Հետազոտական Աշխատանք

Ֆակուլտետ՝ Կիրառական վիճակագրություն և տվյալագիտություն

Կուրս՝ I, մագիստրատուրա

Առարկա՝ Կիրառական վիճակագրություն

Թեմա՝ Ոչ պարամետրական թեստեր

Դասախոս՝ Միքայել Պողոսյան

Ուսանողներ՝ Ավետիսյան Մարիամ, Հարությունյան Իննա, Հունանյան Հարություն

Երևան, 2025

ԲՈՎԱՆԴԱԿՈՒԹՅՈՒՆ

ՆԵՐԱԾՈՒԹՅՈՒՆ

Վիճակագրական թեստը մեթոդաբանական գործիք է, որը նախատեսված է գնահատելու համար՝ արդյոք դիտարկված տվյալները բավարար հիմքեր են տալիս մերժելու համախմբության մասին ձևակերպված որոշակի հիպոթեզը՝ օգտվելով նմուշառված տվյալներից (sample data): Վիճակագրական թեստերը ունեն լայն կիրառություններ գիտության տարբեր ոլորտներում, օրինակ, երբ անհրաժեշտ է գնահատել՝ արդյոք նոր դեղամիջոցը ավելի արդյունավետ է քան արդեն գոյություն ունեցողը, արդյոք երաժշտության ներքո աշխատելն ավելի պրոդուկտիվ է քան լռության պայմաններում, արդյոք հաստոցի արտադրած խոտանի ծավալը միջինում չի գերազանցում առավելագույն շեմը, և այլն:

Վիճակագրական թեստերը կարելի է բաժանել երկու հիմնական խմբի՝ պարամետրական և ոչ-պարամետրական: Պարամետրական թեստերը հիմնվում են համախմբության բաշխման մասին խիստ ենթադրությունների վրա. տվյալները կարելի է մոդելավորել նախօրոք ընտրված բաշխմամբ՝ անհայտ պարամետրով (գրեթե բոլոր պարամետրական թեստերը ենթադրում են, որ տվյալները գալիս ես նորմալ բաշխումից): Չնայած պարամետրական թեստերի լայն կիրառությանը՝ հաճախ իրական տվյալները պարունակում են շեղվածություն (skewness), միջակայքից դուրս ընկած կետեր (outliers), կամ հետերոսկեդաստիկություն, որոնք թույլ չեն տալիս օգտագործել պարամետրական թեստեր ինչպիսիք են t-test-ը [1] կամ ANOVA-ն: Այդ դեպքերում կիրառվում են ոչ-պարամերական թեստերը, որոնք կառուցված են համախմբության բաշխման վերաբերյալ նվազագույն ենթադրությունների վրա: Ոչ-պարամետրական թեստերի առավելություններից է նաև բավականին պարզ լինելը, քանզի ոչ-պարամետրական թեստերը հիմնված են ռանկավորման կամ որոշակի այլ պարզ գործողությունների հաշվման վրա՝ երբեմն բարդ մաթեմատիկական մոդելների փոխարեն: Ոչ պարամետրական թեստերը արդյունավետ են նաև քիչ քանակությամբ դիտարկումների, կարգային (ordinal) և անվանական (nominal) տվյալների վերլուծության համար:

Այս աշխատանքի հիմնական նպատակը որոշ ոչ պարամետրական թեստերի տեսական հիմքերի և կիրառական առանձնահատկությունների ուսումնասիրությունն է:

Գլուխ 1. Ոչ պարամետրական վիճակագրության հիմունքներ

§1. Վիճակագրական թեստերի հիմնական սահմանումներ

Վիճակագրական թեստ կառուցելու համար նախ անհրաժեշտ է սահմանել վարկած՝ հիպոթեզ: Վիճակագրական թեստավորման ժամանակ դիտարկվում են զրոյական և ալտերնատիվ հիպոթեզները: Ալտերնատիվ հիպոթեզն իրենից ներկայացնում է այնպիսի պնդում, որին ցանկանում ենք հասնել զրոյական հիպոթեզի մերժմամբ: Նպատակն է գնահատել, թե արդյոք դիտարկված տվյալները վիճակագրորեն նշանակալի են՝ մերժելու համար գործող զրոյական հիպոթեզը և նախապատվություն տալու փորձարկողի կողմից առաջարկվող այլընտրանքային վարկածին [2]:

Քանի որ հետազոտությունը կատարվում է միայն նմուշառված տվյալների վրա, հավանական է, որ թեստի հիման վրա ստացված եզրակացությունը լինի սխալ: Սխալները կարող են լինել 2 տեսակի՝

- I տեսակի սխալ՝ ժխտել զրոյական հիպոթեզը, երբ այն իրականում ճիշտ է (false positive),
- II տեսակի սխալ՝ չժխտել զրոյական հիպոթեզը, երբ այն իրականում սխալ է (false negative):

Թեստի Նշանակալիության Մակարդակ (Significance Level, α) կանվանենք զրոյական վարկածը մերժելու հավանականությունը այն դեպքում երբ այն ճիշտ է.

$$P(reject\ H_0 |\ H_0\ is\ true)\ =\ \alpha$$

Թեստի հզորություն (Power) կանվանենք հավանականությունը մերժելու H_0-ն, երբ այն իրոք սխալ է, այսինքն՝ ճիշտ եզրակացություն անելու հավանականությունը:

$$Power\ =\ P(reject\ H_0 | H_0\ is\ false)\ =\ 1\ -\ Type\ II\ error\ =\ 1\ -\ \beta$$

§2. Ոչ-պարամետրական թեստերի անհրաժեշտությունն ու առավելությունները

Ինֆերենցիալ վիճակագրության մեջ, պարամետրական թեստերի գերակշիռ մասը կիրառելի են միայն այն դեպքում երբ բավարարված են ստորև բերված ենթադրությունները.

1. Նորմալ կամ ասիմպտոտիկ նորմալ բաշխված տվյալներ:
2. Ընտրանքի հոմոսկեդաստիկություն:
3. Ընտրանքի տվյալների անկախություն [3]:

Միևնույն ժամանակ, ոչ-պարամետրական թեստերը ունեն մի շարք առավելություններ.

1. Չեն պահանջում ընտրանքի բաշխման հստակ պարամետրական օրենք:
2. Հիմնված են համախմբությունների հատկությունների վերաբերյալ նվազագույն ենթադրությունների վրա:
3. Շատ դեպքերում կիրառման տեսակետից պարզ ու գործնական են՝ համեմատությամբ պարամետրական մեթոդների:
4. Ավելի քիչ զգայուն են միջակայքից դուրս ընկած կետերի (outliers) նկատմամբ:

Այնուամենայնիվ, ոչ-պարամետրական թեստերը սովորաբար պակաս ճշգրիտ են քան պարամետրական այլընտրանքները և հնարավոր դեպքերում նախընտրելի է կիրառել պարամետրական թեստեր, չնայած նրան, որ ոչ պարամետրական թեստերը կիրառելի են գրեթե բոլոր դեպքերում [3]:

§3. Ոչ-պարամետրական թեստերի տեսակներ

Ըստ ընտրանքների քանակի և թեստի բնույթի՝ ոչ-պարամետրական թեստերը կարելի բաժանել խմբերի:

1. Մեկ ընտրանքով խնդիրներ (One Sample Problems)
   ա) Մեդիանի վրա հիմնված ոչ պարամետրական թեստեր
      i. Նշանի թեստ (Sign Test)
      ii. Ուիլկոքսոն նշանով ռանկ թեստ մեկ նմուշի համար (Wilcoxon One-Sample Signed-Rank Test)
   բ) Ընտրանքի բաշխման համապատասխանության թեստեր
      iii. Կոլմոգորով - Սմիրնովի՝ 1 ընտրանու թեստ (One Sample Kolmogorov - Smirnov Test)
      iv. Շապիրո - Ուիլկի թեստ (Shapiro–Wilk test)
      v. $\chi 2$ - թեստ (Chi-squared goodness-of-fit test)
   գ) Ընտրանու տվյալների անկախության կամ պատահականության թեստեր
      vi. Ուալդ - Ուոլֆովիցի թեստ (Wald–Wolfowitz runs test)
      vii. Շրջման կետի թեստ (The turning point test for randomness of fluctuations)
2. Երկու ընտրանքով խնդիրներ (Two Sample Problems)
   ա) Բաշխման հավասարության թեստեր (Tests for Equality of Distributions)
      i. Կոլմոգորով - Սմիրնովի՝ 2 ընտրանու թեստ (Two-Sample Kolmogorov–Smirnov Test)
   բ) Ընտրանքների կենտրոնական վիճականիների հավասարության թեստ
      ii. Wilcoxon Rank-Sum Test
   գ) Զույգված տվյալների համար նախատեսված թեստեր (Tests for Paired or Matched Samples)
      iii. Wilcoxon Signed-Rank Paired Test

§4. Տվյալների ռանկավորման գործընթացներ

Ոչ-պարամետրական հետազոտության ժամանակ, այդ թվում՝ ոչ-պարամետրական թեստեր կառուցելիս հաճախ օգտագործում են տվյալների դասավորության ինֆորմացիան: Դիցուք $\{x_1, x_2, ..., x_n\}$ -ը n դիտարկումների բազմություն է: Ռանկն իրենից ներկայացնում է յուրաքանչյուր դիտարկման համապատասխանող իրական թիվ, որը ցույց է տալիս դիտարկման դիրքը կարգավորված բազմության մեջ: Տվյալների ռանկավորումն իրականացվում է հետևյալ կերպ.

1. կարգավորել տվյալները աճման կարգով՝ $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$,

2. x_i-ի ռանկը նշանակել $R(x_i)$, որը հավասար կլինի x_i-ի դիրքաթվին՝ կարգավորած բազմության մեջ՝ $R(x_i) = j$, եթե $x_i = x_{(j)}$,
3. եթե երկու և ավելի դիտարկումներ ունեն նույն արժեքը (անվանում են tied), ապա դրանց վերագրվում է կարգավորված բազմության համապատասխան դիրքաթվերի միջինը:

Եթե ընտրանքները գալիս են տարբեր բաշխումներից և ռանկավորման արդյունքում մեդիաններին համապատասխանող ռանկերը իրարից նշանակալի չափով շեղված են, ապա փոքր մեդիան ունեցող համախմբությունը կձգտի ունենալ ավելի փոքր ռանկեր՝ ի տարբերություն մեծ մեդիան ունեցողի:

Գլուխ 2. Ոչ պարամետրական մեկ ընտրանքի թեստեր

§1. Ուիլկոքսոն նշանով ռանկի թեստ՝ մեկ ընտրանքի դեպքում (Wilcoxon signed rank test)

Մեկ ընտրանքի դեպքում Ուիլկոքսոն նշանով ռանկի թեստը համեմատում է համախմբության մեդիանը (μ) հիպոթետիկ մեդիանի հետ (μ_0):

Ունենք IID դիտարկումներից պատահական $Y_1, Y_2, \ldots Y_n$ ընտրանքը: Ենթադրենք դիտարկումների խտության ֆունկցիան սիմետրիկ է ինչ-որ առանցքի նկատմամբ՝ $f(\mu_0 - x) = f(\mu_0 + x), \forall x$: Ակնհայտ է, որ այդ դեպքում μ_0-ն կլինի այդ բաշխման մեդիանը: Նպատակն է պարզել՝ արդյոք տրված բաշխման մեդիանը համընկնում է հիպոթետիկ μ_0 մեդիանի հետ՝ H_0: $\mu = \mu_0$:

Նկատենք, որ $D_i = Y_i - \mu_0$ պատահական մեծության մեդիանը 0 է, և հավանականային բաշխման ֆունկցիան սիմետրիկ է դրա նկատմամբ:

Եթե $\mu_0 > 0$, կանխատեսելի է, որ դիտարկումների մեծ մասը կլինի դրական, ինչպես նաև շատ ավելի հեռու կլինի 0-ից՝ համեմատած դրանց համապատասխան μ_0-ի նկատմամբ սիմետրիկ դիտարկումների:

Կատարենք հետևյալ նշանակումները.

$$\Psi_i = I(D_i > 0),$$

որտեղ I-ին ինդիկատոր ֆունկցիա է (I=1 երբ D_i >0, և 0՝ հակառակ դեպքում), R_i -ն $|D_i|$-ի ռանկն է (ըստ գլուխ 1.4-ում նկարագրված մեթոդի), $i = 1, 2, \ldots, n$:

Որպես թեստ վիճականի վերցնենք դրական նշանով ռանկերի գումարը՝

$$T^+ = \sum_{i=1}^{n} \Psi_i R_i$$

Սա նույնն է, ինչ μ_0 -ից մեծ տվյալների ռանկերի գումարը:

Սկզբում դիտարկենք այն դեպքը, երբ ռանկերը չեն կրկնվում, այսինքն ընդունում են $1, 2, \ldots, n$ արժեքները:

H_0-ի տեղի ունենալու պայմանում, $\Psi = [\Psi_1, \Psi_2, \cdots, \Psi_n]^T$ և $R = [R_1, R_2, \cdots, R_n]^T$ վեկտորները անկախ են [4]: Այդ դեպքում, քանի որ $\Psi_i \sim Bernoulli(0.5)$ T^+-ի մոմենտ գեներացնող ֆունկցիան՝ $M_{T^+}(t)$, կլինի.

$$M_{T^+}(t) = E[e^{tT^+}] = E[e^{t\sum_{i=1}^{n}\Psi_i R_i}] = \prod_{i=1}^{n} E[e^{t\Psi_i R_i}] = \frac{1}{2^n}\prod_{i=1}^{n}(1 + e^{ti}),$$

քանի որ $E[e^{t\Psi_i R_i}] = \frac{1}{2}\cdot e^{tR_i} + \frac{1}{2}\cdot 1 = \frac{1+e^{tR_i}}{2}$:

Օգտվելով մոմենտ գեներացնող ֆունկցիայի $E[X^n] = M_X^{(n)}(0)$ հատկությունից (որտեղ X-ը պատահական մեծություն է) և $M'_{T^+}(t) = \sum_{i=1}^{n} \frac{R_i e^{tR_i}}{1+e^{tR_i}} \cdot M_{T^+}(t)$ առնչությունից, H_0 -ի ճշմարտացիության ներքո T^+-ի մաթ.սպասումը կլինի.

$$E_{H_0}[T^+] = M'_{T^+}(0) = \sum_{i=1}^{n} \frac{R_i e^{0\cdot R_i}}{1+e^{0\cdot R_i}} \cdot M_{T^+}(0) = \sum_{i=1}^{n} \frac{R_i}{2} = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4},$$

իսկ վարիացիան՝ H_0 -ի ճշմարտացիության պայմաններում կստացվի.

$$M''_{T^+}(0) = \sum_{i=1}^{n} \frac{2R_i^2 - R_i^2}{4} + \frac{n(n+1)}{4}\sum_{i=1}^{n} \frac{R_i}{2} = \frac{n(n+1)(2n+1)}{24} + \frac{n^2(n+1)^2}{16} = \frac{n(n+1)(3n^2+7n+2)}{48}$$

$$Var_{H_0}[T^+] = E_{H_0}[(T^+)^2] - (E_{H_0}[T^+])^2 = \frac{n(n+1)(3n^2+7n+2)}{48} - \frac{n^2(n+1)^2}{16} = \frac{n(n+1)(2n+1)}{24}:$$

Երբ տվյալների ծավալը բավականաչափ մեծ չէ, հիպոթեզի ստուգման համար օգտվում են ճշգրիտ բաշխումից, հակառակ դեպքում հաշվի է առնվում այն հանգամանքը, որ H_0 -ի ներքո T^+-ի ասիմպտոտիկ բաշխումը նորմալ է, այսինքն.

$$T^* = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \to^D N(0, 1), \text{ երբ } n \to \infty:$$

Հետևաբար թեստի ստուգման համար կարելի է օգտագործել ստանդարտ նորմալ բաշխման կրիտիկական կետերի արժեքները:

Եթե գոյություն ունի tied դիտարկումների s հատ խումբ, ապա T^+ թեստ վիճականու վարիացիան կլինի ավելի փոքր [5].

$$Var_{H_0}[T^+] = \frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{i=1}^{s} t_i(t_i - 1)(t_i + 1)\right],$$

որտեղ t_i-ն i-րդ խմբի դիտարկումների քանակն է, $i = 1, ..., s$:

Նկատենք, որ այս դեպքում նույնպես տեղի ունի հետևյալ առնչությունը.

$$T^* = \frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{1}{24}\cdot\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{i=1}^{s} t_i(t_i-1)(t_i+1)\right]}} \to^D N(0, 1), \text{ երբ } n \to \infty:$$

H_0-ի ճշմարտացիության պայմաններում, երբ $n \to \infty$, T^* -ի բաշխումը ասիմպտոտիկ $N(0, 1)$ է [6]: Հաշվի առնելով այդ հանգամանքը.

ա) երբ H_0: $\mu = \mu_0$, H_1: $\mu > \mu_0$, α վստահության մակարդակի դեպքում կմերժենք H_0 -ն եթե $T^* > z_{1-\alpha}$:

բ) երբ H_0: $\mu = \mu_0$, H_1: $\mu < \mu_0$, α վստահության մակարդակի դեպքում կմերժենք H_0 -ն եթե $T^* < z_{\alpha}$:

գ) Երկկողմանի թեստի համար (H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$), α վստահության մակարդակի դեպքում կմերժենք H_0 -ն եթե $|T^*| > z_{1-\alpha/2}$։

§2. Ուիլկոքսոն նշանով ռանկի թեստի կիրառություններ

Դիցուք տրված է 30 հոգու դիաբետի հայտնաբերման տարիքի մասին տվյալներ: Ենթադրելով, որ դիաբետի առաջացման տարիքը տրվում է սիմետրիկ հավանականային խտությամբ բաշխումից, անհրաժեշտ է պարզել՝ արդյոք բաշխման մեդիանը (μ) տարբերվում է 45-ից.

Հիպոթեզներ.

H_0: $\mu = 45$,

H_1: $\mu \neq 45$:

Դիտարկում	Հայտնաբերման տարիք	Տարբերությունը մեդիանից	Բացարձակ տարբերություն	Նշանով ռանկ
1	35.5	-9.5	9.5	-22
2	44.5	-0.5	0.5	-2
3	39.8	-5.2	5.2	-13
4	33.3	-11.7	11.7	-24
5	51.4	6.4	6.4	16
6	51.3	6.3	6.3	15
7	30.5	-14.5	14.5	-25
8	48.9	3.9	3.9	10
9	42.1	-2.9	2.9	-9
10	40.3	-4.7	4.7	-11
11	46.8	1.8	1.8	5
12	38	-7	7	-17
13	40.1	-4.9	4.9	-12
14	36.8	-8.2	8.2	-20
15	39.3	-5.7	5.7	-14
16	65.4	20.4	20.4	30

17	42.6	-2.4	2.4	-8
18	42.8	-2.2	2.2	-6
19	59.8	14.8	14.8	26
20	52.4	7.4	7.4	19
21	26.2	-18.8	18.8	-29
22	60.9	15.9	15.9	27
23	45.6	0.6	0.6	3
24	27.1	-17.9	17.9	-28
25	47.3	2.3	2.3	7
26	36.6	-8.4	8.4	-21
27	55.6	10.6	10.6	23
28	45.1	0.1	0.1	1
29	52.2	7.2	7.2	18
30	43.5	-1.5	1.5	-4

$$T^+ = 16 + 15 + 10 + 5 + 30 + 26 + 19 + 27 + 3 + 7 + 23 + 1 + 18 = 200,$$

$$E_{H_0}[T^+] = \frac{30 \cdot 31}{4} = 232.5,$$

$$\sqrt{Var_{H_0}[T^+]} = \sqrt{\frac{1}{24} \cdot [30 \cdot 31 \cdot 63]} \approx 48.618,$$

$$T^* = \frac{200 - 232.5}{48.618} \approx -0.668,$$

Քանի որ թեստը երկկողմանի է և $\alpha = 0.05$ դեպքում $|T^*| > z_{1-\alpha/2}$ պայմանը տեղի չունի՝ $-z_{0.975} < |T^*| < z_{0.975} \approx 1.96$ և, չենք կարող մերժել H_0 հիպոթեզը:

§3. Ռաների թեստ (Runs test)

Դիցուք տրված է դիխոտոմ հաջորդականություն՝ $X_1, X_2, \cdots, X_n$, կազմված n_1 ($n_1 > 0$) «1»-երից և n_2 ($n_2 > 0$) «0»-ներից, որտեղ $n_1 + n_2 = n$։ Հաջորդականության պատահականությունը որոշելու համար դիտարկում են R թեստ վիճականին (ռան), որը ցույց է տալիս հաջորդականության մեջ անընդմեջ նույն արժեքների՝ «0» կամ «1»-երի, խումբ-շարքերի քանակը։ Օրինակ՝ $0, 0, 1, 0, 0, 1, 1, 1$ հաջորդականության դեպքում $R = 4$։ Զրոյական հիպոթեզը սահմանենք հետևյալ կերպ.

H_0: հաջորդականությունը կազմվել է n_1 «1»-երի և n_2 «0»-ների պատահական ընտրությամբ ($n_1 + n_2 = n$)։

Եթե զրոյական հիպոթեզը պնդում է, որ հաջորդականությունը պատահական է, ապա R թեստ վիճականու չափազանց մեծ կամ փոքր արժեքները աջակցում են այլընտրանքային հիպոթեզին, որը ենթադրում է ոչ պատահական՝ որոշակի միտում ունեցող հաջորդականություն։

Միակողմանի ալտերնատիվ հիպոթեզներ են.

- H_A: R-ը շատ ավելի փոքր է քան սպասվում էր,
- H_B: R-ը շատ ավելի մեծ է քան սպասվում էր։

Զրոյական հիպոթեզի ճշմարտացիության պայմաններում, R-ի առաջին երեք կենտրոնական մոմենտները կլինեն.

- $\mu_R = 1 + \frac{2n_1n_2}{n}$,
- $\sigma_R^2 = \frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)}$,
- $E(R - \mu_R)^3 = -\frac{2n_1n_2(n_2-n_1)^2(4n_1n_2-3n)}{n^3(n-1)(n-2)}$։

H_0 հիպոթեզի ճշմարտացիության պայմաններում R-ի բաշխման pmf-ը կունենա հետևյալ տեսքը [ref ?].

$$f_R(r) = \begin{cases} \dfrac{2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n}{n_1}}, & r = 2k, \\ \dfrac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}}{\binom{n}{n_1}}, & r = 2k+1, \end{cases}$$

որտեղ r-ը ռաների քանակն է ($r = 2, 3, \ldots, n$)։

H_0 հիպոթեզը մերժվում է, եթե ռաների R քանակի համար, որը կարող է ընդունել $r = 2, 3, \cdots, n$ արժեքները, տեղի ունեն հետևյալ առնչությունները.

$$R \le g_1(\alpha; n_1, n_2) \text{ կամ } R \ge g_2(\alpha; n_1, n_2),$$

որտեղ $g_1(\alpha;\, n_1,\, n_2)$-ը և $g_2(\alpha;\, n_1,\, n_2)$ համապատասխանաբար ստորին և վերին կրիտիկական արժեքներն են: g_1-ը և g_2-ը որոշվում են R-ի բաշխմամբ՝ H_0-ի ճշմարտացիության պայմաններում, այնպես, որ

$$P(R \le g_1(\alpha;\, n_1,\, n_2)) \le \frac{\alpha}{2}, \qquad P(R \ge g_2(\alpha;\, n_1,\, n_2)) \le \frac{\alpha}{2}:$$

Այսինքն, g_1 և g_2-ը սահմանում են R-ի բաշխման ստորին և վերին $\frac{\alpha}{2}$ քվանտիլները՝ H_0-ի ճշմարտացիության պայմաններում:

Երբ $n_1 > 15$ և $n_2 > 15$, R ռաների քանակը բաշխումը՝ H_0-ի ճշմարտացիության պայմաններում կարող է մոտարկվել նորմալ բաշխման միջոցով՝ $R \sim N(\mu_R,\, \sigma_R^2)$:

Ասիմպտոտիկ կերպով, երբ $n_1 \to \infty$ և $(\exists \varepsilon,\, s.t.\ 0 < \varepsilon < 1)$ այնպես, որ $\varepsilon \le \frac{n_1}{n_1+n_2} \le 1 - \varepsilon$.

$$P(R \le r) = \Phi(\frac{r+0.5-\mu_R}{\sigma_R}) + O(n_1^{-1/2})$$

Հիմնվելով նորմալ բաշխմամբ մոտարկման վրա, նշված կրիտիկական արժեքները կորոշվեն հետևյալ կերպ.

Ձախակողմյան թեստի դեպքում.

$P(R \le g_1) \approx \Phi(\frac{g_1+0.5-\mu_R}{\sigma_R}) \le \alpha$, հետևաբար $\frac{g_1+0.5-\mu_R}{\sigma_R} \le z_\alpha$, վերցնենք

$g_1 \approx \lfloor z_\alpha \sigma_R + \mu_R - 0.5 \rfloor$:

H_0-ն կմերժենք երբ $R \le g_1(\alpha;\, n_1,\, n_2)$, այսինքն $R \le z_\alpha \sigma_R + \mu_R - 0.5$,

Աջակողմյան թեստի դեպքում.

$P(R \ge g_2) \approx \Phi(\frac{g_2+0.5-\mu_R}{\sigma_R}) \le 1 - \alpha$, հետևաբար $\frac{g_2+0.5-\mu_R}{\sigma_R} \le z_{1-\alpha}$, վերցնենք

$g_2 \approx \lfloor z_{1-\alpha} \sigma_R + \mu_R - 0.5 \rfloor$:

H_0-ն կմերժենք երբ $R \ge g_2(\alpha;\, n_1,\, n_2)$, այսինքն $R \ge z_{1-\alpha} \sigma_R + \mu_R - 0.5$:

Երկկողմանի թեստի դեպքում,

$P(R \le g_1) \approx \Phi(\frac{g_1+0.5-\mu_R}{\sigma_R}) \le \frac{\alpha}{2}$, $g_1 \approx \lfloor z_{\alpha/2} \sigma_R + \mu_R - 0.5 \rfloor$

$P(R \ge g_2) \approx \Phi(\frac{g_2+0.5-\mu_R}{\sigma_R}) \le 1 - \frac{\alpha}{2}$, $g_2 \approx \lfloor z_{1-\alpha/2} \sigma_R + \mu_R - 0.5 \rfloor$

H_0-ն կմերժենք երբ $R \le g_1(\alpha;\, n_1,\, n_2)$ կամ $R \ge g_2(\alpha;\, n_1,\, n_2)$:

Դիտարկենք հետևյալ օրինակը.

Սանտա Քրուզում անհանգստություն կա կապված գլոբալ տաքացման և El Niño էֆեկտի հետ: Հուլիսի 1-21-ը 2003 թ. յուրաքանչյուր օրվա ջերմաստիճանը համեմատվել է 1993–2002 թթ. նույն օրվա համար հաշվարկված միջին ջերմաստիճանի հետ: Եթե 2003-ի նշված օրվա ջերմաստիճանը գերազանցում է նշված ժամանակահատվածի միջին ջերմաստիճանին, ապա գրանցվում է A սիմվոլը, հակառակ դեպքում՝ B: Ստացվել է հետևյալ հաջորդականությունը.

AAABBAA|AABAABA|AAABBBB.

Սահմանենք զրոյական և ալտերնատիվ հիպոթեզները հետևյալ կերպ.

H_0: հաջորդականությունը պատահական է

H_1: հաջորդականությունը պատահական չէ

Խնդրի լուծումը ծրագրային ապահովման միջոցով հասանելի է RunsTest.ipynb-ում: $\alpha = 0.05$ վստահության մակարդակի դեպքում ստացվում է, որ չենք կարող մերժել զրոյական հիպոթեզը:

§4. **Կոլմոգորով - Սմիրնովի՝ 1 ընտրանքի թեստ (One Sample Kolmogorov-Smirnov Test, Kolmogorov-Smirnov Goodness of Fit Test)**

Երբեմն, հետազոտությունների ժամանակ, տվյալների բաշխման վերաբերյալ կարելի է անել նախնական ենթադրություններ՝ հիմնված իրականացված փորձի առանձնահատկությունների կամ փորձարարի՝ նախկինում ունեցած փորձի հիման վրա: Ստուգելու համար՝ արդյոք տվյալները իրոք ունեն ենթադրյալ բաշխումը, կարելի է կիրառել մի շարք ոչ-պարամետրական թեստեր:

Ներկայացնենք նշված նպատակով կիրառվող թեստերից Կոլմոգորով-Սմիրնովի թեստը, որը հիմնված է ընտրանքի $ECDF = F_n(x)$ - ի և որևէ տեսական բաշխման CDF ի՝ $F_0(x)$ համեմատության վրա: Սահմանենք զրոյական և ալտերնատիվ հիպոթեզները՝

$H_0: F(x) = F_0(x),\ \forall x,$

$H_1: F(x) \neq F_0(x),$

Համարենք $F(x)$ անընդհատ ֆունկցիա է: $X_1, X_2, \dots X_n$ ընտրանքի էմպիրիկ բաշխման ֆունկցիան՝ ECDF-ը, նշանակենք $F_n(x)$-ով: Սահմանենք Կոլմոգորով - Սմիրնովի թեստ վիճականին՝

$$D_n = sup_{-\infty<x<\infty} |F_n(x) - F_0(x)|$$

Սահմանենք նաև D_n^+ և D_n^- վիճականիները [4]՝

$$D_n^+ = sup_{-\infty<x<\infty}(F_n(x) - F_0(x)) = max\left\{0,\ max_{1\le i\le n}\left\{\frac{i}{n} - F_0(X_i)\right\}\right\},$$

$$D_n^- = sup_{-\infty<x<\infty}(F_0(x) - F_n(x)) = max\left\{0,\ max_{1\le i\le n}\left\{F_0(X_i) - \frac{i-1}{n}\right\}\right\},$$

$$D_n = max(D_n^+, D_n^-)$$

H_0 վարկածի ներքո, օգտվելով $F_0(x)$-ի անընդհատությունից կարող ենք ասել, որ $Y_i = F_0(X_i)$ պատահական մեծությունները կունենան $Uniform([0, 1])$ բաշխում (նման խնդիր ունեինք Ստատիստիկայի I տնային աշխատանքում :)):

Այսպիսով՝

$$D_n = sup_{-\infty<x<\infty} |F_n(x) - F_0(x)| = sup_{0<y<1} |S_n(y) - y|,$$

որտեղ $S_n(y) = ECDF[Y_1, Y_2, \ldots Y_n]$.

Այսինքն, անկախ $F(x)$-ի տեսքից, թեստի վիճականին՝ D_n կունենա այնպիսի բաշխում, ինչպիսին կլիներ D_n-ի բաշխումը, եթե ընտրանքը լիներ $Uniform([0, 1])$ բաշխումից: D_n-ի համար կարելի է գտնել անալիտիկ տեսքը [7], սակայն հաճախ օգտագործում են $\sqrt{n} \cdot D_n$-ի ասիմպտոտիկ բաշխումը՝ երբ $n \gg 1$ [8]:

$$G(z) = P\left(\sqrt{n}D_n \le z\right) \to 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} \cdot e^{-2i^2z^2},$$

և նմանապես

$$P\left(\sqrt{n}D_n^+ > z\right) \to 1 - e^{-2z^2}.$$

Հետևյալ շարքից կարելի է ստանալ α նշանակալիության մակարդակի վերին կրիտիկական կետը (upper critical point of the Rejection Region of Significance Level α)՝ $d_{1-\alpha}$: Այդ դեպքում՝

$$P\left(\sqrt{n} \cdot D_n > d_{n, 1-\alpha}\right) = \alpha,$$

և կմերժենք H_0 վարկածն՝ ի հավանություն H_1 - ի, երբ $\sqrt{n} \cdot D_n > d_{n, 1-\alpha}$:

Կոլմոգորով-Սմիրնովի թեստի ունակայնության հատկությունների մասին մանրամասն կխոսենք Գլուխ 3.4-ում:

D_n^+ և D_n^- վիճականիները ավելի արդյունավետ է օգտագործել նաև H_0 - նկատմամբ միակողմանի այլընտրանքային վարկածները ստուգելիս:

Հավելենք նաև, որ փոքրածավալ ընտրանքի դեպքում անհրաժեշտ է օգտագործել D_n վիճականու ճշգրիտ բաշխումը, սակայն H_0-ն մերժելու պայմանը չի փոխվի: Բարեբախտաբար, $d_{n, 1-\alpha}$-ի արժեքներն հաշվված են նաև փոքր ընտրանքների համար [1]:

Կոլմոգորով - Սմիրնովի՝ 1 ընտրանքի թեստի կիրառության օրինակ ներկայացված է Kolmogorov.ipynb ֆայլում:

Գլուխ 3. Ոչ պարամետրական երկու ընտրանքի թեստեր

§1. Ոչ պարամետրական, զույգված և չզույգված ընտրանքի թեստեր

Դիցուք $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ դիտարկումներ են՝ վերցված փորձարկման ենթարկված նույն միավորներից: Կասենք, որ տվյալները զույգված կամ համապատասխանեցված են, եթե

1. յուրաքանչյուր (X_i, Y_i) զույգ կազմված է երկու չափումից նույն օբյեկտի վրա, կամ երկու օբյեկտի վրա՝ որոնք ինչ-որ նպատակով դրվել են համապատասխանության մեջ,
2. $D_i = Y_i - X_i$ տարբերությունը իմաստալից է, քանի որ X_i-ն և Y_i-ն նկարագրում են նույն միավորը,
3. $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ զույգերն անկախ են, բայց յուրաքանչյուր զույգի կոմպոնենտներ միմյանցից անկախ չեն:

Դիտարկենք հետևյալ օրինակը. ենթադրենք իրականացվել է հետազոտություն մարդկանց ինչ-որ խմբի մոտ, որից հետո մի քանի ամիս նրանք ստացել են X դեղամիջոցը: Դեղի ընդունումից հետո կրկին իրականացել է նույն հետազոտությունը նույն խմբի մոտ և խնդիրն է պարզել արդյոք դեղամիջոցն ունեցել է դրական ազդեցություն, թե ոչ:

Հետազոտությամբ ստացված թեստի արդյունքները կապված կամ զույգված են, քանի որ յուրաքանչյուր անձ հետազոտվել է երկու անգամ: Այսինքն առաջին հետազոտությամբ ստացված արժեքները համեմատության մեջ են դրվում դեղամիջոցն ընդունելուց հետո ստացած արժեքների հետ.

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n),$$

որտեղ n -ը հետազոտությանը մասնակցած մարդկանց քանակն է, X_i-ն i -րդ անձի հետազոտության արդյունքներն են մինչ դեղամիջոցի ընդունումը, իսկ Y_i-ն՝ ընդունումից հետո $(i = 1, \ldots, n)$:

Չզույգված տվյալները նկարագրենք հետևյալ կերպ. դիցուք $X_1, X_2, \ldots, X_n \sim F(x)$ և $Y_1, Y_2, \ldots, Y_m \sim G(x)$ իրարից անկախ պատահական ընտրանքներ են, որտեղ

- $F(x)$ -ը և $G(x)$ -ը համախմբությունների անհայտ բաշխման ֆունկցիաներն են,
- $X_1, X_2, \ldots, X_n$ և $Y_1, Y_2, \ldots, Y_m$ ընտրանքներն անկախ են: Յուրաքանչյուր X_i և Y_j $(i = 1, \cdots, n, j = 1, \cdots, m)$ գալիս են փորձարկման ենթարկված տարբեր միավորներից,
- դիտարկումների միջև ոչ մի տրամաբանորեն զույգման անհրաժեշտություն չկա:

Դիտարկենք չզույգված տվյալների հետևյալ օրինակը. մեթոդաբանության արդյունավետությունը պարզելու համար ընտրվել է ուսանողների երկու խումբ՝ կազմված համապատասխանաբար n_1 և n_2 ուսանողներից: Խումբ 1-ի դասերն իրականացվել են առաջին մեթոդով և ստացվել են $X_1, X_2, \cdots, X_{n_1}$ արդյունքները, իսկ Խումբ 2-ի դասերը՝ երկրորդ մեթոդով և ստացվել են $Y_1, Y_2, \cdots, Y_{n_2}$ արդյունքները: Խնդիրն է համեմատել մեթոդների արդյունավետությունը և ընտրել լավագույնը:

§2. Ուիլկոքսոն նշանով ռանկի թեստ

Ուիլկոքսոնի նշանով ռանկի այս թեստը նախատեսված է X, Y զույգված տվյալներով ընտրանքների մեդիանների հավասարությունը ստուգելու համար, կամ որ նույնն է՝ ստուգելու արդյոք $D_i = Y_i - X_i, i = 1, \cdots, n$ -ի բաշխումը սիմետրիկ է 0 մեդիանի նկատմամբ: Թեստն ունի կիրառման որոշակի սահմանափակումներ, ինչպիսիք են.

1. $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ զույգված տվյալներն անկախ են և $D_i, i = 1, \cdots, n$ տարբերությունները իմաստալից են,
2. D_i-երը $(i = 1, \cdots, n)$ պատահական մեծություններ են՝ $Y_i - X_i, i = 1, \cdots, n$ տարբերությունների ընդունած հնարավոր արժեքների բազմությունից:
3. համախմբության $D = Y - X$ տարբերությունների բաշխումը սիմետրիկ է համախմբության մեդիան տարբերության նկատմամբ:

Զրոյական հիպոթեզը կսահմանվի հետևյալ կերպ.

$H_0: \mu = 0,$

որտեղ μ-ն տարբերությունների բաշխման մեդիանն է:

$m \leq n$-ով նշանակենք ոչ-զրոյական տարբերությունների քանակը $(D_i \neq 0, i = 1, \cdots, n)$: Կարգավորենք տարբերությունները՝ բացարձակ արժեքներով.

$$|D_{(1)}| \leq |D_{(2)}| \leq \cdots \leq |D_{(m)}|$$

և դրանց համապատասխանեցնենք $R_1, \cdots, R_m$ ռանկեր (եթե տարբերություններում ties կան, ապա դրանց վերագրել համապատասխան ինդեքսների միջինը) ինչպես որ իրականացվել էր մեկ ընտրանքի Ուիլկոքսոն նշանով ռանկի թեստում:

Երկու ընտրանքի Ուիլկոքսոն նշանով ռանկի թեստը վերածվում է մեկ ընտրանքով Ուիլկոքսոն նշանով ռանկի թեստի՝ իրականացված $D_1, D_2, \cdots, D_m$ տարբերությունների վրա:

§3. Ուիլկոքսոն նշանով ռանկի թեստի կիրառություններ

Դիտարկենք 2.1 -ում տրված խնդիրը, սահմանենք զրոյական հիպոթեզը հետևյալ կերպ.

$H_0: \mu_D = 0,$

$H_A: \mu_D \neq 0,$

որտեղ μ_D -ն Նյու Յորքի և Վաշինգթոնի համապատասխան օրերի ջերմաստիճանների տարբերությունների մեդիանն է:

Օր	Վաշինգթոն (°F)	Նյու Յորք (°F)	Տարբերություն	Ռանկերն առանց 0 ների	Նշաններ
Հունվարի 15	31	38	7	11.5	+
Փետրվարի 15	35	33	-2	3	-
Մարտի 15	40	37	-3	5	-
Ապրիլի 15	52	45	-7	11.5	-
Մայիսի 15	70	65	-5	8.5	-
Հունիսի 15	76	74	-2	3	-
Հուլիսի 15	93	89	-4	6.5	-
Օգոստոսի 15	91	85	-6	10	-
Սեպտեմբերի 15	74	69	-5	8.5	-
Հոկտեմբերի 15	55	51	-4	6.5	-
Նոյեմբերի 15	26	25	-1	1	-
Դեկտեմբերի 15	26	24	-2	3	-

Եթե տվյալներում որոշակի օրերի ջերմաստիճանի տարբերություն արձանագրված չլիներ, ապա Հավելված 1 -ում կցված աղյուսակում n-ի փոխարեն կդիտարկվեր $n - m$ -ը, որտեղ m -ը նման օրերի քանակն է:

$T^+ = 11.5$, հետևաբար Հավելված 1 -ի աղյուսակում դիտարկենք $n = 12$, $\alpha = 0.05$ երկկողմանի դեպքը. համապատասխանում է 13 արժեքը: Քանի որ ստացել էինք որ $T = 11.5$, որն էլ $T < 13$, կասենք, որ H_0 հիպոթեզը մերժվում է:

Նշված օրինակի դեպքում, Ուիլկոքսոնի նշանով ռանկի թեստը ($T = 11.5$, $n = 12$, $p < 0.05$) ցույց տվեց, որ քաղաքների ջերմաստիճանները ընդհանուր առմամբ նշանակալիորեն տարբեր են:

Դիտարկենք թեստը մեծ ծավալի տվյալների դեպքում.

Դիցուք դեղագործական ընկերությունը ցանկանում է փորձարկել երկու դեղամիջոցի ազդեցությունների տարբերությունը 34 հոգուց բաղկացած բուժառուների խմբի վրա ովքեր ունեն խրոնիկ ցավեր: Դեղ 1 և Դեղ 2 սյունակները ցույց են տալիս տվյալ դեղամիջոցն ընդունելուց մեկ ժամ անց բուժառուներից յուրաքանչյուրի ցավերի աստիճանը (արժեքներ են ընդունում [1, 10] միջակայքից): Անհրաժեշտ է պարզել որդյոք երկու դեղամիջոցը միջինում ունեն նույն ազդեցությունը, թե զգալիորեն տարբերվում են:

Բուժառու	Դեղ 1	Դեղ 2	Բուժառու	Դեղ 1	Դեղ 2	Բուժառու	Դեղ 1	Դեղ 2
1	2.4	2.5	13	4	6.1	25	4	5.1
2	4.7	3.3	14	2.2	2.9	26	5.5	4.4
3	1.2	5.3	15	2.7	4.3	27	3.6	3.6
4	5.9	5.6	16	2.9	3.3	28	3.8	3.5
5	4.5	5	17	5	5	29	5.4	4.8
6	4	5.3	18	3.1	5.1	30	2.4	3.2
7	2.5	4.6	19	3.3	3.3	31	4.1	2.6
8	3	2.5	20	3	5.9	32	4.5	5.7
9	5	3.4	21	5.4	3.2	33	4	5.8
10	5.8	5.4	22	4.2	5.9	34	6	5
11	1.9	5.1	23	3.6	5.9			
12	3.2	4.3	24	2.2	5.6			

Յուրաքանչյուր բուժառուի դեպքում հաշվենք երկրորդ և առաջին դեղեր ընդունելիս ցավերի աստիճանների տարբերությունը, ռանկավորենք դրանց բացարձակ արժեքներն ու կցենք նշանները.

N	Դեղ 1	Դեղ 2	Տարբ.	Ռանկը՝ նշանով	N	Դեղ 1	Դեղ 2	Տարբ.	Ռանկը՝ նշանով
1	2.4	2.5	0.1	1.0	13	4	6.1	2.1	24.5
2	4.7	3.3	-1.4	-17.0	14	2.2	2.9	0.7	9.0
3	1.2	5.3	4.1	31.0	15	2.7	4.3	1.6	19.5
4	5.9	5.6	-0.3	-2.5	16	2.9	3.3	0.4	4.5
5	4.5	5	0.5	6.5	17	5	5	0.	
6	4	5.3	1.3	16.0	18	3.1	5.1	2.	23.0

7	2.5	4.6	2.1	24.5	19	3.3	3.3	0.	
8	3	2.5	-0.5	-6.5	20	3	5.9	2.9	28.0
9	5	3.4	-1.6	-19.5	21	5.4	3.2	-2.2	-26.0
10	5.8	5.4	-0.4	-4.5	22	4.2	5.9	1.7	21.0
11	1.9	5.1	3.2	29.0	23	3.6	5.9	2.3	27.0
12	3.2	4.3	1.1	13.0	24	2.2	5.6	3.4	30.0

Բուժառու	Դեղ 1	Դեղ 2	Տարբ.	Ռանկերը՝ նշանով
25	4	5.1	1.1	13.0
26	5.5	4.4	1.1	-13.0
27	3.6	3.6	0.0	
28	3.8	3.5	0.3	-2.5
29	5.4	4.8	0.6	-8.0
30	2.4	3.2	0.8	10.0
31	4.1	2.6	1.5	-18.0
32	4.5	5.7	1.2	15.0
33	4	5.8	1.8	22.0
34	6	5	1.0	-11.0

Դրական նշանով ռանկերի գումարը 367.5 է, իսկ բացասականովը՝ 128.5:

Որպես թեստ վիճականի վերցնենք $T = \sum R^{-}$-ը, որը ունի նույն բաշխումը, ինչ T^{+} -ը:

Պետք է հաշվել $T^* = \frac{T - E[T]}{s_T}$, որտեղ $E[T] = \frac{n(n+1)}{4}$ և $s_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$:

$E[T] = \frac{n(n+1)}{4} = \frac{31\times32}{4} = 248$, $s_T = \sqrt{\frac{31\times32\times63}{24}} \approx 51.03$ կստացվի $T^* = \frac{367.5-248}{51.03} \approx -2.35$

Քանի որ $-2.35 < -1.96$, պարզ է որ H_0 վարկածը այն մասին որ դեղամիջոցների ազդեցության միջև նշանակալի տարբերություն չկա, պետք է մերժել:

Ամփոփելով բերված խնդիրը, կասենք, որ Ուիլկոքսոն նշանով ռանկի թեստը ($T = 128.5,\ n = 31,\ p < 0.05$) հայտնաբերեց դեղամիջոցների ազդեցության նշանակելի տարբերություն: Հաշվի առնելով կատարված վերլուծությունը, կասենք, որ երկրորդ դեղամիջոցը առավել նախընտրելի է:

§4. Կոլմոգորով-Սմիրնովի՝ 2 ընտրանքի թեստ (Two Sample Kolmogorov-Smirnov Test)

Կոլմոգորով - Սմիրնովի՝ 2 ընտրանքի թեստը սկզբունքորեն կառուցվում է 1 ընտրանքի թեստի նմանությամբ: Այս պարագայում տրվում է երկու էմպիրիկ բաշխման ֆունկցիաներ՝ 2 ընտրանքների համար, իսկ թեստի վիճականին սահմանվում է որպես նրանց տարբերության սուպրեմումը:

Դիտարկենք $X_1, X_2, \dots X_n$ և $Y_1, Y_2, \dots Y_m$ ընտրանքները համապատասխանաբար $F(x)$ և $G(x)$ անհայտ բաշխման ֆունկցիաններով: Սահմանենք զրոյական հիպոթեզ.

$H_0 : F(x) = G(x),\ \forall x$ ընդդեմ $H_1 : F(x) \neq G(x)$ երկկողմանի ալտերնատիվ հիպոթեզի:

$H_{11} : F(x) \geq G(x)$ և $H_{12} : F(x) \geq G(x)$ կլինեն համապատասխան միակողմանի ալտերնատիվ հիպոթեզները: Դիտարկենք 2 ընտրանքների ECDF - ները՝

$F_n(x) = \frac{1}{n}[\,\#\, of\ X\ observations \leq x]$ և $G_m(x) = \frac{1}{m}[\,\#\, of\ Y\ observations \leq x]$

Թեստի վիճականին կսահմանենք.

$D_{n,m} = sup_{-\infty<x<\infty} |F_n(x) - G_m(x)|.$

H_0 հիպոթեզի և $F(x)$ անընդհատության պայմանի ներքո $D_{n,m}$ վիճականու ընտրանքային բաշխումը կախված չէ $F(x)$ բաշխումից []: Ինչպես 1 ընտրանքի թեստի դեպքում՝ $D_{n,m}$ - ի բաշխումը կարելի է ստանալ կոմբինատոր մեթոդներով, ինչը, սակայն, խնդրահարույց է դառնում մեծ n, m դեպքում. Փոխարենը՝ կարելի է ստանալ նրա ասիմպտոտիկ բաշխումը [7].

$P\left(D^*_{n,m} \leq z\right) \to 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1} \cdot e^{-i^2z^2}$, երբ $n, m \to \infty$, որտեղ

$D^*_{n,m} = \left(\frac{nm}{n+m}\right)^{\frac{1}{2}} D_{n,m}$

Նմանապես, միակողմանի վիճականիները կսահմանվեն՝

$D^+_{n,m} = sup_{-\infty<x<\infty}\left(F_n(x) - G_m(x)\right),\ D^{*+}_{n,m} = \left(\frac{nm}{n+m}\right)^{\frac{1}{2}} D^+_{n,m}$

$D^-_{n,m} = sup_{-\infty<x<\infty}\left(G_m(x) - F_n(x)\right),\ D^{*-}_{n,m} = \left(\frac{nm}{n+m}\right)^{\frac{1}{2}} D^-_{n,m},$

որոնք ունեն նույն ասիմպտոտիկ բաշխումը՝ $H(x) = 1 - e^{-2x^2},\ 0 < x < \infty$:
Իրական բաշխման ֆունկցիան ոչ մեծ n,m դեպքում հաշված է [1] կողմից.
H_0- ն մերժվում է նույն դեպքում, ինչպես 1 ընտրանու թեստում:

Նշենք կարևոր հատկություն. Կոլմոգորով - Սմիրնովի թեստը ունակային է բոլոր $F(x)$, $G(x)$ բաշխումների համար, H_1: $F(x) \neq G(x)$ ալտերնատիվ վարկածի դեպքում [4]. Եթե $\exists x$: $|F(x) - G(x)| = \Delta > 0$, ապա՝

$$D^*_{n,m} := \sqrt{\frac{nm}{n+m}} D_{n,m} > \sqrt{\frac{nm}{n+m}} \Delta \to \infty, \Rightarrow P(D^*_{n,m} > c_\alpha) \to 1.$$

Այսինքն բաշխումների կամայական փոքր տարբերության դեպքում բավականաչափ մեծ քանակի տվյալները թույլ կտան մերժել H_0 վարկածը թեստի կամայական հզորության համար: Սակայն, ունակայնության այս հատկությունը իր թերությունն ունի. ցույց է տրված, որ Կոլմոգորով - Սմիրնովի թեստի հզորությունը հաճախ փոքր է այլ թեստերից՝ քննարկված ալտերնատիվ վարկածների դեպքում:

Կոլմոգորով - Սմիրնովի՝ 2 ընտրանքի թեստի կիրառության օրինակ ներկայացված է Kolmogorov.ipynb ֆայլում:

§6. **Ուոլդ-Ուոլֆովից թեստ** (Wald–Wolfowitz runs test)

Ուոլդ-Ուոլֆովից թեստը (անվանակոչվել է ի պատիվ վիճակագիրներ Abraham Wald-ի և Jacob Wolfowitz-ի) 2 ընտրանքի ոչ պարամետրական թեստ է, որի նպատակն է գնահատել՝ արդյոք տրված 2 ընտրանքները ունեն նույն բաշխումը թե ոչ: Առաջ են բերվում հետևյալ հիպոթեզները՝

$H_0: f(x) = g(x)$

$H_a: f(x) \neq g(x)$

Դիցուք ունենք n IID ընտրանք առաջին համախմբությունից $(x_1, x_2, ..., x_n \sim X)$ և m՝ երկրորդից $(y_1, y_2, ..., y_m \sim Y)$: Թեստը կիրառելու համար անհրաժեշտ է վերջիններս միացնել իրար (merge) և դասավորել աճման կարգով՝ անպայման պահպանելով յուրաքանչյուր արժեքի ընտրանքի պիտակը (label): Կարգավորված տվյալները կունենան հետևյալ տեսքը.

$$x_5, x_2, y_2, y_7, y_m, x_1, x_{n-1}, ... \ :$$

Շարքում հաջորդականորեն հայտված նույն ընտրանքի տարրերը անվանվում են ռան (run), օրինակ՝ առաջին 2 էլեմենտները X ընտրանքից են և կազմում են առաջին ռանը, հաջորդ ռանը կկազմեն y_2, y_7, y_m քանի որ Y-ից են և այլն: Որպես test statistics հանդես է գալիս ընտրանքներով ձևավորված ռաների քանակը և նշանակվում է U-ով:

Թեստի հիմքում ընկած է գաղափարը, որ եթե 2 ընտրանքները ունեն նույն բաշխումը (այսինքն՝ H_0 վարկածը ճիշտ է), ապա աճման կարգով դասավորելուց հետո դրանց տարրերը պատահական կերպով կխառնվեն և կձևավորվեն մեծ քանակությամբ ռաներ: Հակառակ դեպքում՝ ամեն ընտրանքի տարրերը կլինեն իրար մոտ, կազմելով խմբեր (clusters), և ձևավորված ռաները կլինեն քիչ: Այլ կերպ՝ f-ի և g-ի տարբերությունը նվազեցնում է U- ի արժեքը [9]:

H_0-ի տեղի ունենալու պայմանում՝ U-ն ունի ասիմպտոտիկ ստանդարտ նորմալ բաշխում [9]: Հետևաբար, մեծ քանակությամբ տվյալներ ունենալու դեպքում (սովորաբար n,m > 10 համարվում է բավարար մեծ)՝ օգտվելով ստանդարտ նորմալ բաշխման քվանտիլներից, կարող եք ընդունել կամ մերժել հիպոթեզը: Ստորև ներկայացնում ենք ալգորիթմը:

Քայլ 1՝ հաշվենք U-ի մաթ. սպասումը, որը ցույց կտա միջին ռաների քանակը, եթե H_0 վարկածը ճիշտ լիներ (ասել է թե՝ ընտրանքների տարրերը պատահականորեն խառնված կլինեին).

$$E[U] = \frac{2nm}{n+m} + 1$$

Քայլ 2՝ հաշվենք U-ի վարիացիան, որը ցույց կտա H_0-ի տեղի ունենալու պայմանում ռաների ցրվածությունը.

$$\sigma_U^2 = \frac{2nm(2nm-n-m)}{(n+m)^2(n+m+1)}$$

Քայլ 3՝ նորմալիզացնենք տվյալներից ստացված ռաների քանակը.

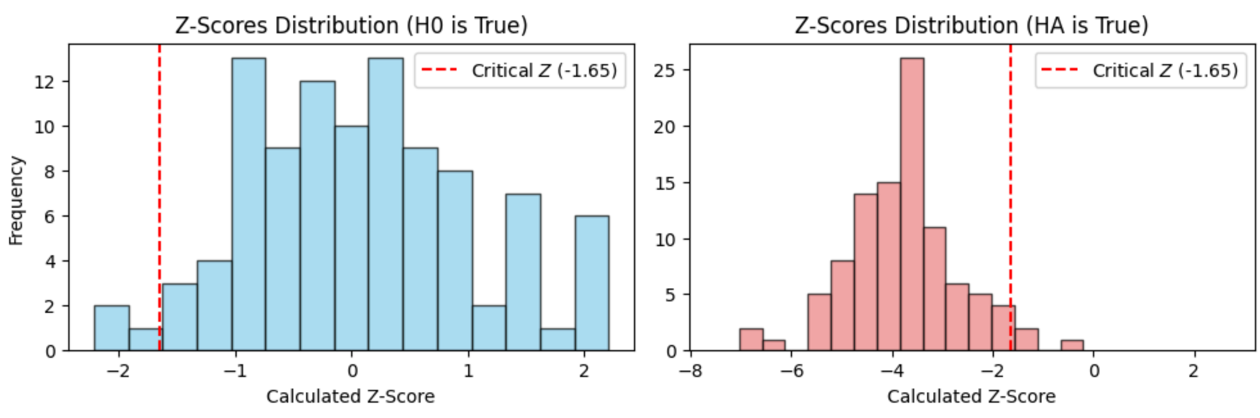
$$Z_U = \frac{U-E[U]}{\sigma_U^2}$$

Քայլ 4՝ ստացված արժեքը համեմատենք ստանդարտ նորմալի α-րդ քվանտիլի հետ (միտված ենք ստուգել արդյոք U-ի արժեքը շատ ավելի փոքր է քան վերջինիս մաթ. սպասումը, այսինքն՝ ունենք ձախակողմյան թեստ).

$$If\ Z_U < -Z_\alpha \Rightarrow reject\ H_0$$

Թեստի արդյունավետության ստուգում

Գեներացնենք 50-ական կետեր յուրաքանչյուր ընտրանքի համար, մի դեպքում նույն բաշխումից (H_0-ն ճիշտ է և $X, Y \sim N(10, 2)$), մեկ այլ դեպքում 2 տարբեր բաշխումներից (H_a-ն ճիշտ է, $X \sim N(10, 2)$ և $Y \sim N(13, 2)$):



Նկ. 1. Wald-Wolfowitz թեստի սիմուլյացիա ($\alpha = 0.05$)

Յուրաքանչյուր դեպքի համար հաշվարկենք Z_U արժեքները: Գործողությունը կրկնենք 100 անգամ:

Նկար 1-ում ամփոփված են ստացված արդյունքները՝ $\alpha = 0.05$ նշանակալիության մակարդակով: Համաձայն ստացված արդյունքների՝ H_0 վարկածը մերժվում է 3% դեպքերում, երբ այն իրականում ճիշտ էր (այսինքն ունենք Type I Error) և 96% դեպքերում մերժվում է H_a վարկածը երբ այն իրոք սխալ է (96.00% Power): Փորձակումը հասանելի է Wald-Wolfowitz_test.ipynb ֆայլում:

ԵԶՐԱԿԱՑՈՒԹՅՈՒՆ

Աշխատանքում ուսումնասիրվել են ոչ-պարամետրական վիճակագրական թեստերի տեսական հիմքերը և դրանց կիրառման մեթոդաբանությունը, մասնավորապես՝ շեշտը դնելով ընտրանքի բաշխման վերաբերյալ նվազագույն ենթադրություններ պահանջող մոտեցումների վրա: Նկարագրվել է պարամետրական թեստերի կարևորությունը այն դեպքերում, երբ տվյալների նորմալության կամ այլ պարամետրական ենթադրությունների ճշմարտացիությունը կասկածելի է կամ չի կարող ստուգվել բավարար հուսալիությամբ:

Աշխատանքի տեսական արդյունքները լրացվել են ծրագրային իրագործմամբ. մշակվել և փորձարկվել են համապատասխան ալգորիթմներ: Թվային փորձարկումները հաստատել են տեսական եզրակացությունների կիրառելիությունը և ցույց են տվել, որ ասիմպտոտիկ մոտարկումները բավարար ճշգրտություն են ապահովում գործնական կիրառությունների համար:

Իրականացված աշխատանքը ցույց է տալիս ոչ-պարամետրական մեթոդների արդյունավետությունը և ճկունությունը, ինչպես նաև դրանց կիրառման կարևորությունը ժամանակակից վիճակագրական վերլուծություններում:

ԳՐԱԿԱՆՈՒԹՅԱՆ ՑԱՆԿ

1. G. K. Kanji “100 Statistical Tests”,
2. M. R. Chernick, R. H. Friis “Introductory Biostatistics for the Health Sciences”,
3. N. Faizi, Y. Alvi “Biostatistics Manual for Health Research”, 2023
4. J.V. Deshpande, U. Naik-Nimbalkar, I. Diwan “Nonparametric Statistics: Theory and Methods”, 2018
5. M. Hollander, Douglas A. Wolfe, Eric Chicken “Nonparametric Statistical Methods”, third edition, 2014
6. W.A. Shewalter, A. Shewhart, S.S. Wilks “Nonparametric statistical methods”, third edition
7. N. V. Smirnov "On the estimation of the discrepancy between empirical curves of distribution for two independent samples", 1933
8. A.Kolmogoroff, “Sulla Determinazione Empirica Di Una Legge Di Distribuzione”, 1933
9. A.Wald, J. Wolfowitz, “On a test whether two samples are from the same population”, 1940
10. P. Kvam, B. Vidakovic, S. Kim, “Nonparametric Statistics with Applications to Science and Engineering with R, 2nd Edition”
11. Tests demonstrations source code: https://github.com/InnHar/StatisticsTests

Հավելված 1

TABLE B.3 Critical Values for the Wilcoxon Signed Rank Test Statistics *T*.

n	$\alpha_{\text{two-tailed}} \leq 0.10$ $\alpha_{\text{one-tailed}} \leq 0.05$	$\alpha_{\text{two-tailed}} \leq 0.05$ $\alpha_{\text{one-tailed}} \leq 0.025$	$\alpha_{\text{two-tailed}} \leq 0.02$ $\alpha_{\text{one-tailed}} \leq 0.01$	$\alpha_{\text{two-tailed}} \leq 0.01$ $\alpha_{\text{one-tailed}} \leq 0.005$
5	0			
6	2	0		
7	3	2	0	
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68
26	110	98	84	75
27	119	107	92	83
28	130	116	101	91
29	140	126	110	100
30	151	137	120	109

Source: Adapted from McCornack, R. L. (1965). Extended tables of the Wilcoxon matched pair signed rank statistic. *Journal of the American Statistical Association*, *60*, 864–871. Reprinted with permission from *The Journal of the American Statistical Association*.