Բովանդակություն

Ներածություն ----------5

Գլուխ 1 Վիճակագրության նշանակությունը տվյալների վերլուծությունում ----------6

1.1 Վիճակագրությունը որպես գիտական դիսցիպլինա ----------6

1.2 Տվյալների վերլուծության հիմնական գործառույթը ----------12

Գլուխ 2 Վիճակագրական ոչ պարամետրական մեթոդներ ----------14

2.1 Վիճակագրական մեթոդների դասակարգում. պարամետրական և ոչ պարամետրական մեթոդներ ----------14

2.2 Երեք և ավելի խմբերի համեմատման Կրասկալ–Ուոլիսի հայտանիշ ----------16

Գլուխ 3 Որակական գործոնների ազդեցության վերլուծության կիրառություն ----------18

3.1 Հաճախորդների վարքագծի վերլուծություն. Վճարման եղանակի և ամսական վճարների կապը ----------18

3.2 Կրթության և ոլորտի ազդեցությունը աշխատավարձերի վրա. համեմատական վերլուծություն նախաճգնաժամային և ճգնաժամային ժամանակահատվածներում ---26

Եզրակացություն ----------46

Օգտագործված գրականության ցանկ ----------47

Բնապահպանություն ----------50

Տնտեսագիտություն ----------60

Կենսագործունեության անվտանգություն ----------69

Ներածություն

Ժամանակակից աշխարհում տվյալների աննախադեպ աճը և տեղեկատվական տեխնոլոգիաների զարգացումը կարևորում են տարբեր ոլորտներում կուտակվող տվյալներից արժեքավոր տեղեկատվության ստացման մեթոդների կիրառումն ու կատարելագործումը: Այս համատեքստում վիճակագրությունը հանդես է գալիս որպես առանցքային միջոց, որը հնարավորություն է տալիս անել հուսալի եզրակացություններ և կայացնել հիմնավորված որոշումներ:

Տվյալների վերլուծությունում կիրառվող վիճակագրական մեթոդները դասակարգվում են պարամետրական և ոչ պարամետրական խմբերի: Այն դեպում, երբ տվյալները չեն բավարարում նորմալ բաշխվածության պայմանին, ունեն արտահայտված անհամաչափություն, պարունակում են արտառոց արժեքներ կամ ներկայացված են կարգային սանդղակով, վերլուծությունը իրականացվում է ոչ պարամետրական վիճակագրական մեթոդներով:

Սույն աշխատանքը նվիրված է ոչ պարամետրական վիճակագրական մեթոդների տեսական հիմքերի և գործնական կիրառությունների ուսումնասիրությանը: Մասնավորապես, դիտարկվել են Կրասկալ–Ուոլիսի, Մանն–Ուիթնիի տեստերը, ինչպես նաև երկգործոն վերլուծության համար կիրառվող ժամանակակից մոտեցումներից մեկը՝ Aligned Rank Transform (ART) մեթոդը: Վերջինս հնարավորություն է տալիս գնահատել ոչ միայն գործոնների գլխավոր ազդեցությունները, այլև դրանց փոխազդեցության էֆեկտը՝ զերծ մնալով ANOVA-ի խիստ սահմանափակումներից:

Python և R ծրագրավորման լեզուների կիրառմամբ իրականացվել է իրական տվյալների վիճակագրական վերլուծություն՝ հիմնավորելով ուսումնասիրված ոչ պարամետրական մեթոդների ընտրությունը տվյալների առկա կառուցվածքի պայմաններում:

Գլուխ 1 Վիճակագրության նշանակությունը տվյալների վերլուծությունում

1.1. Վիճակագրությունը որպես գիտական դիսցիպլինա

Վիճակագրությունը մաթեմատիկական գիտություն է, որը զբաղվում է տվյալների հավաքմամբ, դասակարգմամբ, վերլուծությամբ: Վերջնարդյունքում ընդունվում են որոշումներ, կատարվում եզրակացություններ՝ տալով դրանց հավանականային գնահատական: Այն օգնում է պատահականության հիմքի վրա, տվյալներից որոշումներ կայացնել ուսունասիրվող երևույթների վերաբերյալ:

Երբ ուսունմասիրվում են, օրինակ, բնակչության եկամուտները, ուսանողների գնահատականները, կայքի այցելությունները և այլն, դրանց վերաբերյալ կուտակված տվյալների հիման վրա, բնականաբար անհնար է ամեն մի տվյալ առանձին ուսումնասիրել: Վիճակագրությունը տվյալների վերլուծության փուլերում թույլ է տալիս՝

---Ամփոփ նկարագրել այդ տվյալները (օր.՝ միջին, տատանում, տոկոսներ);

--- Վիզուալիզացնել;

---Գտնել օրինաչափություններ, փոխկապվածություններ;

----Եվ այդ մասնավոր տվյալների հիման վրա անել եզրակացություններ, կայացնել որոշումները ընդհանուր հանուրի վերաբերյալ:

Վիճակագրությունը բաժանվում է երկու խմբի՝ մաթեմատիկական վիճակագրություն և կիրառական վիճակագրություն [1]:

Կիրառական վիճակագրությունը այնպիսի եղանակների, ընթացակարգերի ու հնարքների համակարգ է, որոնք հնարավորություն են տալիս փորձնական տվյալները հիմնավորված ձևով հավաքել, կազմավորել, ի մի բերել, ներկայացնել և վերլուծել՝ դրանց հիման վրա եզրահանգումներ անելու և որոշումներ ընդունելու նպատակով [2]:

1.1.1 Տվյալների թվային բնութագրիչներ: Վիճակագրական վարկած, թեստի հզորություն

Վիճակագրական վերլուծության առաջին փուլերից է տվյալները ամփոփ բնութագրիչների միջոցով նկարագրելը։ Այս ենթաբաժնում կսահմանենք այդ կարևոր հասկացությունները։

Որոշակի ուսումնասիրման ենթակա բոլոր առարկաների բազմությունը կանվանենք համախումբ (անգլերեն՝ population)։ Նրա տարրը կոչվում է անհատ։ Անհատների դիտարկվող հատկությունները կոչվում են հատկանիշներ։ Դրանք կարող են լինել որակական կամ քանակական[2]։

Համախմբի այն մասը, որը վերցվում է քննության համար, կոչվում է նմուշ/sample/, իսկ նմուշի ստացման գործընթացը՝ նմուշահանում/sampling/[2]։

Համախմբից վերցված պատահական նմուշահանմանը համապատասխանող անհատների հատկանիշների $(X_1, X_2, \dots, X_n)$ պատահական վեկտորը կոչվում է նմուշ մինչև փորձարկումը կամ դիտարկումը, իսկ փորձարկումից հետո ստացված $(x_1, x_2, \dots, x_n)$ հաջորդականությունը կոչվում է նմուշ փորձարկումից հետո (հավանականությունների տեսությունում $X_1, X_2, \dots, X_n$-ը անկախ միատեսակ բաշխված պատահական մեծությունների հաջորդականություն է)։

X հատկանիշի չափումներից ստացված տվյալները կոչվում են չմշակված։ Չնվազման կարգով վերադասավորված չմշակված տվյալների հաջորդականությունը կոչվում է փոփոխման (վարիացիոն) շարք[2]։ Փոփոխման շարքի անդամները նշանակվում են $x_{(k)}$-ով և կոչվում են կարգային վիճակներ՝

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}:$$

Հավաքագրելով քանակական տվյալները, բացի աղյուսակներից և գծապատկերներից, պահանջվում է ստանալ նաև այդ տվյալների բաշխումները նկարագրող այլ թվային բնութագրիչներ։ Այդ բնութագրիչների թվին են պատկանում կենտրոնական դիրքը և ցրվածությունը նկարագրող բնութագրիչները։

Շատ դեպքերում տվյալները ցուցաբերում են ակնհայտ միտում խմբավորվելու որոշակի կենտրոնական արժեքների շուրջ։ Կենտրոնական դիրքի բնութագրիչները ներկայացնում են այն թվային ցուցանիշները, որոնք բնութագրում են տվյալների բաշխման կենտրոնական

միտումը՝ սահմանելով այն արժեքները, որոնց շուրջ դիտարկվող մեծության իրականացումները ընդհանուր առմամբ տեղակայված են[2]: Դրանք են՝ *միջինը* (միջին թվաբանականը), *մեդիանը* (կիսողը, որը կոչվում է նաև՝ միջնակետ կամ միջնարժեք) և *մոդը* ՝ամենահաճախ հանդիպող տարբերակը:

Միջինը առավել հաճախ օգտագործվող վիճակագրական բնութագրիչներից է: Այն իրենից ներկայացնում է տվյալների «ծանրության կենտրոնը», որի շուրջ կուտակված/սփռված են բոլոր դիտումները[2]: Միջինը հաշվելու համար բոլոր տվյալների արժեքները գումարվում են, և արդյունքը բաժանվում այդ տվյալների քանակի վրա: Նմուշային միջինը նշանակվում է $\bar{\boldsymbol{x}}$: Այսպիսով, n ծավալ ունեցող $(x_1, x_2, \dots, x_n)$ նմուշի միջինը հավասար է՝

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Նմուշային մեդիանը $(\boldsymbol{x_{med}})$ այն արժեքն է, որը բաժանում է փոփոխման շարքը երկու «հավասար» մասի ($\boldsymbol{x_{med}}$-ից փոքր և $\boldsymbol{x_{med}}$-ից մեծ) այնպես, որ խմբերում լինեն հավասար քանակությամբ արժեքներ[2]: $(x_1, x_2, \dots, x_n)$ նմուշի համար, որին համապատասխանում է $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ փոփոխման շարքը, մեդիանը որոշվում է հետևյալ բանաձևի օգնությամբ՝

$$x_{med} = \begin{cases} \dfrac{x_{(m)} + x_{(m+1)}}{2}, & \text{եթե } n = 2m \\ x_{(m)}, & \text{եթե } n = 2m - 1 \end{cases}$$

Մոդ (նշանակվում է՝ $\boldsymbol{x_{mod}}$) կոչվում է նմուշային տվյալների ամենահաճախ հանդիպող արժեքը: Մոդը կարող է գոյություն չունենալ կամ միակը չլինել, այդ դեպքում բաշխումը կոչվում է բազմամոդալ[2]:

Ցրվածության բնութագրիչները նկարագրում են տվյալների սփռվածությունը (spread): Դրանք են՝ նմուշի *լայնքը, ցրվածքը* (դիսպերսիան), *ստանդարտ* (միջին քառակուսային) *շեղումը, փոփոխման* (վարիացիայի) *գործակիցը*:

Պարզագույն ցրվածության բնութագրիչը *նմուշային լայնքն* է՝ $R = x_{(n)} - x_{(1)}$, այսինքն՝ նմուշի մեծագույն և փոքրագույն արժեքների տարբերությունը:

Տվյալների բաշխվածությունը նկարագրող բնութագրիչներից են՝ նմուշային ցրվածքը (դիսպերսիան) և ստանդարտ կամ միջին քառակուսային շեղումը**:** Այդ բնութագրիչները ցույց են տալիս միջին քառակուսային ցրվածությունը միջինի (սպասելի) շրջակայքում[2]:

Դիցուք X հատկանիշին համապատասխանում է $(x_1, x_2, \dots, x_n)$ նմուշը:

Նմուշային ուղղված ցրվածք (դիսպերսիա) կոչվում է հետևյալ մեծությունը [2].

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

որտեղ $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ նմուշային միջինն է:

Նմուշային ստանդարտ կամ *միջին քառակուսային շեղում* կոչվում է հետևյալ մեծությունը՝

$$s = \sqrt{s^2}$$

Ի տարբերություն դիտարկված ցրվածության բնութագրիչների՝ *փոփոխման (վարիացիայի) գործակիցը* իրենից ներկայացնում է ցրվածության հարաբերական չափը և արտահայտվում է տոկոսներով: Փոփոխման գործակիցը ցույց է տալիս տվյալների միջինի նկատմամբ ցրվածության չափը: Փոփոխման գործակից կոչվում է տոկոսներով արտահայտված նրա միջին քառակուսային շեղման հարաբերությունը նմուշային միջինի մոդուլին[2]՝

$$v = \frac{s}{|\bar{x}|} * 100\%$$

Եթե $v < 30\%$, ապա տվյալները համարվում են համասեռ, դրանց հիմնական մասը գտնվում է միջինից ոչ շատ հեռու, և $\bar{x}$ միջինը լավ է բնութագրում տվյալները: Եթե $v > 30\%$, ապա տվյալները համարվում են անհամասեռ , այսինքն՝ դրանց մեծ մասը գտնվում է $\bar{x}$ միջինից հեռու, և նմուշային միջինը վատ է բնութագրում այդ տվյալները, այսինքն նումշը ներկայացուցչական չէ:

Կառուցվածքային ցուցանիշներ: Կառուցվածքային ցուցանիշների համակարգում, որպես բաշխման ձևի յուրահատկության ցուցանիշ, հանդես են գալիս տարբերակներ,

որոնք որոշակի դիրք են գրավում (յուրաքանչյուր չորրորդը, հինգերորդը, տասներորդը և այլն) վարիացիայի կարգավորված շարքում: Այդպիսի ցուցանիշները կրում են *քանորդիչ* ընդհանուր անվանումը[2]:

Քանորդիչներից են, մասնավորապես, քառորդիչները, որոնք փոփոխման շարքը բաժանում են չորս հավասար մասերի: Տարբերվում են ստորին քառորդիչ (Q_1), որն անջատում է համակցության $\frac{1}{4}$ մասը՝ հատկանիշի ամենափոքր արժեքով, և վերին քառորդիչ (Q_3), որը հատում է $\frac{1}{4}$ մասը՝ հատկանիշի ամենամեծ արժեքով:

Դա նշանակում է, որ համակցության միավորների 25%–ի արժեքները փոքր են Q_1–ից, 25%–ը՝ գտնվում է Q_1–ի և Q_2–ի միջև, 25%–ը՝ Q_2–ի և Q_3–ի միջև, և մնացած 25%–ը գերազանցում է Q_3–ը:

Երկրորդ քանորդիչը՝ Q_2–ը մեդիանն է: Քանորդիչների հաշվարկը նման է մեդիանի հաշվարկմանը:

Ռանգը ոչ պարամետրական վիճակագրության հիմնական հասկացություններից է: Այն ներկայացնում է դիտարկումների դասավորությունը ըստ մեծության: Ռանգը տվյալ արժեքի կարգային համարն է այն շարքում, որտեղ բոլոր տվյալները դասավորված են աճման կարգով [4]:

Օրինակ՝ ենթադրենք ունենք հետևյալ նմուշը՝ 3,2,4,3,5,4: Դասավորենք արժեքները աճման կարգով՝ 2,3,3,4,4,5: Նշենք յուրաքանչյուր արժեքը ռանգով՝ կարգով.

2–1, 3–2,3 (նույն արժեքի համար վերցնում ենք միջին ռանգը՝ $\frac{2+3}{2} = 2.5$), 4–4,5 (միջին ռանգը՝ $\frac{4+5}{2} = 4.5$), 5-6:

Աղյուսակ 2.1. Նմուշային ռանգերը

Նմուշի արժեք	*2*	*3*	*3*	*4*	*4*	*5*
Ռանգ	*1*	*2.5*	*2.5*	*4.5*	*4.5*	*6*

Ռանգը օգտագործվում է ոչ պարամետրական թեստերում՝ Վիլկոկսոնի, Կրասկալ–Ուոլիսի, Սպերմանի և այլն։ Այս թեստերը հաշվի չեն առնում տվյալների բաշխումը, այլ համեմատում են ռանգերը, ոչ թե իրական թվային արժեքները։

Վիճակագրական վարկածներ։ Շատ դեպքերում վիճակագրական մոդելավորումը իրականացվում է վիճակագրական վարկածների ստուգման միջոցով։ *Վիճակագրական վարկածներ* հանուրի բաշխման օրենքի կամ որոշակի պարամետրերի հատկությունների վերաբերյալ ենթադրություններն են [1]։

Հետազոտության նպատակն է լինում որոշել՝ ինչպես են համաձայնեցվում նմուշային տվյալները առաջադրված վարկածի հետ, կարելի է արդյոք նրանց միջև եղած անհամաձայնությունը վերագրել նմուշահանման ընթացքում առաջացած պատահական տատանումներին, թե վարկածը էապես հակասում է փորձնական տվյալներին և պետք է ժխտվի։ Այդ խնդիրը լուծվում է մաթեմատիկական վիճակագրության հատուկ եղանակներով։

Առաջադրված վարկածի հիմնավորված համադրումը նմուշի հետ կատարվում է այս կամ այն *վիճակագրական հայտանիշի* օգնությամբ։ Այդ գործողությունը անվանում են *վարկածի վիճակագրական ստուգում*։ Վարկածի վիճակագրական ստուգման արդյունքը չի նշանակում, որ ենթադրվող պնդումը միակ հարմարն է։

Ստուգման ենթակա *վարկածը* նշանակում են H_0 և անվանում հիմնական կամ զրոյական վարկած։ Յուրաքանչյուր զրոյական վարկածին կարող է հակադրվել *երկընտրանքային վարկածը*, որը նշանակում են H_1–ով։

Վարկածի վիճակագրական ստուգումը հիմնվում է փորձարկումների արդյունքների վրա և կարող է ընդունվել նաև սխալ որոշում։ Վարկածների ստուգման սխալները լինում են *երկու սեռի*։ Եթե ստուգման արդյունքում ժխտվում է H_0 վարկածը, երբ այն իրականում ճիշտ է, ապա ասում են, որ տեղի է ունեցել առաջին սեռի սխալ՝ $\mathrm{P}\{H_1|H_0\}=\alpha$։ Երբ H_0 վարկածը ընդունվում է, երբ իրականում ճիշտ է երկընտրանքային H_1-ը, ապա տեղի է ունեցել երկրորդ սեռի սխալ՝ $\mathrm{P}\{H_0|H_1\}=\beta$։ $\mathrm{P}\{H_1|H_1\}=1-\beta$ H_1-ի վերաբերյալ ճիշտ որոշում կայացնելու հավանականությունն է։ Վերջինս կոչվում է թեստի հզորություն։ Դա այն

հավանականությունն է, որ այլընտրանքային վարկածը իրականում ճիշտ լինելու դեպքում թեստը ճիշտ որոշում է կայացնում և մերժում է H_0-ն։

Վիճակագրական վարկածների ստուգման համար նախապես տրվում է առաջին սեռի սխալի հավանականությունը՝ α, որը կոչվում է նշանակալիության մակարդակ։ Որոշումը կայացվում է հիմնվելով ստուգման հայտանիշի՝ թեստի ընդունած արժեքի վրա։ Եթե այն թույլատրելի տիրույթից է, ապա ընդունվում է H_0 վարկածը։ Եթե կրիտիկական տիրույթից՝ H_1 [4]։ Բնականաբար վարկածի ստուգման համար առաջարկվող հայտնաիշներից լավ համարվում են նրանք, որոնց հզորությունը ավելի մեծ է։

1.2 Տվյալների վերլուծության հիմնական գործառույթը

Տվյալների վերլուծությունը այն գործընթացն է, որի ընթացքում հավաքված տվյալները ուսումնասիրվում, մշակվում և մեկնաբանվում են՝ օգտակար տեղեկատվություն ստանալու, եզրակացություններ անելու կամ որոշումներ ընդունելու համար [3]։

Տվյալների վերլուծությունը կատարվում է հետևյալ փուլերով՝

1.Տվյալների հավաքում. տվյալները կարող են գալ հարցումների, դիտարկումների, ֆայլերի, բազաների, սոցցանցերի կամ սենսորների միջոցով։

2.Տվյալների մաքրում. հեռացվում կամ ուղղվում են սխալ, կրկնվող կամ բացակայող արժեքները։

3.Տվյալների ուսումնասիրում. տեսողական (գրաֆիկներ, աղյուսակներ) և վիճակագրական (միջին, մեդիան, քանորդիչներ, ցրվածք և այլն) ուսումնասիրություն՝ հասկանալու համար տվյալների կառուցվածքն ու օրինաչափությունները։

4.Տվյալների մոդելավորում. օգտագործվում են մաթեմատիկական կամ մեքենայական ուսուցման մոդելներ՝ կանխատեսումներ կամ դասակարգումներ անելու համար։

5.Եզրակացություններ և ներկայացում. արդյունքները ներկայացվում են գրաֆիկների, հաշվետվությունների կամ դեշբորդների տեսքով՝ հեշտ ընկալման համար։

Ստորև ներկայացված է, թե յուրաքանչյուր փուլ ինչ է ներկայացնում իրենից՝

1.Տվյալների վերլուծության ամենակարևոր փուլերից մեկն տվյալների հավաքումն է, որովհետև եթե տվյալները սխալ կամ անորակ են, վերլուծությունն էլ չի կարող ճշգրիտ լինել:

2.Եթե հավաքված տվյալները «կեղտոտ» են՝ ունեն սխալներ, բաց արժեքներ կամ անհամապատասխան ձևաչափեր, ապա վերլուծության բոլոր արդյունքները կարող են լինել սխալ կամ մոլորեցնող:

3.Տվյալների ուսումնասիրումը տվյալների վերլուծության փուլերից է, որի նպատակն է՝ լավ ճանաչել տվյալները, հասկանալ դրանց կառուցվածքը, օրինաչափությունները, կապերը և առանձնահատկությունները մինչև բարդ մոդելների կամ կանխատեսումների կիրառումը: Տվյալների ուսումնասիրության ընթացքում տրվում է տվյալների հիմնական նկարագրություն, կատարվում է վիճակագրական ամփոփում (հաշվարկվում են հիմնական վիճակագրական բնութագրիչները՝ միջին, մեդիան, մաքսիմում և մինիմում արժեքներ, տատանում և այլն): Այնուհետև կատարվում է տվյալների բաշխման վիզուալ ուսումնասիրություն: Այս փուլին են պատկանում անհրաժեշտության դեպքում անոմալիաների և արտառոց արժեքների հայտանբերումը:

4.Տվյալների մոդելավորումը այն փուլն է, երբ արդեն մաքրված և ուսումնասիրված տվյալների հիման վրա կառուցվում է մաթեմատիկական կամ վիճակագրական մոդել, որը կարող է բացատրել, կանխատեսել կամ դասակարգել երևույթները: Եթե մոդելավորման ընթացքում կիրառվում են մեքենայական ուսուցման ալգորիթմները, ապա նախ տվյալները բաժանում ենք երկու մասի՝ ուսուցման տվյալներ և թեստավորման տվյալներ: Այնուհետև ընտրում ենք մոդելը՝ կախված խնդրի բնույթից: Հիմանականում այս մոտեցումը դիտվում է մեծածավալ տվյալների հետ աշխատելու ժամանակ: Վիճակագրական վերլուծությունը, ի տարբերություն մեքենայական ուսուցման, չի պահանջում կատարել տվյալների տրոհում, այլ առկա տվյալների հիման վրա կատարվում է վերլուծություն, վերլուծության արդյունքները գնահատվում են հավանականությամբ, կառուցվում են վստահության միջակայքերը և այլն:

5.Վերջին փուլում տվյալները և մոդելի արդյունքները հասկանալի ձևի ենք վերածում և կատարում ենք եզրակացություններ, որոնք պետք է տանեն հստակ գործողությունների և խնդիրների լուծման:

Գլուխ 2 Վիճակագրական ոչ պարամետրական մեթոդներ

2.1 Վիճակագրական մեթոդների դասակարգում. պարամետրական և ոչ պարամետրական մեթոդներ

Վիճակագրական վերլուծության ժամանակ կախված տվյալներից, վերլուծության պահանջներից, կիրառվում են համապատասխան վիճակագրական մեթոդներ:

Վիճակագրական մեթոդները իրենց հերթին բաժանվում են երկու խմբի՝

---Պարամետրական մեթոդներ;

---Ոչ պարամետրական մեթոդներ:

Պարամետրական մեթոդները այն մեթոդներն են, որոնք ենթադրում են, որ տվյալները գալիս են որևէ հայտնի տեսքով բաշխումից, մասնավորապես, ամենից հաճախ դրվող պայմանը վերաբերում է նորմալ բաշխվածությանը: Եթե հստակ չկա ինֆորմացիա բաշխվածության վերաբերյալ, պարմետրական մեթոդը կիրառելու համար նախապես՝ վարկածների ստուգման միջոցով, հիմնավորվում է նորմալ բաշխման և նրա պարամետրերի՝ միջին, դիսպերսիա, վերաբերյալ ենթադրության հավաստիությունը, որից հետո պարամետրական մեթոդի կիրառման արդյունքում արվող որոշումները կարող ենք համարել հավաստի [1,2]:

Հայտանիշը, որը մեզ թույլ է տալիս տվյալներից կորզել անհրաժեշտ ինֆորմացիան, մաթեմատիկական դրվածքով հանդիսանում է տվյալներից ֆունկցիա, որի բանաձևային նկարագրությունը տրվում է տվյալներից ստացվող միջինի, ուղղված դիսպերսիայի և այլն միջոցով:

Հիմնականում պարամետրական մեթոդները կիրառվում են հենց տվյալների վրա:

Ոչ պարամետրական մեթոդները չեն պահանջում որևէ բաշխման առկայություն: Դրանք չեն օգտագործում խիստ բաշխումներ, ի տարբերություն պարամետրական մեթոդների, ոչ պարամետրական մեթոդները հիմնվում են տվյալների կարգերի՝ ռանգերի վրա, իրականացվում է կարգային համեմատություններ՝ օգտվելով կիսողներից, քանորդիչներից և այլն[4]:

Վիճակագրությունում մեթոդի ընտրությունը պայմանավորված է ոչ միայն բաշխման վերաբերյալ պահանջից, այլ նաև ուսումնասիրվող հատկանիշների տեսակից: Որակական հատկանիշների փոխկապվածության վերաբերյալ հարցերը միշտ դասվում են ոչ պարամետրական մեթոդների ընտանիքին: Իսկ եթե ուսումնասիրվող հատկանիշներից գոնե մեկը քանակական է, ապա այստեղ կա նորմալ բաշխման վերաբերյալ պայմանի պահանջ:

Դիպլոմային աշխատանքում ուսումնասիրվել են այս դասին վերաբերող մեթոդներ, մասնավորապես՝

1. Mann–Whitney U test(Wilcoxon). Անկախ նմուշների համեմատության պարամետրական t-test–ի ոչ պարամետրական անալոգը համապատասխանաբար Մանն Ուիթնիի տեստն է, որը իրականացնում է զույգ դիտարկումներին համապատասխանող տարբերությունների մեդիանային հավասարության վարկածի ստուգումը [5]:
2. Kruskal–Wallis test: Միագործոն պարամետրական վերլուծության՝ one-way-ANOVA, ոչ պարամետրական անալոգը Կրասկալ–Ուոլիսի թեստն է: Այն իրականացնում է գործոնի մակարդակներին համապատասխանող տրոհված խմբերի կիսողների հավասարության վարկածի ստուգումը: Այս թեստը հիմնված է դիտարկումների ռանգավորման վրա [6]:
3. Ոչ պարամետրական զույգ-առ-զույգ ստուգման հայտանիշներ, մասնավորապես Դանիի տեստը[7]:
4. Ոչ պարամետրական երկգործոն ANOVA [8]:

Ոչ պարամետրական մեթոդները ունեն ինչպես առավելություններ, այնպես էլ թերություններ.

Առավելությունները՝

--Կիրառելի են անկախ տվյալների բաշխվածության վերաբերյալ տեղեկության։

--Հնարավորություն են տալիս վերլուծություն կատարել փոքր ծավալի տվյալների՝ նմուշների հիման վրա։

Հիմնական թերությունը կայանում է նրանում, որ դրանք զիջում են իրենց պարամետրական անալոգին թեստի հզորության առումով։

2.2 Երեք և ավելի խմբերի համեմատման Կրասկալ–Ուոլիսի /Kruskal–Wallis/ հայտանիշ

Վիճակագրական հետազոտություններում հաճախ անհրաժեշտ է պարզել՝ արդյո՞ք որակական հատկանիշը ազդեցություն ունի որևէ քանակական արդյունքի վրա։

Որակական հատկանիշը կոչվում է գործոն, իսկ գործոնի ազդեցությանը ենթարկվող հատկանիշը՝ արդյունքային։ Գործոնի յուրաքանչյուր արժեք կոչվում է մակարդակ։ Ըստ էության գործոնը որակական հատկանիշն է, իսկ արդյունքայինը՝ քանակական [6]։ Քանի որ արդյունքային հատկանիշը միայն գործոնի ազդեցությանը չի ենթարկվում, այն նաև կարող է կախված լինել այլ պատահական մեծություններից, օրինակ, բուժման դեպքում. հիվանդի տարիք, իմունիտետ, գենետիկա և այլն, ապա հիմնավորվում է արդյունքային հատկանիշի պատահական բնույթ՝ պատահական մեծություն լինելը։

Վիճակագրական համապատասխան մեթոդների կիրառումը թույլ է տալիս գտնելու հետևյալ հարցադրմամբ խնդրի լուծումը. կա՞ կապ որակական և քանակական հատկանիշների միջև։ Եվ դրանց կիառմամբ լուծման ընթացքը ուղեկցվում է այն գործիքների ու մեթոդական մոտեցումներնի վրա, որոնք իսկ մաս են կազմում տվյալների ուսումնասիրության, համեմատության, կապերի հայտնաբերման և որոշումներ կայացնելու գործընթացին։

Երեք և ավել մակարդակ ունեցող գործոնի ազդեցության բացահայտման խնդիրը բերվում է վարկածների ստուգման խնդրի և իրականացվում է Կրասկալ–Ուոլիսի թեստի միջոցով։

H_0: Խմբային կիսողները հավասար են /որակական գործոնը չունի ազդեցություն/

H_1: Առնվազն երկու խմբերի կիսողները իրար հավասար չեն /գործոնը ազդում է արդյունքային հատկանիշի վրա/:

Թեստի իրականացումը կատարվում է ռանգերի միջոցով. դիտարկվող տվյալները որպես մեկ ամբողջություն դասավորվում են աճման կարգով և ստանում համապատասխան ռանգ՝ R_{ij}, որտեղ j-ն խմբի համարն է, i-ն տվյալ խմբում դիտարկման կարգը[4]:

Թեստը նկարագրելու համար դիտարկվում են հետևյալ մեծությունները[4]՝

Խմբային ռանգերի միջին ռանգը՝

$$\bar{R}_J = \frac{1}{N_i}\sum_{j=1}^{N_i} R_{ij}$$

Միավորված տվյալների ռանգերի միջին արժեքը՝

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} R_{ij} = \frac{1}{N}\,\frac{N(N+1)}{2} = \frac{(N+1)}{2}$$

Ստուգման թեստը հետևյալն է՝

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{L} N_j\left(\bar{R}_J - \bar{R}\right)^2 = \frac{12}{N(N+1)}\sum_{j=1}^{L}\frac{R_j^2}{N_j} - 3(N+1)$$

Երբ H_0 վարկածը ճիշտ է, H-ն մոտարկվում է L-1 ազատության ասիճանով Խի-քառակուսի բաշխմամբ [1,2] (այն պատահական մեծությունների անընդհատ բաշխում է, որն օգտագործվում է հատկապես կատեգորիկ կամ ռանգավորված տվյալների համար): Կրիտիկական տիրույթը աջակողմյան է, ուստի տրված նշանակալիության մակարդակին համապատասխան կրիտիկական կետից դիտարկման արժեքի մեծ լինելու դեպքում ընդունվում է որակական գործոնի ազդեցության վերաբերյալ վարկածը:

Ոչ պարամետրական անալոգի կիրառումը տվյալ դեպքում հնարավորություն է տալիս իրականացնել տվյալների օբյեկտիվ վերլուծություն՝ հաշվի չառնելով բաշխման նորմալության և դիսպերսիաների համասեռության խիստ պահանջները: Փոխարինելով

թվային արժեքները ռանգերով՝ Կրասկալ-Ուոլիսի թեստը նվազեցնում է էքստրեմալ արժեքների ազդեցությունը և ապահովում է վիճակագրական հավաստիություն։

Գլուխ 3 Որակական գործոնների ազդեցության վերլուծության կիրառություն

3.1 Հաճախորդների վարքագծի վերլուծություն. Վճարման եղանակի և ամսական վճարների կապը

Հետազոտության համար օգտագործվել է Telco Customer Churn տվյալների հավաքածուն, որը ձեռք է բերվել Kaggle կայքից։ Տվյալների հավաքածուն պարունակում է հեռահաղորդակցման ընկերության 7043 հաճախորդների վերաբերյալ տեղեկատվություն։ Այն ներառում է 21 փոփոխական, որոնք բնութագրում են հաճախորդների տվյալները, ծառայությունների օգտագործումը, պայմանագրի տեսակը, վճարումների տվյալները, հաճախորդի հեռանալու փաստը։ Դրանցից հատկապես կարևոր են՝

--MonthlyCharges - հաճախորդի ամսական վճարը,

--PaymentMethod – հաճախորդի ընտրած վճարման եղանակը,

Հետազոտության նպատակն է պարզել, թե ինչպես է հաճախորդների վճարման եղանակը ազդում ամսական վճարների վրա, ինչպես նաև գնահատել տվյալ փոփոխականների միջև գոյություն ունեցող վիճակագրական կապերը։

Այսպիսով, առաջադրված են հետևյալ խնդիրը. Հաճախորդների ամսական վճարված գումարը պայմանավորվա՞ծ է վճարման մեթոդով։

Հետազոտության ընթացքում տվյալների վերլուծության և վիզուալիզացիայի համար օգտագործվելու է **Python** ծրագրավորման լեզուն [9] և հետևյալ գրադարանները՝

--pandas – տվյալների կարդալու, մշակման և վերլուծության համար,

--matplotlib.pyplot – գրաֆիկական պատկերների և վիզուալիզացիա ստեղծելու համար,

--seaborn – առավել առաջադեմ գրաֆիկական վիզուալիզացիայի համար, հատկապես բոքսփլոթներ և բաշխվածություն ցույց տալու համար,

--scipy – վիճակագրական թեստեր և գիտական հաշվարկներ իրականացնելու համար, օրինակ՝ Շապիրո-Ուիլքի, Լևենի և Կրասկալ-Ուոլիսի թեստերը։

Վերլուծության առաջին փուլում անհրաժեշտ է իրականացնել վիճակագրական բնութագրիչների հաշվարկ, որպեսզի ստանանք նախնական պատկերացում տվյալների բաշխման և բնութագրերի մասին։ Այդ նպատակով հաշվարկվում են հետևյալ հիմնական ցուցանիշները՝ միջին արժեքը (mean), մեդիանը (median), ստանդարտ շեղումը (standard deviation), ինչպես նաև քվանտիլները և տվյալների նվազագույն ու առավելագույն արժեքները։

Python ծրագրավորման լեզվում տվյալների նման վերլուծությունը հնարավոր է իրականացնել pandas գրադարանի describe() ֆունկցիայի միջոցով, որը միանգամից հաշվարկում է տվյալների հիմնական վիճակագրական բնութագրիչները։

Ստորև ներկայացված է MonthlyCharges (ամսական վճարներ) քանակական փոփոխականի հիմնական վիճակագրական բնութագրիչները, ըստ PaymentMethod (վճարման եղանակ) փոփոխականի մակարդակների.

Աղյուսակ 3.1. Ամսական վճարների վիճակագրական բնութագրիչների հաշվարկ

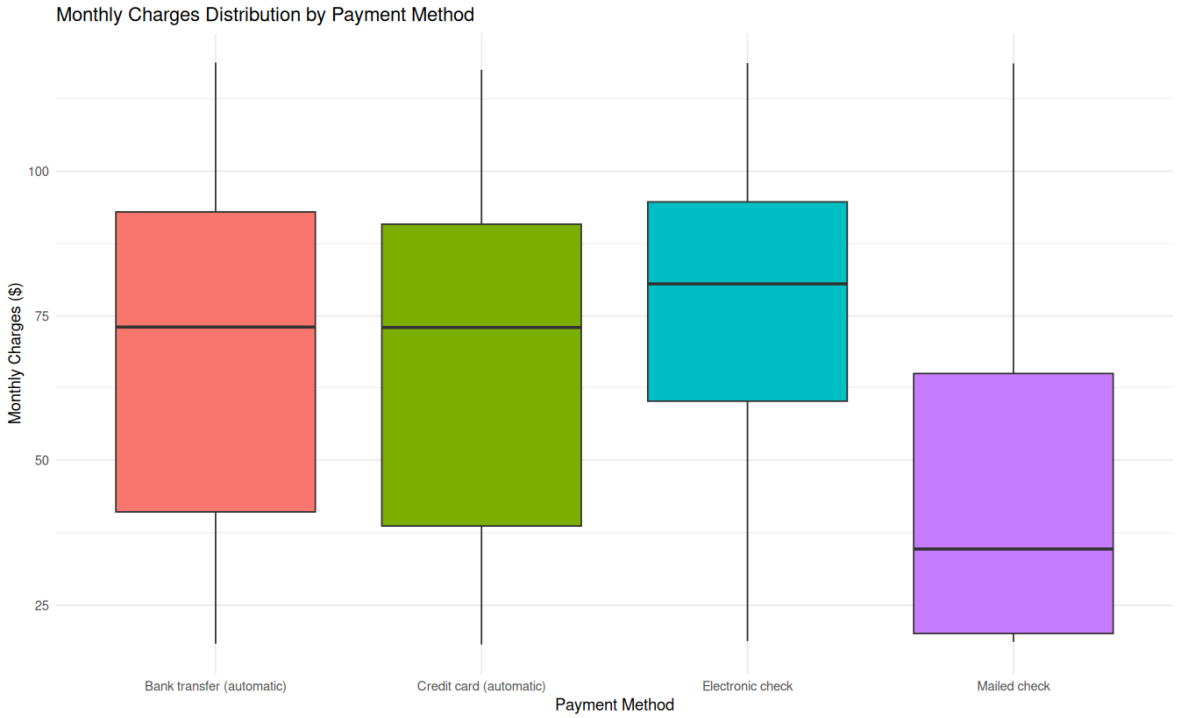
Statistics / PaymentMethod	mean	median	std	min	25%	75%	max
Bank transfer (automatic)	3079.30	2474.65	2357.74	19.25	1052	4934	8684.8
Credit card (automatic)	3071.40	2453.30	2407.40	19.30	989	5016	8670.1
Electronic check	2090.87	1253.90	2155.43	18.85	308	3340	8564.8
Mailed check	1054.48	467.35	1442.87	18.80	114	1294	8331.9

Ընդհանուր վճարների (TotalCharges) քվանտիլների վերլուծությունը ըստ PaymentMethod (վճարման եղանակ) փոփոխականի ցույց է տալիս որոշակի տարբերություններ հաճախորդների խմբերի միջև: Bank transfer (automatic) և Credit card (automatic) վճարման եղանակների դեպքում մեդիան արժեքները համապատասխանաբար կազմում են մոտ 2474.65 և 2453.30 դոլար, իսկ 25% և 75% քվանտիլները գտնվում են մոտավորապես 1052–4934 և 989–5016 դոլար միջակայքներում: Սա ցույց է տալիս, որ տվյալ խմբերում հաճախորդների ընդհանուր վճարումները բավականին բարձր են և ունեն լայն տատանումներ:

Բոլոր խմբերում նկատվում է բարձր ստանդարտ շեղում, ինչը վկայում է տվյալների զգալի տատանողականության մասին: Նվազագույն և առավելագույն արժեքների միջև մեծ տարբերությունը ևս ցույց է տալիս, որ տվյալներում կան ինչպես փոքր, այնպես էլ բավականին մեծ ընդհանուր վճարումներ կատարած հաճախորդներ:

Ստացված նկարագրական վիճակագրության արդյունքները տալիս են նախնական պատկերացում տվյալների բաշխման, ցրվածության և հիմնական միտումների մասին: Սակայն փոփոխականների միջև կապերը և հնարավոր գործոնի ազդեցությունը վիզուալ կարող ենք տեսնել նաև գրաֆիկական ներկայացմամբ: Տվյալների գրաֆիկներից արկղային դիագրամաները հնարավորություն են տալիս արագ գնահատել տվյալների բաշխման ընդհանուր տեսքը, հայտնաբերել հնարավոր արտառոց արժեքները, բաշխման ասիմետրիկությունը կամ այլ առանձնահատկություններ:

Ստորև ներկայացված է արկղային դիագրամա, որը պատկերում է ամսական վճարների բաշխվածությունը՝ ըստ վճարման եղանակների:



Նկ. 1. Վճարման տվյալների ցրվածությունը ըստ վճարման մեթոդի

Նախնական կարող ենք ենթադրել, որ վճարման գործոնը ունի ազդեցություն, սակայն այս ենթադրության նշանակալիությունը կարիք ունի վիճակագրական ստուգման՝ համապատասխան հայտանիշի կիրառմամբ:

Հետազոտության նպատակին համապատասխան իրականացվող վերլուծությունը մեզ հնարավորություն է տալու բացահայտել հաճախորդի ամսական վճարի՝ MonthlyCharges, կապը հաճախորդի ընտրած վճարման եղանակից՝ PaymentMethod: Այն է՝ պարզել վճարման եղանակը ունի ազդեցություն հաճախորդի ամսական վճարի վրա:

Այս հատկանիշների ընտրությունը թույլ է տալիս գնահատել տարբեր վճարման եղանակների կապը ամսական վճարների միջին արժեքների և բաշխման առանձնահատկությունների հետ, քանի որ վճարների մակարդակը և դրանց կատարման հարմարավետությունը կարող են ազդել հաճախորդների բավարարվածության վրա և, հետևաբար, պայմանավորել ծառայությունից հրաժարվելու կամ հաճախորդների հեռանալու հավանականությունը:

Տվյալների բաշխվածության Նկ.1-ում տեսնում ենք, որ տարբեր վճարման եղանակների խմբերում ամսական վճարների բաշխումները ամբողջությամբ սիմետրիկ չեն և պարունակում են որոշ արտառոց արժեքներ: Նաև՝ վճարման մեթոդով պայմանավորված է վճարումների քանակը: Սակայն գրաֆիկական վերլուծությունը ունի հիմնականում դիտողական բնույթ և չի տալիս վիճակագրականորեն հաստատված եզրակացություններ: Այդ պատճառով վիզուալ գնահատումից հետո իրականացվում են համապատասխան վիճակագրական թեստեր, որոնք թույլ են տալիս քանակականորեն ստուգել բաշխման հատկությունները և անհրաժեշտ նախապայմանները հետագա մեթոդների կիրառման համար:

Որակական գործոնի ազդեցությունը բացահայտող վիճակագրական վերլուծության միջոց է միագործոն ցրվածքային վերլուծությունը /ANOVA (Analysis of Variance)/, բայց դրա կիրառման համար անհրաժեշտ է երկու նախապայմանների ստուգում՝

1) խմբային տվյալների նորմալ բաշխվածություն;
2) խմբային տվյալների դիսպերսիաների հավասարություն[4]:

Նորմալ բաշխվածության ստուգումը կատարվել է Շապիրո–Ուիլքի (Shapiro-Wilk) թեստի միջոցով, որի արդյունքում հերքվել է նորմալ բաշխվածության վարկածը[4]: Տվյալները նախ խմբավորվում են ըստ վճարման մեթոդի (PaymentMethod հատկանիշի), որի արդյունքում ձևավորվում են առանձին խմբեր՝ համապատասխան տարբեր վճարման եղանակներին: Այնուհետև յուրաքանչյուր խմբի համար առանձին կիրառվում է Շապիրո–Ուիլքի թեստը՝ գնահատելու համար, թե արդյոք տվյալ խմբում MonthlyCharges փոփոխականի բաշխումը համապատասխանում է նորմալ բաշխմանը (հավելված 1):

Թեստի կիրառման արդյունքում յուրաքանչյուր վճարման եղանակի համար հաշվարկվում և արտածվում է համապատասխան p-value արժեքը, որը հնարավորություն է տալիս վիճակագրորեն գնահատել նորմալության հիպոթեզը՝

Աղյուսակ 3.2. Շապիրո–Ուիլքի թեստի արդյունքները

PaymentMethod	**Shapiro-Wilk test p-value**
Bank transfer	7.14×10^{-28}
Credit card	1.25×10^{-27}
Electronic check	1.15×10^{-30}
Mailed check	7.3×10^{-37}

Քանի որ բոլոր խմբերի համար ստացված p-արժեքները զգալիորեն փոքր են ընդունված $\alpha = 0.05$ նշանակալիության մակարդակից ($p < 0.05$), մենք մերժում ենք զրոյական վարկածը (H_0): Սա նշանակում է, որ MonthlyCharges փոփոխականը ըստ գործոնի մակարդակների չունի նորմալ բաշխում:

Կատարվել է խմբերի դիսպերսիաների հավասարության ստուգում՝ կիրառելով Լևենի թեստը (Levene's test): Նախ տվյալները խմբավորվում են ըստ PaymentMethod հատկանիշի, որի արդյունքում ձևավորվում են առանձին խմբեր՝ համապատասխան վճարման եղանակներին: Այնուհետև յուրաքանչյուր խմբի տվյալները հավաքվում են մեկ ցուցակի մեջ, որը փոխանցվում է stats.levene() ֆունկցիային, որպեսզի ստուգվի, արդյոք տարբեր խմբերի դիսպերսիաները վիճակագրորեն համասեռ են (հավելված 1): Արդյունքում հաշվարկված p-value=1.09×10^{-51}, որը բավականին փոքր է և մեծ նշանակալիությամբ մերժվում է զրոյական վարկածը,այսինքն խմբային դիսպերսիաները համասեռ չեն:

Այսպիսով դասական ANOVA-ի կիրառությունը հնարավոր չէ այս խնդրի լուծման համար: Հետևաբար պետք է կիրառել ոչ պարամետրական մոտեցում՝ Կրասկալ–Ուոլիսի թեստը:

Ի տարբերություն ANOVA-ի, որը ստուգում է խմբային միջինների հավասարությունը, Կրասկալ-Ուոլիսի թեստը հանդիսանում է ոչ պարամետրական այլընտրանք: Այն հիմնված է ռանգերի վրա, ինչը թույլ է տալիս գնահատել խմբերի միջև տարբերությունները՝ առանց բաշխվածության խիստ սահմանափակումների[4]:

Կրասկալ–Ուոլիսի թեստը ստուգում է հետևյալ վարկածները՝

--H_0: Բոլոր խմբերում ամսական վճարների մեդիանները նույնն են:

--H_1: Առնվազն մեկ խմբում ամսական վճարների բաշխումը էականորեն տարբերվում է մյուսներից:

Իրականացվել է Կրասկալ–Ուոլիսի թեստը՝ գնահատելու համար, թե արդյոք հաճախորդների ամսական վճարները (MonthlyCharges) տարբերվում են ըստ վճարման եղանակի (PaymentMethod): Առաջին քայլում տվյալները խմբավորվում են ըստ PaymentMethod հատկանիշի, որի արդյունքում ձևավորվում են առանձին խմբեր («Bank transfer», «Credit card», «Electronic check» և «Mailed check»): Այնուհետև յուրաքանչյուր խմբի MonthlyCharges արժեքները հավաքվում են ցուցակների մեջ և փոխանցվում stats.kruskal() ֆունկցիային՝ որպես առանձին խմբեր (հավելված 1):

Թեստը վերադարձնում է երկու հիմնական արդյունք՝ H-ստատիստիկա և p-value, որոնք թույլ են տալիս գնահատել խմբերի մեդիանների և դասակարգված արժեքների տարբերությունները՝ հաշվի առնելով վճարման մեթոդի ազդեցությունը՝

Աղյուսակ 3.3. Կրասկալ–Ուոլիսի թեստի արդյունքները

Kruskal-Wallis test H-statistics	Kruskal-Wallis test p-value
1103.30	7.0097e-239

Թեև Կրասկալ–Ուոլիսի թեստը հաճախ դիտարկվում է որպես մեդիանների հավասարության ստուգման ոչ պարամետրական անալոգ, այն ավելի ճշգրիտ կերպով գնահատում է խմբերի ռանգային բաշխվածությունների տարբերությունները:

Ստացված բարձր H-ստատիստիկան և չափազանց փոքր p-value-ն (Աղյուսակ 3.4) վկայում են այն մասին, որ ամսական վճարների (MonthlyCharges) բաշխվածությունը վիճակագրորեն նշանակալիորեն տարբերվում է ըստ վճարման եղանակների (PaymentMethod):

Քանի որ $p<0.05$, մենք մերժում ենք զրոյական վարկածը և եզրակացնում, որ առնվազն մեկ վճարման եղանակի դեպքում հաճախորդների ամսական վճարների ռանգերի միջինը

էականորեն տարբերվում է մյուսներից: *Սա նշանակում է, որ հաճախորդի կողմից ընտրված վճարման մեթոդը սերտորեն կապված է ամսական վճարի չափի հետ:*

Կրասկալ–Ուոլիսի արդյունքները ցույց տվեցին, որ առնվազն մեկ խմբի մեդիանները MonthlyCharges–ում զգալիորեն տարբերվում են մյուսներից:

Բնականաբար հաջորդ խնդիրն է, պարզել, որ խմբերի միջև է դիտվում տարբերություն: Խմբային զույգ-առ-զույգ ստուգման տարբեր թեստեր կան, որոնք մեկ ընդհանուր անվամբ են կոչվում՝ Post hoc test: Այս աշխատանքում ներկայացված է զույգ-առ-զույգ համեմատության Դանիի թեստի կիրառությունը:

Իրականացվել է Դանիի զույգ առ զույգ համեմատության թեստը՝ գնահատելու համար, թե վճարման կոնկրետ որ եղանակների (PaymentMethod) միջև են առկա վիճակագրորեն նշանակալի տարբերություններ ամսական վճարների (MonthlyCharges) բաշխման մեջ: Բազմակի համեմատությունների ժամանակ առաջին սեռի սխալը բացառելու նպատակով կիրառվել է Բոնֆերոնիի (Bonferroni) ուղղումը, որը վերահսկում է նշանակալիության մակարդակը բոլոր զույգերի համար (տես աղյուսակ 3,4)՝

Վերլուծությունը ցույց է տալիս, որ գոյություն ունի վիճակագրորեն խիստ նշանակալի տարբերություն վճարման տարբեր եղանակներ նախընտրող հաճախորդների ամսական վճարների (MonthlyCharges) միջև: Դանիի թեստը ցույց տվեց, որ «Bank transfer (automatic)» և «Credit card (automatic)» խմբերի միջև տարբերությունը վիճակագրորեն նշանակալի չէ (p = 1.0): Այսպիսով, կարող ենք հուսալիորեն պնդել, որ ավտոմատ վճարման համակարգերից օգտվող հաճախորդները ձևավորում են մեկ միասնական սեգմենտ, որոնց ամսական վճարների բաշխվածությունը գրեթե նույնական է:

Աղյուսակ 3.4. Դանիի զույգ առ զույգ համեմատությունների թեստի արդյունքները

	Bank transfer	Credit card	Electronic check	Mailed check
Bank transfer	1.00	1.00	2.18×10^{17}	8.24×10^{104}
Credit card	1.00	1.00	4.16×10^{20}	6.73×10^{97}
Electronic check	2.18×10^{17}	4.16×10^{20}	1.00	1.96×10^{234}
Mailed check	8.24×10^{104}	6.73×10^{97}	1.96×10^{234}	1.00

«Electronic check» տարբերակն օգտագործող հաճախորդներն ունեն ամենաբարձր ամսական վճարները: Թեստի արդյունքները հաստատում են, որ այս խումբը վիճակագրորեն նշանակալիորեն տարբերվում է մյուս բոլոր 3 խմբերից (բոլոր դեպքերում $p < 0.05$): «Mailed check» մեթոդն ունի ամենացածր ամսական վճարների բաշխվածությունը: Այն նույնպես վիճակագրորեն նշանակալիորեն տարբերվում է մյուս բոլոր խմբերից:

Եզրակացություն: Կրասկալ-Ուոլիսի թեստի և դրան հաջորդած Դանիի post-hoc վերլուծությունների արդյունքները թույլ են տալիս եզրակացնել, որ հաճախորդների ամսական վճարների չափը զգալիորեն կախված է նրանց ընտրած վճարման մեթոդից: Ստացված արդյունքները կարող են հիմք հանդիսանալ մարքեթինգային ռազմավարությունների մշակման և հաճախորդների հեռացման ռիսկերը կանխատեսելու համար, քանի որ վճարների բաշխվածության նման կտրուկ տարբերությունները հաճախ կապված են նաև հաճախորդների հավատարմության աստիճանի հետ:

3.2 Կրթության և ոլորտի ազդեցությունը աշխատավարձերի վրա. համեմատական վերլուծություն նախաճգնաժամային և ճգնաժամային ժամանակահատվածներում

Հետազոտության համար դիտարկվել է ԱՄՆ Միջին Ատլանտյան տարածաշրջանում 2003-2009 թվականներին աշխատող 3000 տղամարդ աշխատակիցների վերաբերյալ տվյալները: Այն ներառում է ինչպես աշխատավարձի տվյալներ, այնպես էլ մի շարք սոցիալ-դեմոգրաֆիկ և աշխատանքային բնութագրիչներ, որոնք կարող են ազդել աշխատավարձի ձևավորման վրա:

Տվյալների հավաքածուն բաղկացած է 9 հատկանիշներից, որոնցից հետազոտության համար առավել նշանակալի են հետևյալ հատկանիշները՝

— education՝ կրթական մակարդակը (<HS Grad, HS Grad, Some College, College Grad, Advanced Degree),

— jobclass՝ տնտեսության ոլորտը (Industrial, Information/IT),

— wage՝ աշխատավարձը:

Տարածաշրջանը առանձնանում է իր ընդգրկած ընկերությունների կարևորությամբ, քանի որ այն հանդիսանում է ԱՄՆ-ի տնտեսական և ֆինանսական զարկերակը: Այստեղ են կենտրոնացված աշխարհի խոշորագույն ֆինանսական հաստատությունները (Wall Street-ի բանկերն ու ներդրումային ընկերությունները), դեղագործական հսկաները (օրինակ՝ Johnson & Johnson, Merck), տեխնոլոգիական կորպորացիաները և բազմաթիվ Fortune 500 ընկերությունների կենտրոնակայանները: Այս տվյալների բազան առանձնահատուկ կարևորություն է ձեռք բերում, քանի որ այն ծածկում է 2003-2009 թվականները: Այս տարածաշրջանում աշխատավարձերը սովորաբար ավելի բարձր են, քան ԱՄՆ-ի միջինը:

2003-2006 թթ. այս ընկերությունները մեծ պահանջարկ ունեին մագիստրոսի և դոկտորի աստիճան ունեցողների նկատմամբ (Advanced Degree)՝ բարդ ալգորիթմներ և համակարգեր ստեղծելու համար [10]: Սակայն 2008 թ. ճգնաժամը շրջադարձային էր, այն մեծ վնաս հասցրեց ընկերություններին և շատ ընկերություններ ընտրեցին աշխատատեղերի և աշխատավարձերի կրճատման քաղաքականությունը [11]:

Հաշվի առնելով այս դրված են հետևյալ հետազոտական հարցերը՝

1. Ինչպե՞ս են փոխազդում կրթական մակարդակը և աշխատանքային ոլորտը (Արդյունաբերական և IT) աշխատավարձի ձևավորման վրա:
2. Արդյո՞ք բարձրագույն գիտական աստիճանը ունի՞ ազդեցություն աշխատավարձի վրա երկու ոլորտներում էլ:
3. Արդյո՞ք 2008 թվականի համաշխարհային տնտեսական ճգնաժամը ազդեցություն ունեցել է այս երկու ոլորտներում ըստ կրթական մակարդակների աշխատավարձի բաշխվածության վրա:

Հետազոտության ընթացքում տվյալների մշակումը, վիճակագրական վերլուծությունը և վիզուալիզացիան իրականացվել են R ծրագրավորման միջավայրում՝ կիրառելով հետևյալ գրադարանները [12].

-- ISLR – օգտագործվել է Wage տվյալների հավաքածուի ներմուծման և նախնական ուսումնասիրության համար,

-- tidyverse – կիրառվել է տվյալների մաքրման, տրանսֆորմացիայի և կառուցվածքային վերափոխման նպատակով,

-- ggplot2 – օգտագործվել է գրաֆիկական ներկայացումների կառուցման համար, ,

-- dplyr – կիրառվել է տվյալների ֆիլտրացիայի, խմբավորման համար,

-- stats – օգտագործվել է հիմնական վիճակագրական թեստերի իրականացման համար (օրինակ՝ Kruskal–Wallis և Wilcoxon թեստեր),

-- FSA – կիրառվել է Դանի post–hoc զույգ առ զույգ համեմատությունների իրականացման համար,

-- ARTool – օգտագործվել է ոչ պարամետրական երկգործոն դիսպերսիոն վերլուծության (ART ANOVA) իրականացման համար։

Հետազոտական հարցին պատասխանելու համար կիրառվել է փուլային վերլուծություն, որը ներառում է ինչպես նկարագրական, այնպես էլ վիճակագրական մեթոդներ։

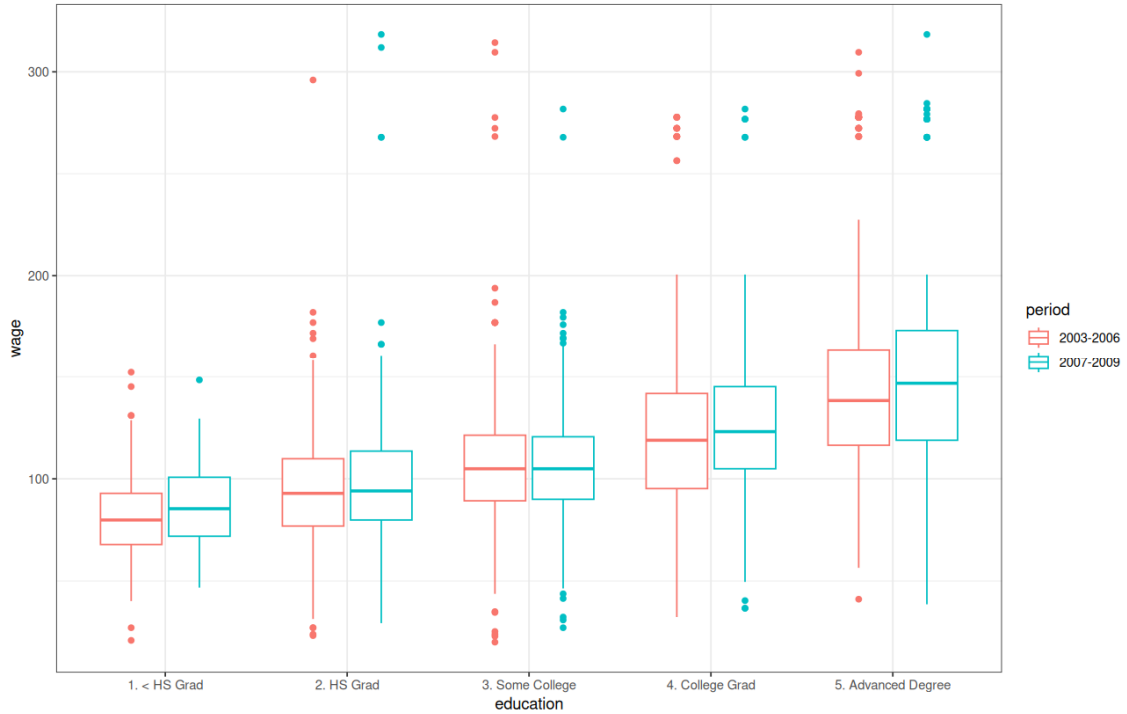
Հետազոտության մեթոդաբանական հիմքում ընկած է տվյալների ժամանակային բաժանման սկզբունքը՝ նախաճգնաժամային (2003-2006թթ.) և ճգնաժամային (2007-2009թթ.) ժամանակահատվածների, ինչը բացառում է ճգնաժամի հետևանքով առաջացած փոփոխությունների լուծվելուն ընհանուր տվյալների մեջ։ Այսպիսի ռազմավարությունը թույլ է տալիս խուսափել տվյալների ագրեգացման արդյունքում տեղեկատվական կորստից։

3.2.1 Տվյալների նախնական վերլուծություն

Վերլուծության առաջին փուլում, որպես տվյալների հետազոտական վերլուծության գործիք, կիրառվել են արկղային դիագրամները (boxplots)։ Արկղային դիագրամները հնարավորություն են տալիս վիզուալ կերպով գնահատել աշխատավարձերի բաշխման ոչ

միայն կենտրոնական միտումը (մեդիանը), այլև դիսպերսիան, բաշխման ասիմետրիան և արտառոց արժեքների (outliers) առկայությունը:

Նկ 2.-ը հաստատում է կրթական մակարդակի և աշխատավարձերի միջև ուղիղ կապը. անկախ ժամանակահատվածից, կրթության աստիճանի բարձրացմանը զուգընթաց դիտվում է աշխատավարձի կայուն աճ: 2007-2009թթ. տնտեսական ճգնաժամի ընթացքում աշխատավարձերի անվանական մակարդակը չի կրել կտրուկ անկում, սակայն 2003-2009թթ. ընթացքում գնաճի առկայության պայմաններում, անվանական աշխատավարձի անփոփոխ լինելը փաստացի նշանակում է իրական աշխատավարձի և գնողականության նվազում:

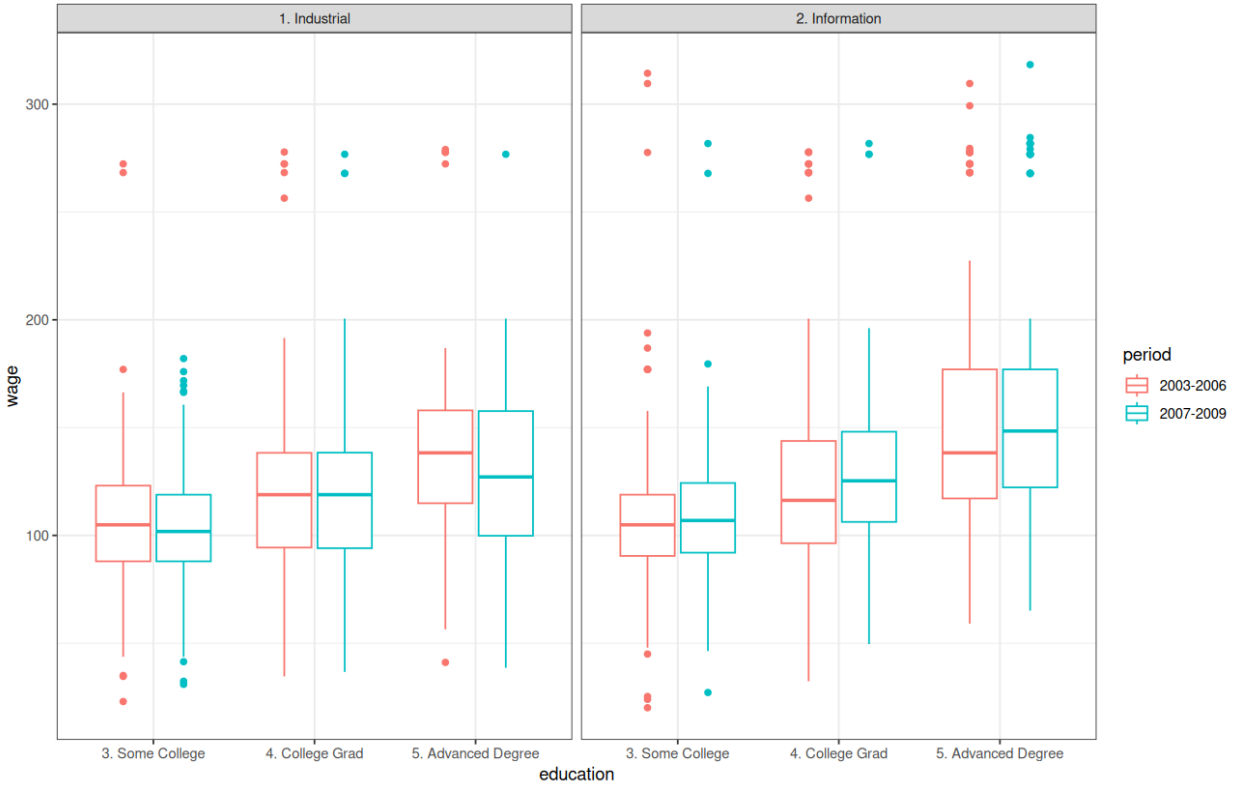


Նկ. 2. Աշխատավարձերի բաշխումը ըստ կրթական մակարդակի (2003-2006 և 2007-2009 թթ.)

Հատկանշական է, որ բարձրագույն կրթություն (Advanced Degree) ունեցողների մոտ ճգնաժամային տարիներին դիտվում է մեդիան աշխատավարձի փոքր աճի միտում, ինչը կարող է վկայել այն մասին, որ բարձր որակավորում ունեցող աշխատուժը ճգնաժամի ժամանակ ավելի լավ է պահպանել իր դիրքերը: Գրաֆիկում նաև արտացոլվում են

արտառոց արժեքներ. դրանք այն բարձր վարձատրվող աշխատակիցներն են, որոնց վրա չի ազդել ճգնաժամը:

Կրթական մակարդակները ֆիլտրվում են առանձնացնելով միայն բարձր մասնագիտական կրթություն ունեցող խմբերը: Նպատակն է առանձնացնել և վերլուծել այն աշխատուժի դինամիկան, որն առավելագույնս ներգրավված է մտավոր և կառավարչական ոլորտներում, ինչը թույլ է տալիս գնահատել, թե ճգնաժամային պայմաններում որքանով են պաշտպանված կամ խոցելի այս մասնագիտական խմբերը՝ համեմատած ընդհանուր շուկայի հետ:



Նկ. 3. Աշխատավարձերի բաշխումը ըստ ոլորտների և կրթական մակարդակների (2003-2006 և 2007-2009 թթ.)

Ըստ տնտեսության ոլորտի վերլուծությունը վկայում է ճգնաժամի անհամաչափ ազդեցության մասին: Մինչ տեղեկատվական (Information) ոլորտում աշխատավարձերի մակարդակը պահպանել է իր կայունությունը և անգամ ցուցաբերել աճի միտումներ, արդյունաբերական (Industrial) ոլորտի բարձր որակավորում ունեցող աշխատուժի

շրջանում արձանագրվել է աշխատավարձերի անվանական նվազում: Սա թույլ է տալիս եզրակացնել, որ տնտեսական ցնցումների ժամանակ արդյունաբերական ձեռնարկությունները ավելի զգայուն են:

Գրաֆիկական ներկայացման միջոցով ստացված նախնական արդյունքների վիճակագրական հավաստիացման և խորքային ուսումնասիրության նպատակով՝ հետազոտության շրջանակներում յուրաքանչյուր ժամանակահատվածի համար իրականացվում է վերլուծության հաջորդական գործընթաց:

3.2.2 Նախաճգնաժամային ժամանակահատվածի (2003—2006 թթ.) աշխատավարձի տվյալների վերլուծություն

Առաջին փուլով կատարվում է նախաճգնաժամային ժամանակահատվածի ուսումնասիրություն: Իրականացվում է նկարագրական վիճակագրության բնութագրիչների հաշվարկ, որը թույլ է տալիս քանակական առումով ֆիքսել յուրաքանչյուր կրթական մակարդակի և ոլորտի համար բնորոշ աշխատավարձային ցուցանիշները:

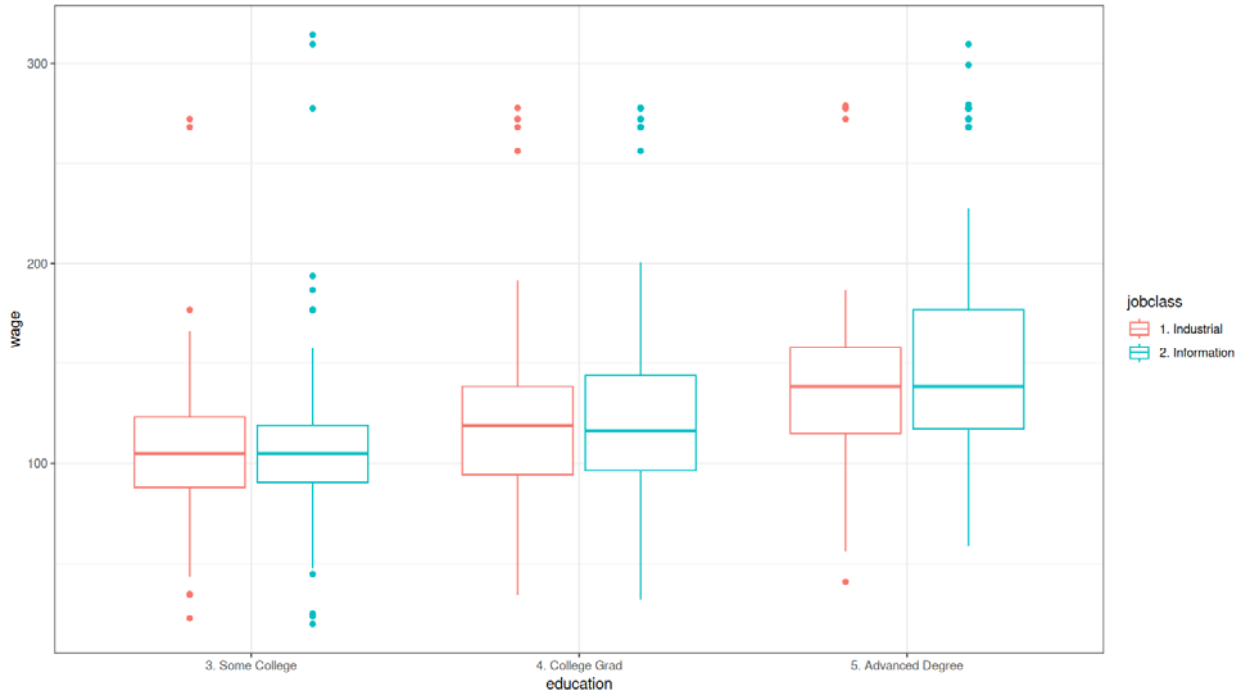
Աղյուսակ 3.5. Աշխատավարձի վիճակագրական բնութագրիչների հաշվարկ

	2003--2006			**2003—2006**		
	Industrial			**Information**		
	Some College	College Grad	Advanced Degree	Some College	College Grad	Advanced Degree
Min	22.96	34.60	41.10	20.09	32.37	59.10
Q_1	87.98	94.07	109.70	90.48	94.22	114.70
Median	105.43	117.30	128.20	105.59	117.61	136.67
Mean	104.41	118.88	134.71	104.41	114.48	134.34
Q_3	123.09	136.16	152.52	118.88	138.83	152.52
Max	176.99	191.57	186.88	193.87	200.54	227.46
Sd	26.80	30.64	33.36	26.17	33.12	33.59

Աղյուսակ 3.7-ը փաստում է, որ ինչպես Industrial, այնպես էլ Information ոլորտներում աշխատավարձի մեդիանը հետևողականորեն աճում է կրթության մակարդակին զուգընթաց: Սա հիմնավորում է, որ նախաճգնաժամային շրջանում շուկան գնահատում

էր բարձրագույն կրթությունը: Information ոլորտում աշխատավարձերի տարածվածությունը ավելի մեծ է, ինչը երևում է թե՛ առավելագույն արժեքներից, թե՛ ստանդարտ շեղման ցուցանիշներից:

Արկղային դիագրամի միջոցով վիզուալ ներկայացնենք աղյուսակ 3.7–ի արդյունքները.



Նկ. 4. Աշխատավարձերի բաշխումն ըստ տնտեսական ոլորտների և կրթական մակարդակի (2003-2006թթ.)

Արկղերի տեղակայումը և ծավալը լիովին համապատասխանում են աղյուսակի հաշվարկային տվյալներին (մեդիան, քառորդներ), ինչը վկայում է դիտարկվող շրջանի տնտեսական ցուցանիշների կայունության մասին:

3.2.2.1 Կրթական մակարդակի աշխատավարձի չափի ազդեցության վերլուծություն ըստ ոլորտների

Վիզուալ վերլուծության արդյունքում ձևավորված նախնական վարկածների վիճակագրական նշանակալիությունը հիմնավորելու նպատակով իրականացվել է վիճակագրական թեստավորում:

Քանի որ աշխատավարձերի տվյալների բաշխումը չի համապատասխանում նորմալ բաշխմանը, կիրառվել է Կրասկալ–Ուոլիսի թեստը՝ ստուգելու համար, թե արդյոք կրթական մակարդակների միջև աշխատավարձերի տարբերությունները վիճակագրորեն նշանակալի են, թե ոչ:

Թեստը կիրառվում է տնտեսության ոլորտի առանձին մակարդակների համար: Ըստ զորյական վարկածի կրթական մակարդակները որևէ էական ազդեցություն չունեն աշխատավարձերի բաշխման վրա:

Աղյուսակ 3.6. Կրասկալ–Ուոլիսի թեստի արդյունքները

Ոլորտ	Kruskal-Wallis test p-value	Արդյունք
Industrial	2.95×10^{-8}	Տարբերվում են (p-value < 0.05)
Information	2.2×10^{-16}	Տարբերվում են (p-value < 0.05)

Ստացված p–value-ի չափազանց ցածր արժեքը թույլ է տալիս մերժել զրոյական վարկածը, հետևաբար, արձանագրված վիճակագրական արդյունքները հաստատում են վիզուալ վերլուծության միջոցով ձևավորված այն ենթադրությունը, որ կրթական մակարդակը հանդիսանում է և՛ տեղեկատվական, և՛ արդյունաբերական ոլորտում աշխատավարձերի տարբերությունը պայմանավորող էական գործոն:

*Աղյուսակ 3.7. Արդյունաբերական (*Industrial*) ոլորտի տվյալների Դանիի զույգ առ զույգ համեմատությունների թեստի արդյունքները*

Համեմատություն	p-value	p-value.adj	Արդյունք
Some College vs College Grad	7.97×10^{-5}	0.000159	Տարբերվում են (p-value < 0.05)
Some College vs Advanced Degree	6.95×10^{-8}	0.0000002	Տարբերվում են (p-value < 0.05)
College Grad vs Advanced Degree	9.61×10^{-3}	0.0096	Տարբերվում են (p-value < 0.05)

Կրասկալ–Ուոլիսի թեստի արդյունքը ցույց չի տալիս, թե կոնկրետ որ խմբերի միջև է էական տարբերություն նկատվում: Այդ նպատակով կիրառվում է Դանիի զույգ առ զույգ համեմատությունների մեթոդը:

Դանի թեստի կիրառմամբ ստացվում է արդյունքը արդյունաբերական (Industrial) ոլորտի համար բերված է աղյուսակ 3,7-ում:

Զույգային համեմատությունների իրականացումը ցույց տվեց, որ Industrial ոլորտում աշխատավարձերի տարբերությունները բոլոր կրթական մակարդակների միջև վիճակագրորեն նշանակալի են: Մասնավորապես, Advanced Degree ունեցող մասնագետների վարձատրությունը էապես գերազանցում է թե՛ College Grad, և թե՛ Some College մակարդակի աշխատողների եկամուտներին: Սա թույլ է տալիս եզրակացնել, որ կրթական մակարդակը ոչ միայն ընդհանուր առմամբ է ազդում վարձատրության վրա, այլև յուրաքանչյուր կրթական աստիճան ապահովում է աշխատավարձի վիճակագրորեն շոշափելի հավելում:

Դանիի զույգ առ զույգ համեմատությունների թեստը արդյունքը տեղեկատվական (Information) ոլորտի համար բերված է աղյուսակ 3,8-ում:

Աղյուսակ 3.8. Տեղեկատվական (Information) ոլորտի տվյալների Դանիի զույգ առ զույգ համեմատությունների թեստի արդյունքները

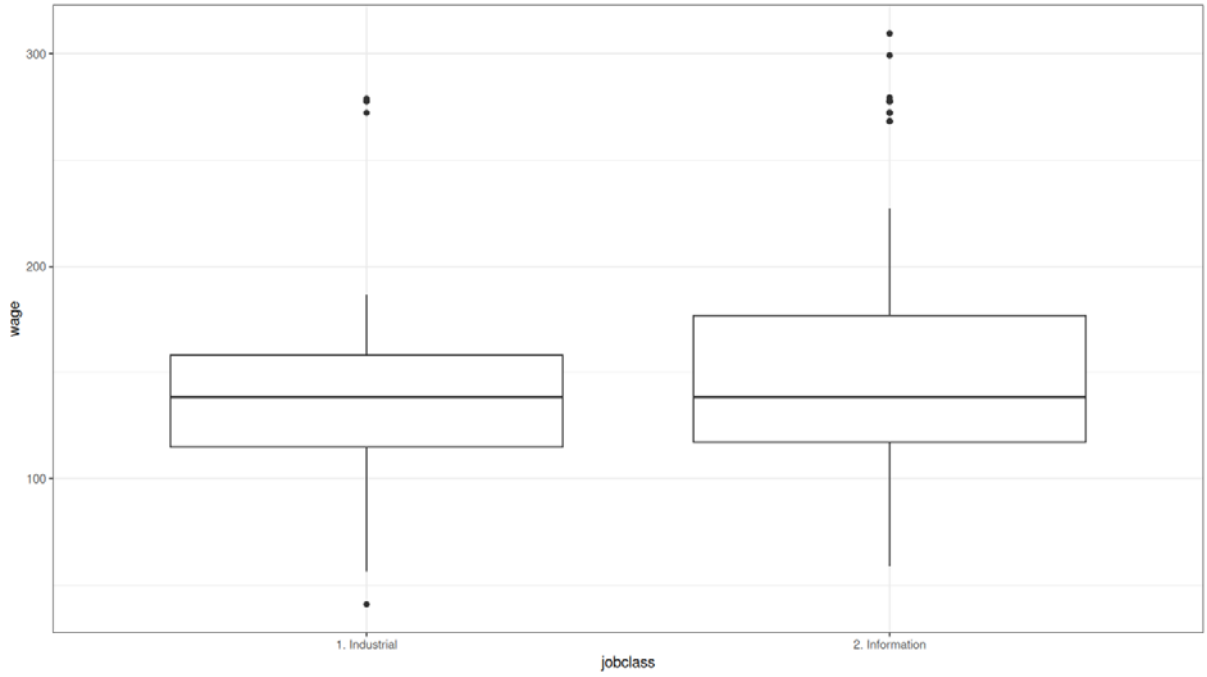
Համեմատություն	p-value	p-value.adj	Արդյունք
Some College vs College Grad	6.68×10^{-6}	0.00002	Տարբերվում են (p-value < 0.05)
Some College vs Advanced Degree	2.43×10^{-23}	7.28×10^{-23}	Տարբերվում են (p-value < 0.05)
College Grad vs Advanced Degree	4.86×10^{-9}	0.000000015	Տարբերվում են (p-value < 0.05)

Զույգային համեմատությունների իրականացումը ցույց տվեց, որ տեղեկատվական ոլորտում աշխատավարձերի տարբերությունները բոլոր կրթական խմբերի միջև վիճակագրորեն նշանակալի են: Հատկանշական է Some College և Advanced Degree խմբերի միջև արձանագրված առավելագույն տարբերությունը, ինչը ընդգծում է մարդկային կապիտալի և ակադեմիական որակավորման բարձր եկամտաբերությունը տեղեկատվական ոլորտում: Ստացված արդյունքները թույլ են տալիս եզրակացնել, որ Information ոլորտում կրթական մակարդակը աշխատավարձի ձևավորման հիմնարար գործոններից է:

3.2.2.2 Կրթական ամենաբարձ մակարդակի (Advanced Degree) աշխատավարձերի միջոլորտային համեմատություն

Նախորդ փուլերում հաստատվեց, որ կրթությունը յուրաքանչյուր ոլորտի ներսում ազդում է աշխատավարձի վրա: Ոլորտային առանձնահատկություններն ավելի խորությամբ ուսումնասիրելու նպատակով՝ իրականացվել է միջոլորտային համեմատություն՝ կրթական ամենաբարձր մակարդակի (Advanced Degree) շրջանակներում: Վիլկոկսոնի (Wilcoxon) ոչ պարամետրական թեստի կիրառումը հնարավորություն է տալիս պարզելու, թե արդյոք առկա է վիճակագրորեն նշանակալի տարբերություն նույն

որակավորումն ունեցող մասնագետների վարձատրության միջև՝ կախված տնտեսության ոլորտից: Ըստ զրոյական վարկածի տեղեկատվական (Information) և արդյունաբերական (Industrial) ոլորտներում բարձրագույն գիտական աստիճան ունեցող մասնագետների աշխատավարձերի բաշխումների միջև վիճակագրորեն նշանակալի տարբերություն չկա:



Նկ. 5. Աշխատավարձերի բաշխումը ըստ ոլորտների (Advanced Degree)

Գրաֆիկական ներկայացումը ցույց է տալիս, որ տեղեկատվական և արդյունաբերական ոլորտներում բարձրագույն գիտական աստիճան ունեցող մասնագետների աշխատավարձերի բաշխումները մեծապես համընկնում են: Մեդիանների և միջքառորդային միջակայքների նմանությունը հաստատում է Վիլկոկսոնի թեստի արդյունքը՝ $p = 0.1971$, ինչը վկայում է վիճակագրորեն նշանակալի տարբերության բացակայության մասին:

Չնայած նախորդ թեստերը ցույց տվեցին, որ երկու ոլորտներում էլ կրթությունը մեծ դեր ունի, այս կոնկրետ արդյունքը փոխում է պատկերը բարձրագույն որակավորում ունեցողների համար: Քանի որ $p > 0.05$, մենք չենք կարող մերժել զրոյական վարկածը: $W = 5008$ թեստի վիճակագրական արժեքն է, որը տվյալ դեպքում ցույց է տալիս երկու խմբերի աշխատավարձերի բաշխվածության համադրելիությունը: Սա թույլ է տալիս եզրակացնել,

որ բարձրագույն կրթական աստիճանը հանդիսանում է «հավասարեցնող» գործոն, որը մասնագետի համար ապահովում է բարձր վարձատրություն՝ անկախ տնտեսության ճյուղից: Այսպիսով, աշխատաշուկայի տարանջատումը ըստ ոլորտների ավելի ցայտուն է դրսևորվում կրթական ցածր և միջին մակարդակներում, մինչդեռ բարձրագույն որակավորման տիրույթում աշխատավարձերը միտված են մերձեցման:

3.2.2.3 Աշխատավարձի վրա կրթական մակարդակի և ոլորտի համատեղ ազդեցության վերլուծություն

Համապարփակ հետազոտության համար վերջին փուլում կատարվում է երկգործոն վերլուծություն: ART ANOVA-ն ոչ պարամետրական մոտեցում է, որը թույլ է տալիս գնահատել գլխավոր ազդեցությունները (main effects) և փոխազդեցության էֆֆեկտը (interaction effect) և միևնույն ժամանակ չպահանջել տվյալների նորմալ բաշխում:

Երկգործոն վերլուծությունը կիրառելու նպատակով ձևակերպում են հետևյալ վիճակագրական վարկածները՝

Կրթության ազդեցությունը (Main Effect 1)

H_0: Կրթական մակարդակը չի ազդում աշխատավարձի մեդիանի վրա:

H_1: Առնվազն մեկ կրթական մակարդակի մեդիան **տարբերվում է մյուսներից:**

Ոլորտի ազդեցությունը (Main Effect 2)

H_0: IT և Industry ոլորտների միջև աշխատավարձերի մեդիանային տարբերություն չկա:

H_1: IT ոլորտի աշխատավարձերը նշանակալիորեն տարբերվում են արդյունաբերական ոլորտի աշխատավարձերից:

Փոխազդեցության էֆեկտը (Interaction Effect)

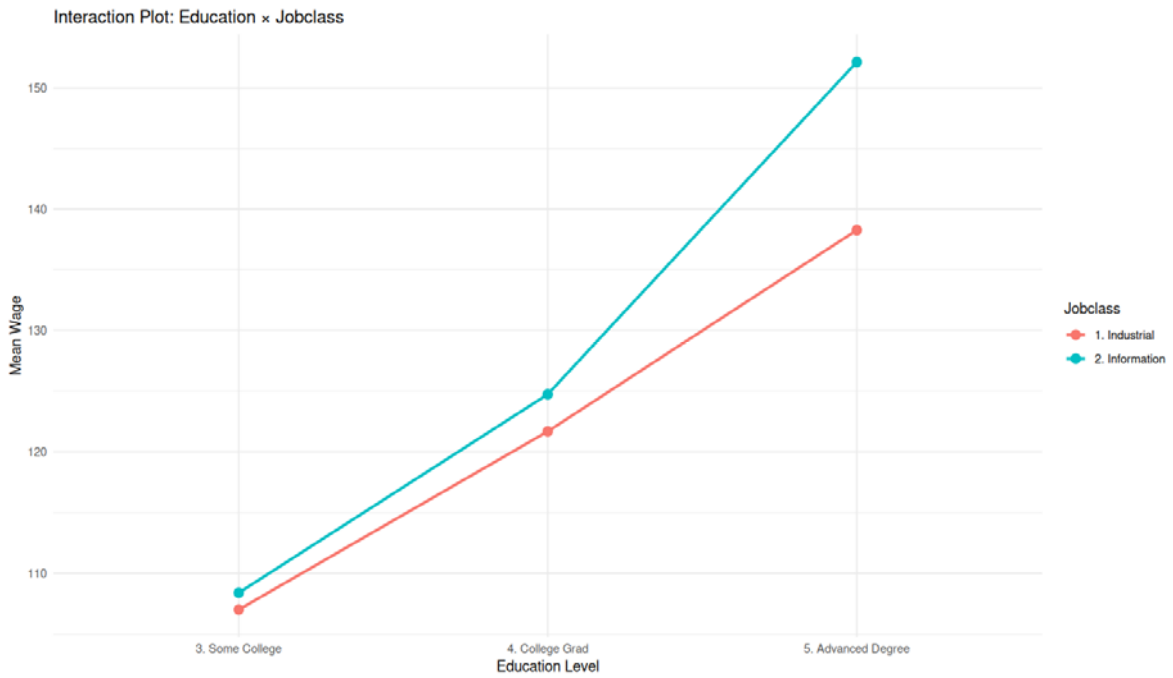
H_0: Կրթության ազդեցությունը աշխատավարձի վրա կախված չէ ոլորտից (IT թե Industry):

H_1: Գոյություն ունի փոխազդեցություն. կրթության ազդեցությունը աշխատավարձի վրա տարբեր է կախված ընտրված ոլորտից (օրինակ՝ Industry-ում դիպլոմն ավելի շատ է բարձրացնում աշխատավարձը, քան IT-ում):

Աղյուսակ 3.9. Երկգործոն վերլուծության արդյունքները

Գործոն	p-value	Ազդեցություն
Կրթության մակարդակ	2.14×10^{-29}	Նշանակալի
Տնտեսության ոլորտ	8.45×10^{-4}	Նշանակալի
Փոխազդեցություն	0.287	Բացակայում է

Կրթական մակարդակը և ոլորտը անկախաբար և վիճակագրորեն նշանակալի կերպով ազդում են աշխատավարձի վրա: Education:Jobclass փոխազդեցության p-արժեքը հիմք է տալիս պնդելու, որ կրթության ազդեցությունը աշխատավարձի վրա միատեսակ է ինչպես Industrial, այնպես էլ Information ոլորտներում: Քանի որ փոխազդեցությունը նշանակալի չէ, դա նշանակում է, որ եթե մենք կառուցենք աշխատավարձերի փոփոխության գրաֆիկը, ապա ոլորտների գծերը կլինեն գրեթե զուգահեռ՝ ցույց տալով, որ կրթության դերը չի փոխվում ոլորտից կախված:



Նկ. 6. Աշխատավարձի վրա կրթական մակարդակի և տնտեսության ոլորտի փոխազդեցության գրաֆիկ

Գծերը գրեթե զուգահեռ են։ Դա նշանակում է, որ կրթական մակարդակի ազդեցությունը աշխատավարձի վրա նույնն է ինչպես արդյունաբերական, այնպես էլ տեղեկատվական ոլորտներում։ Սա վիզուալ հաստատումն է այն բանի, որ ANOVA մոդելում education:jobclass փոխազդեցությունը նշանակալի չէր։

Կապույտ գիծը (Information) ամբողջությամբ կարմիր գծի (Industrial) վերևում է՝ բոլոր կրթական մակարդակների համար։ Սա ցույց է տալիս, որ Information ոլորտում վարձատրությունն ավելի բարձր է՝ անկախ կրթական մակարդակից։ Սա հաստատում է jobclass-ի նշանակալիությունը։

Երկու գծերն էլ ունեն կտրուկ վերընթաց տեսք։ Սա ակնհայտ ցույց է տալիս կրթական մակարդակի ուժեղ ազդեցությունը աշխատավարձի վրա, որը վիճակագրության մեջ ամենանշանակալի գործոնն էր։

Նախաճգնաժամային շրջանում պարզվեց, որ կրթական մակարդակը աշխատավարձի վրա ամենաուժեղ ազդեցություն ունեցող գործոնն է, և այդ ազդեցությունը նույնն է թե՛ Industrial, թե՛ Information ոլորտներում։ Միևնույն ժամանակ, բարձրագույն գիտական աստիճանը (Advanced Degree) «հավասարեցնում» է ոլորտային տարբերությունները՝ ապահովելով համադրելի աշխատավարձեր։

3.2.3 Ճգնաժամային ժամանակահատվածի (2007—2009 թթ.) աշխատավարձի տվյալների վերլուծություն

Այժմ անցնում ենք ճգնաժամային ժամանակաշրջանի (2007-2009թթ.) վերլուծությանը՝ գնահատելու, թե արդյոք 3.2.2 բաժնում ստացված օրինաչափությունները պահպանվում են տնտեսական անկման պայմաններում, թե ոչ։

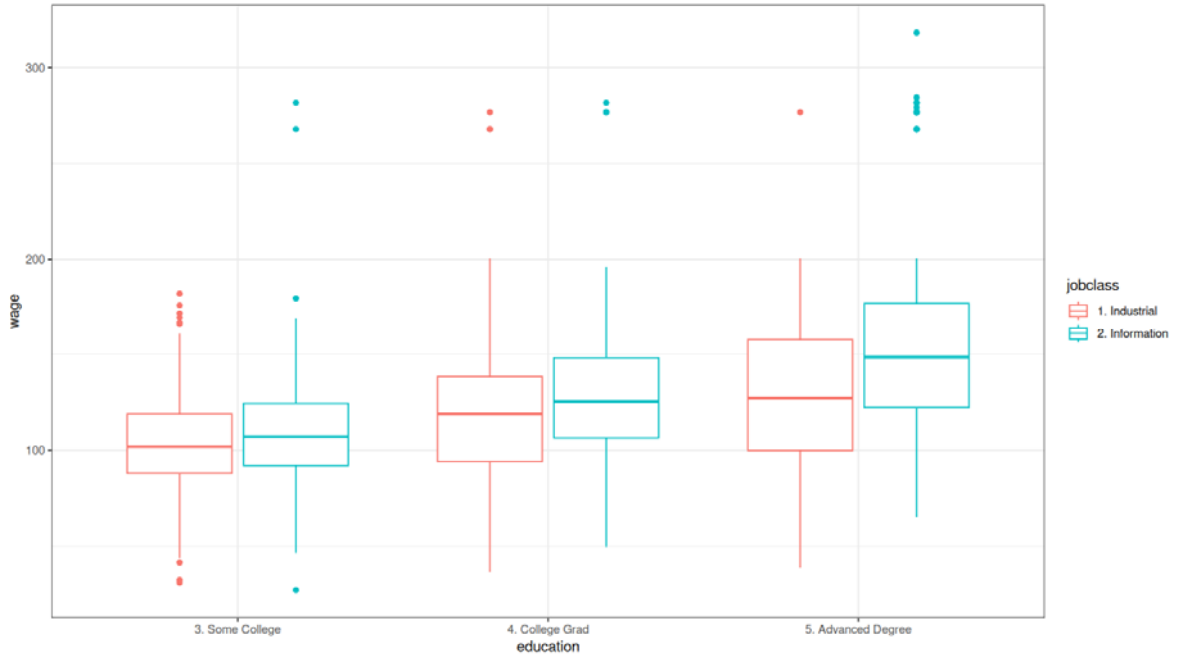
Առաջին քայլում իրականացվում է նկարագրական վիճակագրության բնութագրիչների հաշվարների արդյունքները բերված են աղյուսակ 3,10-ում։

Աղյուսակ 3.10. Աշխատավարձի վիճակագրական բնութագրիչների հաշվարկ

	2007--2009			2007--2009		
	Industrial			Information		
	Some College	College Grad	Advanced Degree	Some College	College Grad	Advanced Degree
Min	30.92	36.67	38.61	27.14	49.56	65.11
Q_1	87.98	94.07	99.85	91.99	106.26	122.26
Median	101.82	118.88	127.12	106.92	125.32	148.41
Mean	105.16	120.35	132.95	111.49	129.12	160.60
Q_3	118.88	138.38	157.66	124.31	148.09	176.99
Max	182.02	276.78	276.78	281.76	281.75	318.34
Sd	28.95	41.76	43.38	34.41	35.42	56.10

Information ոլորտի միջին և մեդիան աշխատավարձերը բոլոր կրթական մակարդակներում ավելի բարձր են, քան Industrial ոլորտինը: Advanced Degree ունեցողների մոտ, հատկապես Information ոլորտում, ստանդարտ շեղումը շատ ավելի մեծ է, քան այլ խմբերում: Սա կարող է նշանակել, որ ճգնաժամի ժամանակ տեղեկատվական ոլորտի բարձրակարգ մասնագետների եկամուտները դարձել են ավելի անկայուն:

Արկղային դիագրամի միջոցով վիզուալ ներկայացնենք աղյուսակ 3.12–ի արդյունքները.



Նկ. 7. Աշխատավարձերի բաշխումն ըստ տնտեսական ոլորտների և կրթական մակարդակի (2007-2009թթ.)

Վիճակագրական վերլուծությունը հաստատում է, որ կրթական մակարդակը և ոլորտային պատկանելությունը ճգնաժամային տարիներին շարունակել են մնալ աշխատավարձի որոշիչ գործոններ՝ առավել բարձր եկամուտ ապահովելով տեղեկատվական ոլորտի և բարձրագույն կրթությամբ մասնագետների համար:

3.2.3.1 Կրթական մակարդակի աշխատավարձի չափի ազդեցության վերլուծություն ըստ ոլորտների

Համանման միջճգնաժամային ժամանակաշրջանի վերլուծության, ճգնաժամային ժամանակաշրջանի համար Կրասկալ–Ուոլիսի և Դանիի տեստի արդյունքները բերված աղյուսակ 3,11-ում և 3,12- 3,13-ում համապատասխանաբար:

Աղյուսակ 3.11. Կրասկալ–Ուոլիսի թեստի արդյունքները

Ոլորտ	Kruskal-Wallis test p-value	Արդյունք
Industrial	5.38×10^{-5}	Տարբերվում են (p-value < 0.05)
Information	2.2×10^{-16}	Տարբերվում են (p-value < 0.05)

Աղյուսակ 3.12. Դանի զույգ առ զույգ համեմատությունների թեստի արդյունքները

Համեմատություն	p-value	p-value.adj	Արդյունք
Some College vs ollege Grad	2.34×10^{-3}	0.007	Տարբերվում են (p-value < 0.05)
Some College vs Advanced Degree	5.34×10^{-5}	0.00016	Տարբերվում են (p-value < 0.05)
College Grad vs Advanced Degree	9.33×10^{-2}	0.28	Չեն տարբերվում (p-value > 0.05)

Ըստ բերված վիճակլագրական տվյալների, ի տարբերություն Industrial ոլորտի, որտեղ բարձրագույն և գիտական աստիճանների միջև վարձատրության էական տարբերություն

չէր նկատվել, Information ոլորտում կրթական յուրաքանչյուր հաջորդ աստիճանը (ներառյալ Advanced Degree-ն) բերում է աշխատավարձի վիճակագրորեն նշանակալի և հաստատուն աճի: Սա փաստում է տեղեկատվական ոլորտի մեջ բարձրագույն մասնագիտական որակավորման նկատմամբ շարունակական և բարձր պահանջարկի մասին՝ նույնիսկ ճգնաժամային պայմաններում:

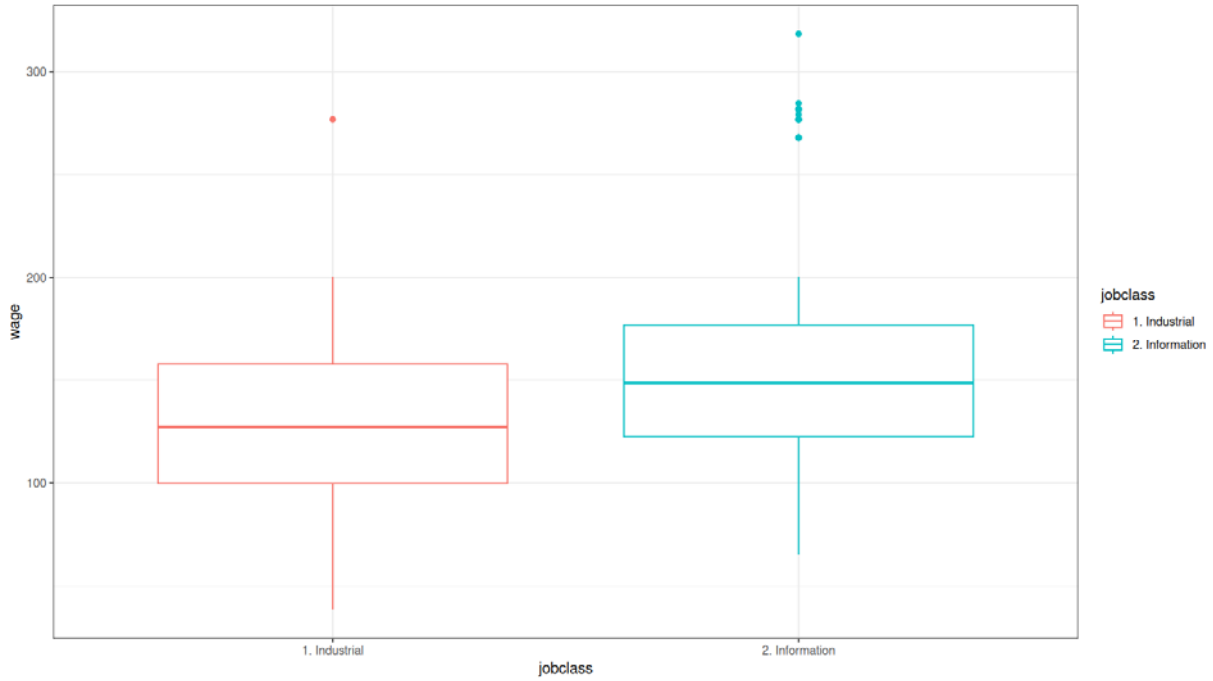
Աղյուսակ 3.13. Դանիի զույգ առ զույգ համեմատությունների թեստի արդյունքները

Համեմատություն	p-value	p-value.adj	Արդյունք
Some College vs College Grad	8.89×10^{-6}	0.000027	Տարբերվում են (p-value < 0.05)
Some College vs Advanced Degree	4.86×10^{-18}	1.46×10^{-17}	Տարբերվում են (p-value < 0.05)
College Grad vs Advanced Degree	3.04×10^{-7}	0.0000009	Տարբերվում են (p-value < 0.05)

3.2.3.2 Կրթական ամենաբարձ մակարդակի (Advanced Degree) աշխատավարձերի միջոլորտային համեմատություն

Ոլորտային առանձնահատկություններն ավելի խորությամբ ուսումնասիրելու նպատակով՝ իրականացվել է միջոլորտային համեմատություն՝ կրթական ամենաբարձր մակարդակի (Advanced Degree) շրջանակներում: Վիլկոկսոնի (Wilcoxon) ոչ պարամետրական թեստի կիրառումը հնարավորություն է տալիս պարզելու, թե արդյոք առկա է վիճակագրորեն նշանակալի տարբերություն նույն որակավորումն ունեցող մասնագետների վարձատրության միջև՝ կախված տնտեսության ոլորտից: Ըստ զրոյական վարկածի տեղեկատվական (Information) և արդյունաբերական (Industrial) ոլորտներում բարձրագույն գիտական աստիճան ունեցող մասնագետների աշխատավարձերի բաշխումների միջև վիճակագրորեն նշանակալի տարբերություն չկա:

Վիլկոկսոնի թեստի p-value–ն ստացվել է 0.00398, իսկ W ստատիստիկան 2023։ Վիլկոկսոնի ոչ պարամետրական թեստի արդյունքները թույլ են տալիս հաստատել, որ 2007-2009 թթ. ճգնաժամային ժամանակաշրջանում, նույնիսկ կրթական ամենաբարձր մակարդակի (Advanced Degree) պայմաններում, պահպանվել է վարձատրության վիճակագրորեն նշանակալի միջոլորտային անհավասարություն։ Տեղեկատվական ոլորտի բարձրագույն որակավորում ունեցող մասնագետները շարունակել են վարձատրվել ավելի բարձր, քան արդյունաբերական ոլորտի նույն որակավորմամբ մասնագետները, ինչը վկայում է տեղեկատվական ոլորտի ավելի բարձր դիմադրողականության և պահանջարկի մասին՝ անկախ տնտեսական ճգնաժամից։



Նկ. 8. Աշխատավարձերի բաշխումը ըստ ոլորտների (Advanced Degree)

Գրաֆիկական ներկայացումը վերջնականապես հաստատում է ստացված p=0.0039 արժեքի նշանակալիությունը. նույնիսկ ամենաբարձր որակավորման դեպքում ոլորտների միջև առկա է եկամտային տարբերություն՝ հօգուտ տեղեկատվական ոլորտի։

3.2.3.3 Աշխատավարձի վրա կրթական մակարդակի և ոլորտի համատեղ ազդեցության վերլուծություն

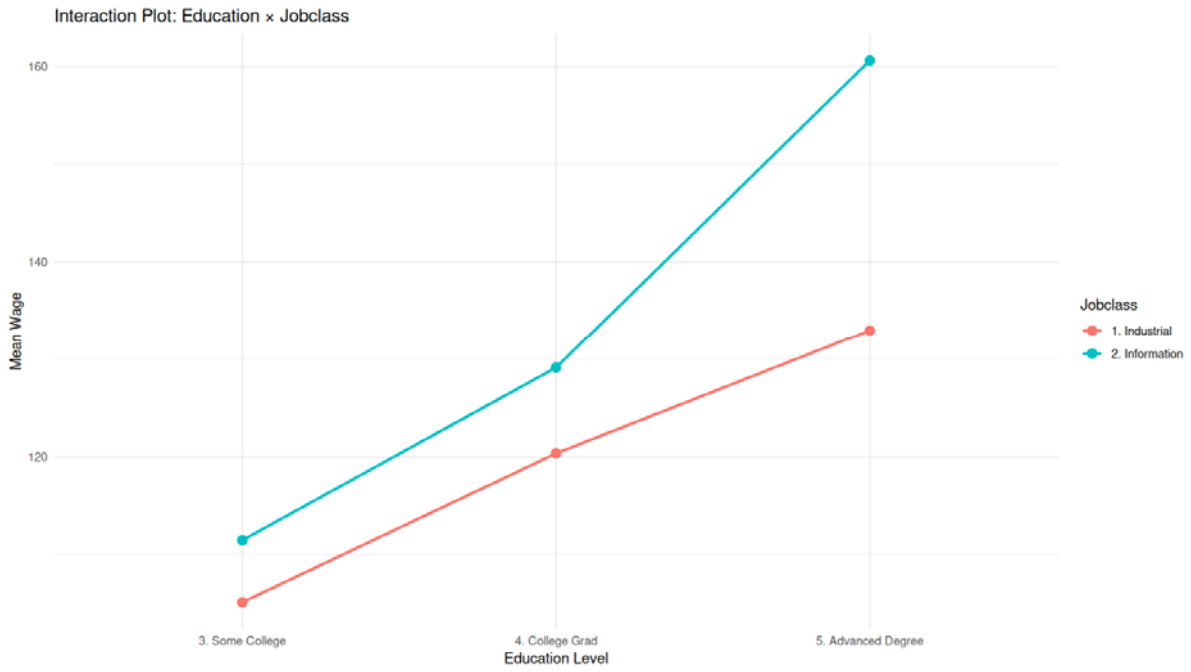
Համապարփակ հետազոտության համար վերջին փուլում կրկին կատարվում է երկգործոն վերլուծություն: ART ANOVA կիրառումը թույլ է տալիս ինտեգրել նախորդ փուլերում ստացված արդյունքները և մոդելավորել կրթության և ոլորտային պատկանելության համատեղ ազդեցությունը վարձատրության վրա:

Աղյուսակ 3.14. Երկգործոն վերլուծության արդյունքները

Գործոն	p-value	Ազդեցություն
Կրթության մակադրդակ	4.79×10^{-26}	Նշանակալի
Տնտեսության ոլորտ	2.93×10^{-10}	Նշանակալի
Փոխազդեցություն	0.056	Բացակայում է

Երկգործոն վերլուծության արդյունքները ցույց են տալիս (տես աղյուսակ 3,14), որ թե՛ կրթական մակարդակը, թե՛ տնտեսության ոլորտը հանդիսանում են աշխատավարձի վրա նշանակալի ազդեցություն ունեցող անկախ գործոններ: Փոխազդեցության էֆեկտը վիճակագրորեն նշանակալի չէ, ինչը նշանակում է, որ կրթության դրական ազդեցությունը եկամուտների վրա պահպանվում է նմանատիպ կերպով թե՛ արդյունաբերական, թե՛ տեղեկատվական ոլորտներում՝ չնայած տեղեկատվական ոլորտում նկատվող միջին աշխատավարձերի ընդհանուր բարձր մակարդակին:

Փոխազդեցության գրաֆիկը (տես նկ.9) տեսողականորեն հիմնավորում է, որ տեղեկատվական ոլորտում կրթական բարձր որակավորման աշխատավարձը ավելի զգալի է, քան արդյունաբերական ոլորտում: Մինչ արդյունաբերական ոլորտում կրթության ազդեցությունը աշխատավարձերի վրա համեմատաբար կայուն է, տեղեկատվական ոլորտում նկատվում է աճի ավելի սրընթաց կորագիծ, ինչը վկայում է գիտելիքահենք տնտեսության մեջ բարձր որակավորման մասնագետների նկատմամբ բարձր պահանջարկի մասին:



Նկ. 9. Աշխատավարձի վրա կրթական մակարդակի և տնտեսության ոլորտի փոխազդեցության գրաֆիկ

Կրթական մակարդակը հանդիսանում է աշխատավարձի ձևավորման հիմնական և կայուն գործոնը՝ անկախ ժամանակահատվածից և ոլորտից։ Սակայն տնտեսական ճգնաժամը բացահայտում է ոլորտային տարբերությունները. եթե նախաճգնաժամային շրջանում բարձրագույն գիտական աստիճանը հավասարեցնում էր աշխատավարձերը տարբեր ոլորտներում, ապա ճգնաժամային պայմաններում տեղեկատվական ոլորտը ցուցաբերում է ավելի բարձր դիմադրողականություն և ապահովում է զգալիորեն բարձր վարձատրություն նույնիսկ նույն որակավորում ունեցող մասնագետների համար։

Եզրակացություն

Դիպլոմային աշխատանքում իրականացվել է ոչ պարամետրական վիճակագրական մեթոդների տեսական և գործնական հետազոտություն՝ կիրառելով դրանք իրական տվյալների վերլուծության մեջ։

Ուսումնասիրվել է քանական փոփոխականի վրա որակական գործոնների ազդեցությունը բացահայտող վիճակագրական ոչ պարամետրական մեթոդների ընտանիքը, մասնավորապես Կրասկալ–Ուոլիսի, Մանն–Ուիթնիի տեստերը, զույգ-առ-զույգ համեմատության Դանիի տեստը, ինչպես նաև երկգործոն վերլուծության համար կիրառվող ժամանակակից մոտեցումներից մեկը՝ Aligned Rank Transform (ART) մեթոդը։

Օգտագործելով Python և R ծրագրավորման լեզուները, իրականացվել է երկու տարբեր տվյալների հավաքածուների վերաբերյալ վիճակագրական վերլուծություն քննարկված մեթոդների կիրառմամբ։ Հիմնավորվել է ոչ պարամետրական մեթոդների կիրառման անհրաժեշտությունը և հավաստիությունը դիտարկված տվյալների համար, մասնավորապես Շապիրո-Ուիլկի տեստի միջոցով հիմնավորվել է խմբային տվյալների նորմալ բաշխվածությանը չբավարարելու ենթադրությունը, ցույց է տրվել արտառոց արժեքների առկայությունը, ինչը դասական պարամետրական մեթոդների կիրառումը դարձնում էր անհնարին։

Դիպլոմային աշխատանքում մշակված վերլուծական ալգորիթմը և ստացված արդյունքները կարող են ծառայել որպես մեթոդական հիմք այլ ոլորտների նմանատիպ տվյալների վերլուծության և հիմնավորված որոշումների կայացման համար։

Օգտագործված գրականության ցանկ

1. S .M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*. 6th ed. London: Academic Press, 2021.
2. M. L. Berenson,et al. *Basic Business Statistics: Concepts and Applications*. 14th ed., Pearson, 2020.
3. Bruce P., Bruce A. *Practical Statistics for Data Scientists*: – Sebastopol: O'Reilly Media, 2017.
4. Gibbons J. D., Chakraborti S. *Nonparametric Statistics: Theory and Methods*: – Singapore: World Scientific Publishing Co. Pte. Ltd., 2018.
5. N. Nachar, The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution, Tutorials in Quantitative Methods for Psychology, vol. 4(1), p. 13-20, 2008. https://doi.org/10.20982/tqmp.04.1.p013
6. E. Ostertagová, O. Ostertag, and J. Kováč, Methodology and Application of the Kruskal-Wallis Test, Applied Mechanics and Materials, vol. 611, pp. 115–120, 2014, doi: 10.4028/www.scientific.net/amm.611.115.
7. A. Dinno, Nonparametric pairwise multiple comparisons in independent groups using Dunn's test, The Stata Journal, vol. 15, no. 1, pp. 292—300, 2015. https://doi.org/10.1177/1536867X1501500117
8. Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). Aligned Rank Transform for Nonparametric Factorial ANOVAs. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*, pp. 143–146:
9. Wes McKinney, Python for Data Analysis, https://wesmckinney.com/book/
10. D. H. Autor, , Katz, L. F., & Kearney, M. S. Trends in U.S. Wage Inequality: Revising the Revisionists. *The Review of Economics and Statistics*, 90(2), 300-323, 2008, https://doi.org/10.1162/rest.90.2.300
11. Beaudry, P., Green, D. A., & Sand, B. M. (2016). The Great Reversal in the Demand for Skill and Cognitive Tasks. *Journal of Labor Economics*, 34(S1), pp 199-247, 2016, https://doi.org/10.1086/682347
12. Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund, R for Data Science, https://r4ds.hadley.nz/