to\infty]{} F(y)\} = 1.$$

Vadinasi, $\mathcal{F}_n(y)$ yra nepaslinktasis ir suderintasis $F(y)$ įvertis.

Dar daugiau: iš centrinės ribinės teoremos (III.11.2 teoremos 2 išvados) išplaukia, kad kiekvienam $y \in R$ su sąlyga $0 < F(y) < 1$

$$\mathcal{P}\left\{\frac{\sum_{k=1}^{n}(Y_{ky} - MY_{ky})}{\left(\sum_{k=1}^{n} DY_{ky}\right)^{1/2}} < u\right\} = \mathcal{P}\left\{\frac{\sqrt{n}\big(\mathcal{F}_n(y) - F(y)\big)}{\sqrt{F(y)\big(1-F(y)\big)}} < u\right\} \xrightarrow[n\to\infty]{} \Phi(u).$$

Šie teiginiai rodo, kad kiekvienam fiksuotam $y \in R$ empirinė pasiskirstymo funkcija $\mathcal{F}_n(y)$ artėja prie tikrosios pasiskirstymo funkcijos $F(y)$. V. Glivenka[1] 1933 m. įrodė, kad $\mathcal{F}_n(y)$ su tikimybe 1 konverguoja į $F(y)$ *tolygiai* visiems $y \in R$. Įrodysime tą teoremą.

1 lema. *Jei turime baigtinę arba skaičią įvykių sistemą* $\{A_k\}$ *ir kiekvieno jų tikimybė* $P(A_k) = 1$, *tai ir tos sistemos sankirtos tikimybė*

$$P\Big(\bigcap_k A_k\Big) = 1.$$

Į r o d y m a s. Kadangi

$$P\Big(\bigcup_k A_k^c\Big) \le \sum_k P(A_k^c)$$

ir $P(A_k^c) = 0$, tai

$$P\Big(\bigcup_k A_k^c\Big) = 0.$$

Iš čia

$$P\Big(\bigcap_k A_k\Big) = 1 - P\bigg(\Big(\bigcap_k A_k\Big)^c\bigg) = 1 - P\Big(\bigcup_k A_k^c\Big) = 1. \ \square$$

1(Glivenkos) teorema. *Jei* $F(y)$ *yra atsitiktinio dydžio pasiskirstymo funkcija, o* $\mathcal{F}_n(y)$ *– jo empirinė pasiskirstymo funkcija, tai*

$$\mathcal{P}\Big\{\sup_{-\infty<y<\infty} |\mathcal{F}_n(y) - F(y)| \xrightarrow[n\to\infty]{} 0\Big\} = 1.$$

Į r o d y m a s. Imkime bet kurį natūralųjį skaičių r. Pažymėkime y_{rk} $(k = 1, ..., r)$ mažiausią y, tenkinantį nelygybes

[1] Valerijus Glivenka (1897–1940) – ukrainiečių kilmės matematikas.

$$F(y) \le \frac{k}{r} \le F(y+0).$$

Tarkime, kad

$$E_{rk} = \Big\{\mathcal{F}_n(y_{rk}) \xrightarrow[n\to\infty]{} F(y_{rk})\Big\},$$
$$E_r = E_{r1} \cap \ldots \cap E_{rr} = \Big\{\max_{1\le k\le r} |\mathcal{F}_n(y_{rk}) - F(y_{rk})| \xrightarrow[n\to\infty]{} 0\Big\},$$
$$E = E_1 \cap E_2 \cap \ldots = \Big\{\max_{1\le k\le r} |\mathcal{F}_n(y_{rk}) - F(y_{rk})| \xrightarrow[n\to\infty]{} 0; r = 1,2,\ldots\Big\}.$$

Iš stipriojo didžiųjų skaičių dėsnio išplaukia, kad

$$\mathcal{P}(E_{rk}) = 1 \quad (r = 1,2,\ldots; k = 1,\ldots,r).$$

Pagal 1 lemą

$$\begin{aligned} \mathcal{P}(E_r) &= 1, \\ \mathcal{P}(E) &= 1. \end{aligned} \tag{1}$$

Pažymėkime

$$C = \{\sup_{-\infty<y<\infty} |\mathcal{F}_n(y) - F(y)| \xrightarrow[n\to\infty]{} 0\}.$$

Jei įrodytume, kad $E \subset C$, tai iš (1) išplauktų teoremos teiginys.

Kiekvienam $y \in (y_{rk}, y_{r,k+1}]$ teisingos nelygybės

$$\mathcal{F}_n(y_{rk}+0) \le \mathcal{F}_n(y) \le \mathcal{F}_n(y_{r,k+1}) \tag{2}$$

ir

$$F(y_{rk}+0) \le F(y) \le F(y_{r,k+1}), \tag{3}$$

be to,

$$0 \le F(y_{r,k+1}) - F(y_{rk}+0) \le \frac{1}{r}. \tag{4}$$

Atėmę iš (2) nelygybės (3), gauname

$$\mathcal{F}_n(y_{rk}+0) - F(y_{r,k+1}) \le \mathcal{F}_n(y) - F(y) \le \mathcal{F}_n(y_{r,k+1}) - F(y_{rk}+0).$$

Iš čia ir iš (4) nelygybės išplaukia

$$|\mathcal{F}_n(y) - F(y)| \le \max_{1\le k\le r} |\mathcal{F}_n(y_{rk}) - F(y_{rk})| + \frac{1}{r}.$$

Todėl

$$\sup_{-\infty<y<\infty} |\mathcal{F}_n(y) - F(y)| \le \max_{1\le k\le r} |\mathcal{F}_n(y_{rk}) - F(y_{rk})| + \frac{1}{r}.$$

Kadangi ši nelygybė yra teisinga kiekvienam r, tai galime padaryti išvadą, kad $E \subset C$. □

Pažymėję

$$\mathcal{D}_n = \sup_y |\mathcal{F}_n(y) - F(y)|,$$

Glivenkos teoremą galime užrašyti šitaip:

$$\mathcal{P}(\mathcal{D}_n \xrightarrow[n\to\infty]{} 0) = 1.$$

Statistika $\mathcal{D}_n$ yra funkcijos $\mathcal{F}_n$ nuokrypio nuo F matas. Pasirodo, kad tolydžiosioms pasiskirstymo funkcijoms (įprasta matematinėje analizėje prasme) $\mathcal{D}_n$ pasiskirstymas nepriklauso nuo F.

2 lema. *Jei Z yra atsitiktinis dydis su tolydžia pasiskirstymo funkcija $H(z)$, tai atsitiktinio dydžio $Y = H(Z)$ pasiskirstymo funkcija yra*

$$G(y) = \begin{cases} 0, & \textit{kai } y \le 0, \\ y, & \textit{kai } 0 < y \le 1, \\ 1, & \textit{kai } y > 1. \end{cases}$$

Vadinasi, Y yra tolygiai pasiskirstęs intervale $(0, 1)$.

Į r o d y m a s. Tolydi funkcija $y = H(z)$ atvaizduoja tiesę R į vieną iš intervalų [0, 1], [0, 1), (0, 1], (0, 1). Jei $H(z)$ yra (griežtai) didėjanti, tai tas atvaizdis yra abipus vienareikšmis. Bendresniu atveju (kai $H(z)$ yra nemažėjanti ir egzistuoja intervalai, kuriuose ji yra pastovi) vieną y reikšmę gali atitikti daugiau z reikšmių – visas neišsigimęs intervalas. Kiekvienam $y \in (0, 1]$ apibrėžkime

$$z_y = \inf\{z : H(z) = y\}.$$

Iš funkcijos $H(z)$ tolydumo išplaukia, kad $H(z_y) = y$.

Lemą pakanka įrodyti tik visiems $y \in (0, 1)$. Turime

$$\begin{aligned} G(y) &= P(Y < y) = P\{H(Z) < y\} = P\{H(Z) < H(z_y)\} = \\ &= P(Z < z_y) = H(z_y) = y. \ \square \end{aligned}$$

2 teorema. *Kiekvienai tolydžiajai pasiskirstymo funkcijai $F(y)$ statistika $\mathcal{D}_n$ turi tą patį pasiskirstymą.*

Į r o d y m a s. Kaip ir visame šiame skyrelyje, $(X_1, ..., X_n)$ žymėsime atsitiktinę imtį. Stebimojo atsitiktinio dydžio pasiskirstymo funkcijos $F(y)$ pastovumo intervalu vadinsime kiekvieną uždarą intervalą $[b, c]$, jei $\mathcal{P}\{X_1 \in \in [b, c)\} = 0$ ir nėra kito uždaro intervalo, kuriam priklausytų $[b, c]$ su ta

savybe. Išmeskime iš R visus pastovumo intervalus. Gautą aibę pažymėkime A. Tada

$$\mathcal{D}_n = \sup_{y\in R} |\mathcal{F}_n(y) - F(y)| = \sup_{y\in A} |\mathcal{F}_n(y) - F(y)|$$

ir kiekvienam $y \in A$

$$\{X_k < y\} = \{F(X_k) < F(y)\}.$$

Pažymėkime $U_k = F(X_k)$ ir

$$\mathcal{G}_n(y) = \frac{1}{n} \sum_{U_k<y} 1.$$

Tada visiems $y \in A$

$$\mathcal{G}_n\big(F(y)\big) = \frac{1}{n} \sum_{F(X_k)<F(y)} 1 = \frac{1}{n} \sum_{X_k<y} 1 = \mathcal{F}_n(y).$$

Vadinasi,

$$\begin{aligned}\mathcal{D}_n &= \sup_{y\in A} \big|\mathcal{G}_n\big(F(y)\big) - F(y)\big| = \\ &= \sup_{y\in R} \big|\mathcal{G}_n\big(F(y)\big) - F(y)\big| = \sup_{0\le u\le 1} \big|\mathcal{G}_n(u) - u\big|.\end{aligned}$$

Teorema išplaukia iš 2 lemos. □

Teoremos įrodymas yra ir būdas statistikos $\mathcal{D}_n$ pasiskirstymo funkcijai rasti. Tam reikia imti atsitiktinio dydžio, tolygiai pasiskirsčiusio intervale $(0, 1)$, empirinę pasiskirstymo funkciją $\mathcal{G}_n(u)$ ir apskaičiuoti statistikos

$$\sup_{0\le u\le 1} |\mathcal{G}_n(u) - u|$$

pasiskirstymo funkciją. Ji bus ir $\mathcal{D}_n$ pasiskirstymo funkcija.

Galima rasti gana paprastą ribinį statistikos $\mathcal{D}_n$ pasiskirstymą.

3 (Kolmogorovo) teorema. *Kiekvienai tolydžiajai pasiskirstymo funkcijai* $F(y)$

$$\mathcal{P}(\sqrt{n}\mathcal{D}_n < y) \to K(y);$$

čia

$$K(y) = \begin{cases} 0, & \text{kai } y \le 0, \\ \displaystyle\sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2y^2}, & \text{kai } y > 0. \end{cases}$$

Šios teoremos įrodymo idėja išdėstyta 14 skyrelyje.

Funkcija $K(y)$ yra tabuliuota (žr. [17], XII lentelę, [2], 6.1, 6.2 lenteles).

Remdamiesi šiais rezultatais, pagal jau gerai žinomą schemą galime sudaryti Kolmogorovo kriterijų tikrinti hipotezei, kad stebimasis dydis yra pasiskirstęs pagal tolyduįį dėsnį $F(x)$. Tarkime, kad konkretūs stebėjimo rezultatai surašyti nemažėjančia tvarka. Turime variacinę seką $x_1^* \le x_2^* \le \le ... \le x_n^*$. Apskaičiuojame dydžius

$$D_n^+ = \max_{1\le k\le n}\Big(\frac{k}{n} - F_n(x_k^*)\Big), \quad D_n^- = \max_{1\le k\le n}\Big(F_n(x_k^*) - \frac{k-1}{n}\Big).$$

Tada

$$D_n = \max(D_n^+, D_n^-).$$

Paėmę reikšmingumo lygmenį α, iš lentelių randame $\mathcal{D}_n$ pasiskirstymo $(1-\alpha)$--kvantilį u_α. Kai n dideli, galima remtis Kolmogorovo teorema. Kai $n \ge 10$ ir $0,01 \le \alpha \le 0,2$, galima naudotis apytiksle formule

$$u_\alpha \approx \sqrt{\frac{1}{2n}\Big(v - \frac{2v^2-4v-1}{18n}\Big)} - \frac{1}{6n} \approx \sqrt{\frac{v}{2n}} - \frac{1}{6n};$$

čia $v = -\ln(\alpha/2)$.

Jei $D_n > u_\alpha$, tai hipotezė atmestina, priešingu atveju galime manyti, kad ji neprieštarauja stebėjimo duomenims.

Empirinės funkcijos nuokrypį nuo teorinės galima apibūdinti ir kitais būdais. H. Krameras 1928 m. ir nepriklausomai nuo jo R. Mizesas[1] 1931 m. pasiūlė tam reikalui naudoti statistiką

$$\int_{-\infty}^{\infty} \big(\mathcal{F}_n(y) - F(y)\big)^2 dL(y)$$

su nemažėjančia funkcija $L(y)$. Paprastai naudojamos dvi statistikos

$$\omega_n^2 = \int_{-\infty}^{\infty} \big(\mathcal{F}_n(y) - F(y)\big)^2 dF(y),$$

$$\Omega_n^2 = \int_{-\infty}^{\infty} \frac{\big(\mathcal{F}_n(y) - F(y)\big)^2}{F(y)\big(1-F(y)\big)} dF(y).$$

Čia funkcija $F(y)$, kaip ir anksčiau, yra tolydi. Galima būtų įrodyti, kad

$$n\omega_n^2 = \sum_{k=1}^{n}\Big(F(X_k^*) - \frac{2k-1}{2n}\Big)^2 + \frac{1}{12n},$$

$$n\Omega_n^2 = -2\sum_{k=1}^{n}\Big[\frac{2k-1}{2n}\ln F(X_k^*) + \Big(1 - \frac{2k-1}{2n}\Big)\ln\Big(1 - F(X_k^*)\Big)\Big] - n$$

[1] Richard von Mises (1883–1953) – vokiečių matematikas.

ir kad tų statistikų pasiskirstymai nepriklauso nuo funkcijos $F(y)$. Įrodyta, kad egzistuoja ir tų statistikų ribiniai pasiskirstymai

$$\mathcal{P}(n\omega_n^2 < u) \to a_1(u),$$
$$\mathcal{P}(n\Omega_n^2 < u) \to a_2(u).$$

Funkcijų $a_1(u)$ ir $a_2(u)$ analizinės išraiškos yra gana sudėtingos. Jos – tabuliuotos (žr. [17], XV, XVI lenteles, [2], 6.4 lentelę ir artutines formules). Šiais rezultatais pagrįstas vadinamasis ω^2, arba Kramero–Mizeso, kriterijus tikrinti hipotezei, kad stebimasis dydis yra pasiskirstęs pagal tolyduįį dėsnį $F(y)$. Jis sudaromas jau mums žinomais metodais.

P a v y z d y s. Automatinės staklės gamina rutuliukus. Atsitiktinai parinkti 25 rutuliukai ir išmatuotas jų skersmuo (milimetrais). Gautieji rezultatai surašyti lentelėje didėjančia tvarka. Su reikšmingumo lygmeniu $\alpha = 0,05$ reikia patikrinti hipotezę, kad rutuliukų skersmenys pasiskirstę pagal $N(12; 0,1)$. Skaičiavimų rezultatai surašyti lentelėje. Čia $z_k = 10(x_k^* - 12)$. Rezultatai apvalinami 0,001 tikslumu.

k	x_k^*	k/n	z_k	$\Phi(z_k)$	$k/n - \Phi(z_k)$	$\Phi(z_k) - (k-1)/n$
1	11,790	0,04	-2,10	0,018	0,022	0,018
2	11,834	0,08	-1,66	0,048	0,032	0,008
3	11,862	0,12	-1,38	0,084	0,036	0,004
4	11,882	0,16	-1,18	0,119	0,041	-0,001
5	11,902	0,20	-0,98	0,164	0,036	0,004
6	11,912	0,24	-0,88	0,189	0,051	-0,011
7	11,916	0,28	-0,84	0,200	0,080	-0,040
8	11,944	0,32	-0,56	0,288	0,032	0,008
9	11,954	0,36	-0,46	0,323	0,037	0,003
10	11,970	0,40	-0,30	0,382	0,018	0,022
11	11,986	0,44	-0,14	0,444	-0,004	0,044
12	11,990	0,48	-0,10	0,460	0,020	0,016
13	12,002	0,52	0,02	0,508	0,012	0,028
14	12,006	0,56	0,06	0,524	0,036	0,004
15	12,018	0,60	0,18	0,571	0,029	0,011
16	12,030	0,64	0,30	0,618	0,022	0,018
17	12,034	0,68	0,34	0,633	0,047	-0,007
18	12,040	0,72	0,40	0,655	0,017	-0,025
19	12,052	0,76	0,52	0,698	0,062	-0,022
20	12,062	0,80	0,62	0,732	0,068	-0,028

k	x_k^*	k/n	z_k	$\Phi(z_k)$	$k/n - \Phi(z_k)$	$\Phi(z_k) - (k-1)/n$
21	12,072	0,84	0,72	0,764	0,076	-0,036
22	12,090	0,88	0,90	0,816	0,064	-0,024
23	12,100	0,92	1,00	0,841	0,079	-0,039
24	12,122	0,96	1,22	0,889	0,071	-0,031
25	12,130	1,00	1,30	0,903	0,097	-0,057

Skaičiavimai rodo, kad

$$D_n = 0,097.$$

Iš lentelių randame, kad $\mathcal{D}_n$ pasiskirstymo 0,95-kvantilis yra apytiksliai lygus 0,264. Hipotezė atitinka stebėjimo duomenis.

Analogišką išvadą gautume ir pritaikę ω^2 kriterijų.

14. SMIRNOVO KRITERIJUS

Sakykime, stebime du atsitiktinius dydžius ir turime dvi tų dydžių nepriklausomų stebėjimų serijas

$$X_1, ..., X_n,$$
$$Y_1, ..., Y_m.$$

Remiantis stebėjimų duomenimis, reikia patikrinti hipotezę, kad abu dydžiai yra vienodai pasiskirstę. Su tokiais uždaviniais jau susidūrėme, kai abu stebimieji dydžiai turėjo tą patį pasiskirstymo tipą ir reikėjo patikrinti hipotezę, kad pasiskirstymų parametrai yra lygūs. Sprendžiant šį uždavinį, galima buvo taikyti ir χ^2 kriterijų.

Jei apie abiejų dydžių pasiskirstymą žinoma tik tiek, kad jie yra tolydūs, tai galima taikyti Smirnovo[1] kriterijų, su kurio teorija dabar susipažinsime.

Taigi tarkime, kad abu dydžiai turi tą pačią tolydžią pasiskirstymo funkciją $F(u)$. Pažymėkime pirmosios imties empirinę pasiskirstymo funkciją $\mathcal{F}_n(u)$, o antrosios – $\mathcal{G}_m(u)$ ir sudarykime statistikas

$$\mathcal{D}_{nm}^{+} = \sup_{-\infty<u<\infty} \big(\mathcal{F}_n(u) - \mathcal{G}_m(u)\big),$$
$$\mathcal{D}_{nm} = \sup_{-\infty<u<\infty} |\mathcal{F}_n(u) - \mathcal{G}_m(u)|.$$

Pasirodo, kad abiejų tų statistikų pasiskirstymas priklauso tik nuo n ir m, bet nepriklauso nuo F, jei tik ji yra tolydi.

Kad būtų paprasčiau, nagrinėsime tik specialų atvejį $n = m$. Žymėsime

[1] Nikolajus Smirnovas (1900–1966) – rusų matematikas.

$$\mathcal{D}_n^+ = \mathcal{D}_{nn}^+, \quad \mathcal{D}_n = \mathcal{D}_{nn}.$$

1 teorema. *Jei abu stebimieji dydžiai yra tolydūs ir vienodai pasiskirstę, tai sveikiesiems s*

$$\mathcal{P}(n\mathcal{D}_n^+ < s) = \begin{cases} 0, & \text{kai } s \le 0, \\ 1 - \dfrac{\binom{2n}{n-s}}{\binom{2n}{n}}, & \text{kai } 0 < s \le n, \\ 1, & \text{kai } s > n. \end{cases}$$

Į r o d y m a s. Kadangi stebimieji dydžiai yra tolydūs, tai tarp stebėjimo rezultatų $X_1, ..., X_n, Y_1, ..., Y_n$ lygūs galės būti tik su tikimybe 0. Todėl laikysime visus stebėjimo rezultatus skirtingais. Surašysime juos didėjančia tvarka

$$Z_1 < Z_2 < ... < Z_{2n}.$$

Įvesime pagalbinius dydžius $U_1, ..., U_{2n}$, imdami $U_k = 1$, kai Z_k yra iš pirmosios, ir $U_k = -1$, kai Z_k yra iš antrosios imties. Pažymėkime $V_0 = 0, V_k = U_1 + ... + U_k$ $(k = 1, ..., 2n)$. Skaičius $n\big(\mathcal{F}_n(u) - \mathcal{G}_n(u)\big)$ yra imties $X_1, ..., X_n$ elementų, mažesnių už u, ir imties $Y_1, ..., Y_n$ elementų, mažesnių už u, skaičių skirtumas. Jei u prabėga visą realiųjų skaičių tiesę, tai $n(\mathcal{F}_n(u) - \mathcal{G}_n(u))$ pasikeičia tik tada, kai u peržengia reikšmes Z_k $(k = 1, ..., 2n)$; pokytis yra U_k. Todėl

$$n\mathcal{D}_n^+ = \max_{1 \le k \le 2n} n\big(\mathcal{F}_n(Z_k + 0) - \mathcal{G}_n(Z_k + 0)\big) = \max_{1 \le k \le 2n} V_k.$$

Skaičius galimų sekų $U_1, ..., U_{2n}$ yra lygus skaičiui derinių iš $2n$ elementų po n, t. y.

$$\binom{2n}{n}.$$

Kadangi $X_1, ..., X_n, Y_1, ..., Y_n$ yra nepriklausomi ir turi tuos pačius pasiskirstymus, tai kiekviena seka $U_1, ..., U_{2n}$ turi tą pačią tikimybę

$$1 \Big/ \binom{2n}{n}.$$

Reikia rasti skaičių tų sekų $U_1, ..., U_{2n}$, kurioms $\max V_k < s$. Tam pravers geometrinė interpretacija. Plokštumoje (v, t) atidėkime taškus su koordinatėmis (k, V_k) $(k = 0, 1, ..., 2n)$ ir sujunkime juos atkarpomis. Gausime laužtę. Jos pradžia bus taške $(0, 0)$, galas – taške $(2n, 0)$. Atkarpos su abscisių ašimi sudaro $45°$ arba $-45°$ kampus; abiejų rūšių atkarpų bus po n. Reikia rasti skaičių laužčių, nekertančių tiesės $t = s$. Tuo tikslu kiekvieną laužtę,

pasiekiančią tiesę $t = s$, pakeiskime nauja laužte, kuri nuo (0, 0) iki pirmojo susikirtimo su $t = s$ sutampa su pirmykšte laužte, o vėliau yra pastarosios veidrodinis atspindys tos tiesės atžvilgiu. Naujoji laužtė prasidės taške (0, 0) ir baigsis taške $(2n, 2s)$. Joje turi būti $n + s$ pakilimų ir $n - s$ nusileidimų. Todėl tokių laužčių bus

$$\binom{2n}{n-s}.$$

Vadinasi, skaičius pirmykščių laužčių, nepasiekiančių tiesės $t = s$, yra

$$\binom{2n}{n} - \binom{2n}{n-s}. \quad \square$$

2 teorema. *Jei išpildytos* 1 *teoremos sąlygos, tai*

$$\mathcal{P}\Big(\mathcal{D}_n^+ \sqrt{\frac{n}{2}} < z\Big) \to \begin{cases} 0, & \text{kai } z \le 0, \\ 1 - e^{-2z^2}, & \text{kai } z > 0. \end{cases}$$

Į r o d y m a s. Pastebėsime, kad

$$\mathcal{P}\Big(\mathcal{D}_n^+ \sqrt{\frac{n}{2}} < z\Big) = \mathcal{P}(n\mathcal{D}_n^+ < z\sqrt{2n}).$$

Tirdami šios tikimybės asimptotiką, remsimės 1 teorema. Pakanka nagrinėti atvejį, kai $z > 0$. Pažymėkime $s = z\sqrt{2n}$, kai tas skaičius yra sveikasis, ir $s = [z\sqrt{2n}] + 1$, kai jis nėra sveikas. Taigi $s = z\sqrt{2n} + \delta$, $0 \le \delta < 1$. Laikysime, kad n yra didelis.

Iš Stirlingo formulės (žr. I.6 skyrelį), pažymėję

$$p = \frac{\binom{2n}{n-s}}{\binom{2n}{n}} = \frac{(n!)^2}{(n+s)!(n-s)!},$$

gauname

$$\begin{aligned} \ln p = (2n+1)\ln n - \Big(n + s + \frac{1}{2}\Big)\ln(n+s) - \\ - \Big(n - s + \frac{1}{2}\Big)\ln(n-s) + O\Big(\frac{1}{n}\Big). \end{aligned} \tag{1}$$

Kadangi pagal 12 skyrelio lemą

$$\ln(n+s) = \ln n + \ln\Big(1+\frac{s}{n}\Big) = \ln n + \frac{s}{n} - \frac{s^2}{2n^2} + O\Big(\frac{s^3}{n^3}\Big),$$
$$\ln(n-s) = \ln n - \frac{s}{n} - \frac{s^2}{2n^2} + O\Big(\frac{s^3}{n^3}\Big),$$

tai iš (1), atlikę paprastus skaičiavimus, gauname

$$\ln p = -\frac{s^2}{n} + O(n^{-1/2}) = -2z^2 + O(n^{-1/2}).$$

Vadinasi,

$$\mathcal{P}\Big(\mathcal{D}_n^+\sqrt{\frac{n}{2}} < z\Big) = 1 - \exp\{-2z^2 + O(n^{-1/2})\} \to 1 - e^{-2z^2},$$

kai $n \to \infty$. □

3 teorema. *Jei teisingos* 1 *teoremos sąlygos, tai sveikiesiems* s

$$\mathcal{P}(n\mathcal{D}_n < s) = \begin{cases} 0, & \text{kai } s \le 1, \\ \dfrac{1}{\binom{2n}{n}} \displaystyle\sum_{k=-[n/s]}^{[n/s]} \binom{2n}{n-ks}, & \text{kai } 1 < s \le n, \\ 1, & \text{kai } s > n. \end{cases}$$

Į r o d y m a s panašus į 1 teoremos įrodymą. Vartosime vėl tuos pačius žymėjimus ir remsimės ta pačia geometrine interpretacija. Šį kartą ieškosime skaičiaus N_0 laužčių, kurios telpa tarp tiesių $t = -s$ ir $t = s$, jų nepasiekdamos.

Kaip žinome, laužčių yra iš viso

$$N = \binom{2n}{n}.$$

Ieškomąjį laužčių skaičių N_0 gausime, atmetę iš N skaičių laužčių, kurios turi bendrų taškų su tiesėmis $t = -s, t = s$. Pirmiausia atmesime skaičių $N(+)$ laužčių, kurios turi bendrą tašką su $t = s$, ir skaičių $N(-)$ laužčių, kurios turi bendrą tašką su $t = -s$. Tačiau tada bus du kartus atmestos laužtės, turinčios bendrų taškų su abiem tiesėm. Todėl pridėsime skaičių $N(+,-)$ laužčių, kurios po bendro taško su $t = s$ turi bendrą tašką su $t = -s$, ir skaičių $N(-,+)$ laužčių, kurios po bendro taško su $t = -s$ turi bendrą tašką su $t = s$. Kai kurios laužtės bus pridėtos du kartus. Todėl tęsime šiuos samprotavimus. Vartodami lengvai suvokiamus žymėjimus, gausime

$$N_0 = N - N(+) - N(-) + N(+,-) + N(-,+) - N(+,-,+) -$$
$$- N(-,+,-) + ...$$

1 teoremos įrodyme gavome

$$N(+) = \binom{2n}{n-s}.$$

Skaičius $N(-) = N(+)$. Tai išplaukia iš simetriškumo.

Apskaičiuosime $N(+,-)$. Kiekvieną laužtę, išeinančią iš taško (0, 0) ir pasiekiančią tiesę $t = s$, pakeisime nauja laužte, kuri sutampa su pirmykšte nuo (0, 0) iki bendro taško su $t = s$, o toliau yra jos veidrodinis atspindys tiesės $t = s$ atžvilgiu. Taip gauta laužtė pasibaigs taške $(2n, 2s)$. Jei pirmykštė laužtė iš pradžių turi bendrą tašką su $t = s$, o po to su $t = -s$, tai ką tik gauta naujoji laužtė pasieks tiesę $t = 3s$. Konstruojame dar vieną laužtę, kuri sutaps su naująja iki bendro taško su $t = 3s$, o po to sutaps su pirmykštės laužtės veidrodiniu atspindžiu tiesės $t = 3s$ atžvilgiu. Vadinasi, naujausioji laužtė baigsis taške $(2n, 4s)$. Tokių naujausių laužčių yra

$$\binom{2n}{n-2s}.$$

Taigi

$$N(+,-) = N(-,+) = \binom{2n}{n-2s} = \binom{2n}{n+2s}.$$

Analogiškai samprotaudami, galime gauti

$$N(\varepsilon_1, \varepsilon_2, ..., \varepsilon_k) = \binom{2n}{n-ks} = \binom{2n}{n+ks},$$

kai $\varepsilon_1, ..., \varepsilon_k$ yra alternuojanti + ir − seka.

Galutinai

$$N_0 = \sum_{k=-[n/s]}^{[n/s]} (-1)^k \binom{2n}{n-ks}. \ \square$$

4 teorema. *Jei teisingos* 1 *teoremos sąlygos, tai*

$$\mathcal{P}\Big(\mathcal{D}_n \sqrt{\frac{n}{2}} < z\Big) \to K(z),$$

kai $n \to \infty$*; čia* $K(z)$ *yra* 13.3 *teoremoje apibrėžta funkcija.*

Į r o d y m a s. Pakanka nagrinėti atvejį, kai $z > 0$. Toliau z laikysime fiksuotu, o n – pakankamai dideliu. Atkreipsime dėmesį, kad

$$\mathcal{P}\Big(\mathcal{D}_n \sqrt{\frac{n}{2}} < z\Big) = \mathcal{P}(n\mathcal{D}_n < z_n);$$

čia $z_n = z\sqrt{2n}$, kai pastarasis skaičius yra sveikasis, ir $z_n = [z\sqrt{2n}] + 1$, kai jis nėra sveikasis. Tada

$$\mathcal{P}\Big(\mathcal{D}_n\sqrt{\frac{n}{2}} < z\Big) = \sum_{k=-[n/z_n]}^{[n/z_n]} (-1)^k \frac{\binom{2n}{n-kz_n}}{\binom{2n}{n}} =$$

$$= \sum_{k=-[n/z_n]}^{[n/z_n]} (-1)^k \frac{(n!)^2}{(n-kz_n)!(n+kz_n)!}.$$

Paėmę bet kokį $\varepsilon > 0$, parinkime tokį sveikąjį teigiamą n_0, kad būtų

$$e^{-2n_0^2z^2} < \frac{\varepsilon}{16}, \quad \Big|\sum_{|k|>n_0} (-1)^k e^{-2k^2z^2}\Big| < \frac{\varepsilon}{6}.$$

Kaip ir 2 teoremos įrodyme, iš Stirlingo formulės gauname (skaičiavimus paliekame skaitytojui)

$$\frac{(n!)^2}{(n-kz_n)!(n+kz_n)!} = e^{-2k^2z^2}\big(1+o(1)\big)$$

tolygiai visiems k su sąlyga $|k| \le n_0$. Todėl

$$\Big|\sum_{k=-n_0}^{n_0} (-1)^k e^{-2k^2z^2} - \sum_{k=-n_0}^{n_0} (-1) \frac{(n!)^2}{(n-kz_n)!(n+kz_n)!}\Big| =$$

$$= o(1) \sum_{k=-n_0}^{n_0} e^{-2k^2z^2} < \frac{\varepsilon}{2}$$

pakankamai dideliems n. Kadangi

$$\binom{2n}{n-kz_n} > \binom{2n}{n-(k+1)z_n},$$

tai

$$\Big|\sum_{n_0<|k|\le z_n} (-1)^k \frac{(n!)^2}{(n-kz_n)!(n+kz_n)!}\Big| <$$

$$< \frac{4(n!)^2}{(n-n_0z_n)!(n+n_0z_n)!} = 4e^{-2n_0^2z^2}\Big(1+o(1)\Big) < \frac{\varepsilon}{3},$$

kai n yra pakankamai didelis. Todėl pakankamai dideliems n

$$\Big|\mathcal{P}\Big(\mathcal{D}_n\sqrt{\frac{n}{2}}<z\Big)-K(z)\Big|\le\Big|\sum_{k=-n_0}^{n_0}(-1)^k e^{-2k^2z^2}-$$
$$-\sum_{k=-n_0}^{n_0}(-1)^k\frac{(n!)^2}{(n-kz_n)!(n+kz_n)!}\Big|+$$
$$+\Big|\sum_{n_0<|k|\le z_n}(-1)^k\frac{(n!)^2}{(n-kz_n)!(n+kz_n)!}\Big|+$$
$$+\Big|\sum_{|k|>n_0}(-1)^k e^{-2k^2z^2}\Big|<\frac{\varepsilon}{2}+\frac{\varepsilon}{3}+\frac{\varepsilon}{6}=\varepsilon.\ \square$$

Pateiksime be įrodymo statistikų $\mathcal{D}^+_{nm}$ ir $\mathcal{D}_{nm}$ pasiskirstymo ribines teoremas, kai n ir m yra bet kokie.

5 teorema. *Jei stebimieji dydžiai yra tolydūs ir vienodai pasiskirstę, tai*

$$\mathcal{P}\Big(\mathcal{D}^+_{nm}\sqrt{\frac{nm}{n+m}}<z\Big)\to\begin{cases}0, & kai\ z\le 0,\\ 1-e^{-2z^2}, & kai\ z>0,\end{cases},$$
$$\mathcal{P}\Big(\mathcal{D}_{nm}\sqrt{\frac{nm}{n+m}}<z\Big)\to K(z),$$

kai $n\to\infty,\ m\to\infty$*; čia* $K(z)$ *yra* 13.3 *teoremoje apibrėžta funkcija.*

Statistikų $\mathcal{D}^+_{nm}$ ir $\mathcal{D}_{nm}$ pasiskirstymu grindžiamas Smirnovo kriterijus tikrinti hipotezei, kad abu stebimi tolydieji atsitiktiniai dydžiai turi tą patį pasiskirstymą. Tų statistikų pasiskirstymai yra tabuliuoti (žr. [17], XVII lentelę; [2], 6.5^a lentelę). Skaičiavimams suprastinti naudojamos formulės

$$\mathcal{D}^+_{nm}=\max_{1\le k\le n}\Big(\frac{k}{n}-\mathcal{G}_m(X^*_k)\Big)=\max_{1\le k\le m}\Big(\mathcal{F}_n(Y^*_k)-\frac{k-1}{m}\Big),$$
$$\mathcal{D}^-_{nm}=\max_{1\le k\le n}\Big(\mathcal{G}_m(X^*_k)-\frac{k-1}{n}\Big)=\max_{1\le k\le m}\Big(\frac{k}{m}-\mathcal{F}_n(Y^*_k)\Big),$$
$$\mathcal{D}_{nm}=\max(\mathcal{D}^+_{nm},\mathcal{D}^-_{nm});$$

čia $\{X^*_k\}$ yra pirmojo stebimo dydžio, o $\{Y^*_k\}$ – antrojo dydžio variacinės sekos.

15. ŽENKLŲ KRITERIJUS

11 skyrelyje nagrinėjome šitokį uždavinį. Stebėjome atsitiktinį dydį, įgyjantį dvi reikšmes 1 ir 0 atitinkamai su tikimybėmis p ir $1-p$. Tikimybė p buvo nežinoma. Nurodėme metodą tikrinti hipotezei, kad p yra koks nors konkretus skaičius $p_0\in(0,1)$. Dabar mums pravers specialus atvejis $p_0=1/2$. Priminsime tą kriterijų.

Atliekame n nepriklausomų stebėjimų. Pažymėkime κ_n skaičių atvejų, kai stebimasis atsitiktinis dydis įgyja reikšmę 1. Statistika κ_n turi binominį pasiskirstymą

$$\mathcal{P}(\kappa_n = k) = \binom{n}{k}\left(\frac{1}{2}\right)^n \quad (k = 0, 1, ..., n).$$

Nerandomizuotas kriterijus sudaromas šitaip. Imkime reikšmingumo lygmenį α. Pažymėkime k_n statistikos κ_n realizaciją. Jei alternuojanti hipotezė yra $p = p_1 < 1/2$, tai hipotezę atmetame, kai

$$\sum_{k=0}^{k_n}\binom{n}{k}\left(\frac{1}{2}\right)^n \le \alpha.$$

Jei alternatyva yra $p = p_1 > 1/2$, tai hipotezę atmetame, kai

$$\sum_{k=k_n}^{n}\binom{n}{k}\left(\frac{1}{2}\right)^n \le \alpha.$$

Pagaliau, jei alternatyva yra $p = p_1 \neq 1/2$, tai hipotezę atmetame, kai

$$\sum_{k=0}^{k_n}\binom{n}{k}\left(\frac{1}{2}\right)^n \le \frac{\alpha}{2}$$

arba

$$\sum_{k=k_n}^{n}\binom{n}{k}\left(\frac{1}{2}\right)^n \le \frac{\alpha}{2}.$$

Šiuo kriterijumi yra pagrįstas vadinamasis *ženklų kriterijus*, vienas iš paprasčiausių statistikoje. Panagrinėsime porą jo taikymo atvejų.

1. Sakykime, stebime tolydųjį atsitiktinį dydį. Reikia patikrinti hipotezę, kad jo mediana yra lygi z_0. Kaip paprastai, tarkime, kad $X_1, ..., X_n$ yra atsitiktinė imtis. Kadangi stebimasis dydis yra tolydus, tai jo mediana tenkina sąlygą

$$P(X_1 < z_0) = P(X_1 > z_0) = \frac{1}{2}.$$

Pažymėkime κ_n skaičių tų k, kuriems $X_k - z_0$ yra neigiamas. Ši statistika yra pasiskirsčiusi pagal binominį dėsnį. Todėl hipotezei tikrinti galima taikyti anksčiau aprašytą procedūrą. Praktiškai ją taikant, tenka suskaičiuoti, kiek yra neigiamų ir kiek yra teigiamų skirtumų $X_k - z_0$. Iš čia ir kilęs ženklų kriterijaus pavadinimas.

2. Sakykime, stebime dvimatį tolydųjį atsitiktinį dydį su tankio funkcija $p(u, v)$. Reikia patikrinti hipotezę, kad $p(u, v) = p(v, u)$. Jei stebimojo dydžio

komponentai yra nepriklausomi, tai tikrinamoji hipotezė reiškia, kad jie yra vienodai pasiskirstę. Imkime n nepriklausomų stebėjimų $(X_1, Y_1), ..., (X_n, Y_n)$. Jei hipotezė yra teisinga, tai skirtumai $X_k - Y_k$ įgyja teigiamas ir neigiamas reikšmes su ta pačia tikimybe $1/2$. Todėl galime taikyti ženklų kriterijų.

16. RANGINIAI KRITERIJAI

Sakykime, tiriame du tolydžiuosius atsitiktinius dydžius. Jų nepriklausomų stebėjimų rezultatai yra $X_1, ..., X_n$ ir $Y_1, ..., Y_m$. Reikia patikrinti hipotezę, kad tie dydžiai turi tą pačią pasiskirstymo funkciją.

Sujunkime visus stebėjimo rezultatus ir iš jų sudarykime vieną variacinę seką. Joje bus $N = n + m$ narių. Praleiskime indeksus. Gausime, pavyzdžiui, šitokio tipo seką

$$\begin{array}{cccccccccc} X & X & Y & X & Y & Y & \dots & X & Y & Y. \\ 1 & 2 & 3 & 4 & 5 & 6 & & \text{N-2} & \text{N-1} & \text{N} \end{array}$$

Apačioje surašyti narių numeriai – jų rangai. Tiesa, gali pasitaikyti vienodų pirmosios ir antrosios imties narių. Tada gautoji seka nebus vienareikšmiškai nusakyta. Tačiau tokie atvejai galės pasitaikyti tik su tikimybe 0, nes dydžiai yra tolydieji. Jei vis dėlto taip atsitiko, tai lygius stebėjimo rezultatus išdėstome bet kaip.

Tarkime, kad $R_1 < R_2 < ... < R_n$ yra X numeriai. Parenkame kokią nors funkciją $f(r)$, apibrėžtą visiems $r = 1, ..., N$, ir imame statistiką

$$W = f(R_1) + ... + f(R_n).$$

Atitinkamai konkretizavę funkciją $f(r)$, gauname Vilkoksono ir Van der Vardeno kriterijus.

1. V i l k o k s o n o[1] k r i t e r i j u s. Tarkime, kad $s(1), s(2), ..., s(N)$ yra skaičiai $1, 2, ..., N$, surašyti kokia nors iš anksto fiksuota (nepriklausančia nuo stebėjimo rezultatų) tvarka, kitaip tariant, $s(1), s(2), ..., s(N)$ yra skaičių $1, 2, ..., N$ kėlinys. Paėmę $f(r) = s(r)$, turime statistiką

$$W = f(R_1) + ... + f(R_n).$$

Galima įrodyti: jei hipotezė yra teisinga, tai taip sudarytos statistikos pasiskirstymas nepriklauso nuo teorinės pasiskirstymo funkcijos ir funkcijos $s(r)$, priklauso tik nuo n ir m. Statistikos W pasiskirstymo funkcija yra tabuliuota (žr. [2], 6.8 lentelę). Kai n ir m dideli, galima naudotis asimptotine formule

[1] F. Wilcoxon – anglų matematikas.

$$\mathcal{P}\left(\frac{W-\dfrac{m}{2}(m+n+1)}{\sqrt{\dfrac{mn}{12}(m+n+1)}}<u\right)\to\Phi(u),$$

kai $N=n+m\to\infty$.

2. V a n d e r V a r d e n o[1] k r i t e r i j u s yra panašus į Vilkoksono, tik funkcija $f(r)$ parenkama šiek tiek kitaip:

$$f(r)=\Phi^{-1}\Big(\frac{s(r)}{n+m+1}\Big);$$

čia Φ^{-1} reiškia funkciją, atvirkštinę standartinei normaliajai pasiskirstymo funkcijai Φ. Ir šiuo atveju statistikos

$$W=\sum_{k=1}^{n}\Phi^{-1}\Big(\frac{s(R)}{n+m+1}\Big)$$

pasiskirstymas priklauso tik nuo n ir m. Kai $N\to\infty$,

$$\mathcal{P}(W<u\sqrt{DW})\to\Phi(u);$$

čia

$$DW=\frac{nm}{n+m+1}\ \frac{1}{n+m}\sum_{k=1}^{n+m}\left(\Phi^{-1}\Big(\frac{k}{n+m+1}\Big)\right)^2.$$

Statistikos W pasiskirstymo funkcija yra tabuliuota (žr. [17], XIX lentelę, [2], 6.9 lentelę).

Kriterijus, pagrįstas šia statistika, yra gana tikslus, kai stebimieji atsitiktiniai dydžiai pasiskirstę pagal normalųjį dėsnį arba artimą jam.

[1] Van der Waerden (g. 1903) – olandų matematikas, šiuo metu dirbąs Šveicarijoje.