Kazalo vsebine

Kazalo tabel in slik

2. INFERENČNA STATISTIKA

2. 1. Uvod

V drugem delu statistične naloge želim natančneje analizirati prvo spremenljivko ENV2 na podlagi računalniško generiranega vzorca (Tabela 2. 1). V ta vzorec je vključnih deset slučajno izbranih držav članic EU, na podlagi katerih bom analizirala količino odpadkov v kg na posameznika v letu 2004.

Na podlagi te analize bom podala vsebinske ugotovitve o srednji vrednosti in razpršenosti spremenljivke na vzorcu v primerjavi s celotno populacijo. Na osnovi danega vzorca bom nato skušala oceniti populacijsko aritmetično sredino in stardardni odklon ter ugotoviti uspešnosti ocen (t.j. intervalov zaupanja) pri zajemanju pravih populacijskih vrednosti. Eden izmed ciljev drugega dela naloge je tudi preučiti razlike med aritmetičnimi sredinami spremenljivke med starimi in novimi članicami Evropske Unije na populaciji držav na osnovi razlik, dobljenih na vzorcu. Pod stare članice uvrščamo držav članice, pridružene pred 2004, med nove članice pa članice, pridružene leta 2004.

Tabela 2.1: Količina odpadkov v kg na posameznika v letu 2004 (ENV2). Podatki za slučajni vzorec desetih držav članic EU (n=10).

Država	ENV2
Belgija	469,0
Italija	559,5
Litva	355,0
Malta	537,7
Nizozemska	667,7
Poljska	256,0
Portugalska	416,6
Slovenija	456,8
Švedska	510,4
Velika Britanija	618,0

2. 2. Ocena populacijske aritmetične sredine in standardnega odklona

2. 2. 1 Vzorčna aritmetična sredina

Vzorčna aritmetična sredina $\overline{X_i}$ izračunana na i—tem vzorcu je ocena populacijske aritmetične sredine μ. Vzorčno aritmetično sredino imenujemo tudi cenilka populacijske aritmetične sredine μ. Vrednosti cenilke se od ocenjevanega parametra bolj ali manj odklanjajo (Ferligoj 1995: 112). Primerna je za številske, približno normalno porazdeljene spremenljivke.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10} \cdot 4846{,}700 = 484{,}670$$

2. 2. 2 Standardni odklon

Standardni odklon imenujemo tudi standardna napaka statistike (Ferligoj 1995: 112). Je napaka, ki jo napravimo, ko s statistiko ocenjujemo parameter (Brvar 1997: 158). Primeren je za številske, približno normalno porazdeljene spremenljivke.

Tabela 2. 2: Odkloni posameznih vrednosti spremenljivke ENV2 od aritmetične sredine in kvadrati teh odklonov ter varianca spremenljivke

Država	ENV2 $\left(x_i - \overline{X}\right)$	ENV2 $\left(x_i - \overline{X}\right)^2$
Belgija	-15,670	245,5489
Italija	74,830	5599,5289
Litva	-129,670	16814,3089
Malta	53,030	2812,1809
Nizozemska	183,030	33499,9809
Poljska	-228,670	52289,9689
Portugalska	-68,070	4633,5249
Slovenija	-27,870	776,7369
Švedska	25,730	662,0329
Velika Britanija	133,330	17776,8889
VARIANCA		135110,701

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^{10}(x_i - 484{,}670)^2}{10-1} = \frac{135110{,}701}{9} = 15012{,}300$$

$$s = \sqrt{15012{,}300} = 122{,}525$$

2. 2. 3 Interval zaupanja

Interval zaupanja je interval, v katerem se ocenjevani parameter (γ) nahaja z določeno verjetnostjo zaupanja $(1-\alpha)$ (Brvar 1997: 164). Z verjetnostjo tveganja α se parameter γ nahaja v tem intervalu. Verjetnost tveganja je pomembna, saj nas opozarja na dejstvo, da lahko trditev, ki jo postavljamo, ni pravilna, hkrati pa pove, kolikšen je delež vseh vrednosti, ki ne zagotavljajo pravilnega sklepa (Košmelj in Rovan 1997: 98).

$$P(a < \gamma < b) = 1-\alpha$$

Interval zaupanja za aritmetično sredino

Parameter, ki ga ocenjujemo je populacijska aritmetična sredina. Izračunati želimo spodnjo in zgornjo mejo intervala, v katerem se bo z 95% verjetnostjo zaupanja oz. 5% tveganjem nahajala populacijska aritmetična sredina $\mu = 520{,}504$.

Če tveganje α porazdelimo polovico na levo in polovico na desno na konce normalne porazdelitve, dobimo naslednjo formulo, kjer je $t_{\alpha/2}$ določen s stopnjo tveganja α.

$$P\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Iz tabele za t —porazdelitev preberemo, da je $t_{\alpha/2}(n-1) = t_{0,025}(9) = 2,26$

Spodnja meja intervala

$$a_a = \overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} = 484,670 - 2,26 \cdot \frac{122,525}{\sqrt{10}} = 397,104$$

Zgornja meja intervala

$$b_a = \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}} = 484,670 + 2,26 \cdot \frac{122,525}{\sqrt{10}} = 572,234$$

S 5% tveganjem lahko torej trdimo, da se aritmetična sredina populacije nahaja na intervalu s spodnjo mejo 397,104 in zgornjo mejo 572,234.

Interval zaupanja za standardni odklon

V tem primeru vzamemo za ocenjevani parameter standardni odklon populacije. Izračunati želimo interval, v katerem se bo z 95% gotovostjo oz. 5% tveganjem nahajal standardni odklon populacije $\sigma = 148,745$.

Najprej po spodnji formuli izračunamo meje intervala zaupanja za variacijo populacije, ju nato korenimo in dobimo meje intervala zaupanja za standardni odklon populacije.

$$P\left(\frac{(n-1)s^2}{x^2_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{x^2_{\alpha/2}}\right) = 1-\alpha$$

Spodnja meja intervala

Iz tabele za x^2 −porazdelitev razberemo, da je $x^2_{1-\alpha/2}(n-1) = x^2_{0,975}(9) = 19,02$

$$a_v = \frac{(n-1)s^2}{x^2_{1-\alpha/2}} = \frac{9 \cdot 122,525^2}{19,02} = 7103,612$$
$$a_s = \sqrt{7103,612} = 84,283$$

Zgornja meja intervala

Iz tabele za x^2 −porazdelitev razberemo, da je $x^2_{\alpha/2}(n-1) = x^2_{0,025}(9) = 2,70$

$$b_v = \frac{(n-1)s^2}{x^2_{\alpha/2}} = \frac{9 \cdot 122,525^2}{2,70} = 50041,252$$
$$b_s = \sqrt{50041,252} = 223,699$$

2. 3. Razlike med starimi in novimi članicami EU

Naključni vzorec razdelimo v dva podvzorca. V prvi vzorec uvrstimo države, ki so postale članice EU pred letom 2004, v drugega pa tiste ki so postale članice v letu 2004.

Tabela 2. 3: Količina odpadkov v kg na posameznika v letu 2004. Podatki za stare članice EU vključene v vzorec desetih držav ($n = 6$).

Država	ENV2
Belgija	469,000
Italija	559,500
Nizozemska	667,700
Portugalska	416,600
Švedska	510,400
Velika Britanija	618,000
SKUPAJ	3241,200

Tabela 2. 4: Količina odpadkov v kg na posameznika v letu 2004. Podatki za nove članice EU vključene v vzorec desetih držav ($n = 4$)

Država	ENV2
Litva	355,000
Malta	537,700
Poljska	256,000
Slovenija	456,800
SKUPAJ	1605,500

2. 3. 1 Aritmetična sredina in standardni odklon prvega vzorca

$$\overline{X_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{3241,200}{6} = 540,200$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \overline{X_1})^2}{n_1 - 1} = \frac{43916,020}{5} = 8783,204$$

$$s_1 = \sqrt{8783,204} = 93,719$$

2. 3. 2 Aritmetična sredina in standardni odklon drugega vzorca

$$\overline{X_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1605,500}{4} = 401,375$$

$$s_2^2 = \frac{\sum_{i=1}^{n_2}(x_i - \overline{X_2})^2}{n_2 - 1} = \frac{44940{,}968}{3} = 14980{,}323$$
$$s_2 = \sqrt{122{,}394} = 122{,}394$$

2. 3. 3 Preverjanje domneve o enakosti aritmetičnih sredin

Na osnovi prej dobljenih podatkov želimo preveriti domnevo o enakosti aritmetičnih sredin na populaciji starih in novih članic EU pri 5% stopnji značilnosti.

Domneva ali hipoteza je nedokazana, zgolj verjetna trditev o obstoju neke lastnosti populacije, ki je izražena kot nedokazano, verjetno razmerje med dvema spremenljivkama (Brvar 1997: 179) Postopek, ki ga uporabimo za preverjanje domnev imenujemo postopek preverjanja ali testiranja domnev. Pri tem postavimo osnovno in ničelno hipotezo. Osnovna hipoteza H_1 izraža trditev, ki ji domnevamo. Toda s statističnimi testi preverjamo ničelno hipotezo, ki zanika našo trditev. Ničelna hipoteza je namreč preprosta, razumljiva, nedvoumna in natančna, saj preprosti trdi, da povezanosti ni (Brvar 1997: 180).

Najprej postavimo ničelno in osnovno hipotezo

$$H_0 : \overline{X_1} = \overline{X_2} \quad oziroma \quad \overline{X_1} - \overline{X_2} = 0$$
$$H_1 : \overline{X_1} \neq \overline{X_2} \quad oziroma \quad \overline{X_1} - \overline{X_2} \neq 0$$

Pri preverjanju domneve predpostavimo enakost varianc v obeh vzorcih in dobimo skupni standardni odklon

$$s_s^2 = \frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{n_1 + n_2 - 2} = \frac{3\cdot 122{,}394^2 + 5\cdot 93{,}719^2}{8} = 11107{,}141$$
$$s_s = 105{,}390$$

ter eksperimentalno vrednost testne statistike t (matematično upanje je za oba vzorca enako $\mu_1 = \mu_2$)

$$t = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{s_s} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{540{,}200 - 401{,}357 - 0}{105{,}390} \cdot \sqrt{\frac{4 \cdot 6}{10}} = 2{,}041$$

Našo domnevo preizkušamo z dvosmernim testom, saj moramo pri našem preizkusu upoštevati oboje vrednosti, torej večje in manjše od parametra v ničelni domnevi (Košmelj in Rovan 1997: 235). Pri dvosmernem testu moramo vrednost značilnosti $\alpha = 0{,}05$ razpoloviti. Iz tabele razberemo vrednost testne statistike pri značilnosti $\alpha/2 = 0{,}025$

$t_{\alpha/2}(n_1 + n_2 - 2) = t_{0{,}025}(8) = 2{,}31$

Eksperimentalna vrednost testne statistike $t = 2{,}041$ ni enaka vrednosti testne statistike pri 5% stopnji značilnosti $t_{\alpha/2} = 2{,}31$, kar pomeni da smo potrdili ničelno in s tem ovrgli našo hipotezo o enakosti aritmetičnih sredin danih vzorcev.

2. 4 Vsebinske ugotovitve

Na podlagi računalniško generiranega slučajnega vzorca desetih držav EU sem analizirala prvo spremenljivko, količino odpadkov na posameznika v kg v letu 2004 (ENV2). Pri tem sem se osredotočila predvsem na aritmetično sredino in razpršenost spremenljivke na vzorcu v primerjavi s celotno populacijo. Izračunala sem vzorčno aritmetično sredino za dani vzorec, ki je znaša 484,670 kg. Upoštevati moramo, da se vrednost vzorčne aritmetične sredine ali cenilke bolj ali manj odklanja od prave vrednosti populacijske aritmetične sredine. Dobljena vzorčna aritmetična sredina tako od populacijske aritmetične sredine, ki znaša 520,504 kg, odstopa za 35,830 kg. Slučajni vzorec je torej dobro izbran, saj je odstopanje le 6,88%.

Vzorčni standardni odklon znaša 122,525 kar pomeni, da se vrednosti prve spremenljivke slučajnega vzorca odklanjajo v povprečju za 122,525 kg. Če vzročni standardni odklon primerjamo s populacijskim standardnim odklonom, ki znaša 148,745, tudi tu opazimo odstopanja. Vzorčni standardni odklon odstopa od populacijskega za 26,195 oziroma za 17,61 %. Vendar ta odstopanja vzorčne aritmetične sredine in standardnega odklona lahko zmanjšamo, s tem ko večamo vzorec.

Interval zaupanja je interval, v katerem se z določeno stopnjo gotovosti $1-\alpha$ oziroma tveganja α nahaja ocenjevani parameter. Ocenjevala sem populacijsko aritmetično sredino in standardni odklon populacije, zato sem izračunala meje za intervala zaupanja omenjenih parametrov, znotraj katerih se s 5% tveganjem nahajata parametra. Ugotovila sem, da lahko z 95% verjetnostjo oziroma 5% tveganjem trdimo, da se aritmetična sredina populacije $\mu=520{,}504$ nahaja na intervalu s spodnjo mejo 397,104 in zgornjo mejo 572,234. Interval zaupanja torej vsebuje pravo populacijsko vrednost aritmetične sredine, saj se ta nahaja znotraj meja dobljenega intervala. Z 95% verjetnostjo oziroma 5% tveganjem lahko trdimo tudi, da se standardni odklon populacije $\sigma=148{,}745$ nahaja na intervalu s spodnjo mejo 84,283 in zgornjo mejo 223,699. Tudi interval zaupanja za standardni odklon torej vsebuje pravo populacijsko vrednost standardnega odklona, saj se tudi ta nahaja znotraj meja intervala.

Na koncu me je zanimalo še kakšna je razlika aritmetičnih sredin spremenljivke med starimi in novimi članicami Evropske Unije. Države so razdeljene v dva vzorca. V prvi vzorec so vključene stare članice EU oziroma članice pridružene pred letom 2004, v drugi vzorec pa nove članice EU oziroma članice pridružene leta 2004. V prvem vzorcu se tako nahaja šest enot, v drugem pa le štiri. Za prvi vzorec je aritmetična sredina $\overline{X}_1=540{,}240$ in standardni odklon $s_1=93{,}719$. Aritmetična sredina drugega vzorca pa $\overline{X_2}=401{,}375$ in standardni odklon $s_2=122{,}394$. Aritmetična sredina vzorca starih članic je večja od aritmetične sredine drugega vzorca, vrednosti drugega vzorca pa se bolj odklanjajo od aritmetične sredine kot v prvem. Od populacijske odstopa aritmetična sredina vzorca starih članic za 119,129, medtem ko aritmetična sredina vzorca novih članic le za 19,736. V drugem vzorcu se namreč nahajajo države z precej nizkim BDP-jem, medtem ko so v drugem, z izjemo Portugalske, države z

visokim BDP-jem. Večja količina odpadkov je lahko torej posledica večjega BDP-ja, višje razvitosti držav in potrošniške družbe. Nato sem preverila še postavljeno domnevo o enakosti aritmetičnih sredin na populaciji starih in novih članic EU pri 5% stopnji značilnosti. Najprej sem postavila ničelno in osnovno hipotezo. Ničelna hipoteza nasprotuje naši osnovni domnevi o enakosti aritmetičnih sredin. Dobljena eksperimentalna vrednost testne statistike $t = 2{,}041$ ni enaka vrednosti testne statistike pri 5% stopnji značilnosti $t_{\alpha/2} = 2{,}31$, kar pomeni da smo potrdili ničelno in s tem ovrgli našo hipotezo o enakosti aritmetičnih sredin danih vzorcev.

2. 5 Literatura

Brvar, Bogo (1997) *Osnove statistike*. Ljubljana: Visoka policijsko-varnostna šola.

Ferligoj, Anuška (1995) *Osnove statistike na prosojnicah.* Ljubljana: Samozaložba.

Košmelj, Blaženka in Rovan Jože (1997) *Statistično sklepanje.* Ljubljana: Ekonomska fakulteta.

3. ANALIZA POVEZANOSTI NOMINALNIH SPREMENLJIVK

3. 1. Uvod

V tretjem delu naloge se ponovno osredotočimo na celotno populacijo držav članic EU. Preučevali bomo 25 držav članic EU, glede na količino odpadkov v kg na posameznika v letu 2004. Analizirali bomo dve dihotomni spremenljivki in na osnovi teh razporedili države v kontingenčno tabelo ter podatke ustrezno analizirali. Na podlagi analize dobljenih rezultatov bomo podali vsebinske ugotovitve o nekaterih razlikah med starimi in novimi članicami ter jih uprizorili s strukturnimi stolpci. S pomočjo različnih koeficientov pa bomo skušali tudi ugotoviti, če sta spremenljivki med seboj povezani.

Prva spremenljivka: Pripadnost starim ali novim članicam EU. Stare članice EU so države, ki so postale članice EU pred letom 2004, nove pa tiste, ki so postale članice v letu 2004.

Druga spremenljivka: Vrednost prve spremenljivke glede na mediano. Lahko zavzame vrednosti »pod mediano« in »enako« ali »nad mediano«.

Tabela 3. 1: Kontignenčna tabela: Stare in nove članice EU glede na mediano z dejanskimi empiričnimi frekvencami f_i

	Stare članice	Nove članice	Skupaj
manjše od mediane	5	7	12
enako ani večje od mediane	10	3	13
Skupaj	15	10	$N = 25$

3.2. Analiza povezanosti med dihotomnima spremenljivkama

Najprej moramo izračunati teoretične frekvence $f_i^{'}$, ki jih bomo primerjali z empiričnimi. Teoretična frekvenca je verjetnost določnega dogodka, pomnožena s številom enot v vzorcu (Ferligoj 1995: 163).

Stare članice označimo s S, nove članice z N, članice z vrednostjo manjšo od mediane z M in tiste z enako ali večjo vrednostjo z V.

Izračunamo teoretične frekvence, jih vnesemo v kontingenčno tabelo in jih ponazorimo s strukturnimi stolpci.

$$f^{'}(S \cap M) = n \cdot P(S \cap M) = 25 \cdot \frac{15}{25} \cdot \frac{12}{25} = 7{,}2$$

$$f^{'}(N \cap M) = n \cdot P(N \cap M) = 25 \cdot \frac{10}{25} \cdot \frac{12}{25} = 4{,}8$$

$$f^{'}(S \cap N) = n \cdot P(S \cap V) = 25 \cdot \frac{15}{25} \cdot \frac{13}{25} = 7{,}8$$

$$f^{'}(N \cap V) = n \cdot P(N \cap V) = 25 \cdot \frac{10}{25} \cdot \frac{13}{25} = 5{,}2$$

Tabela 3. 2: Kontignenčna tabela: Stare in nove članice EU glede na mediano z teoretičnimi frekvencami $f_i^{'}$

	Stare članice	Nove članice	Skupaj
manjše od mediane	7	5	12
enako ali večje od mediane	8	5	13
Skupaj	15	10	$N = 25$

3. 2. 1 Strukturni odstotki in strukturni stolpci

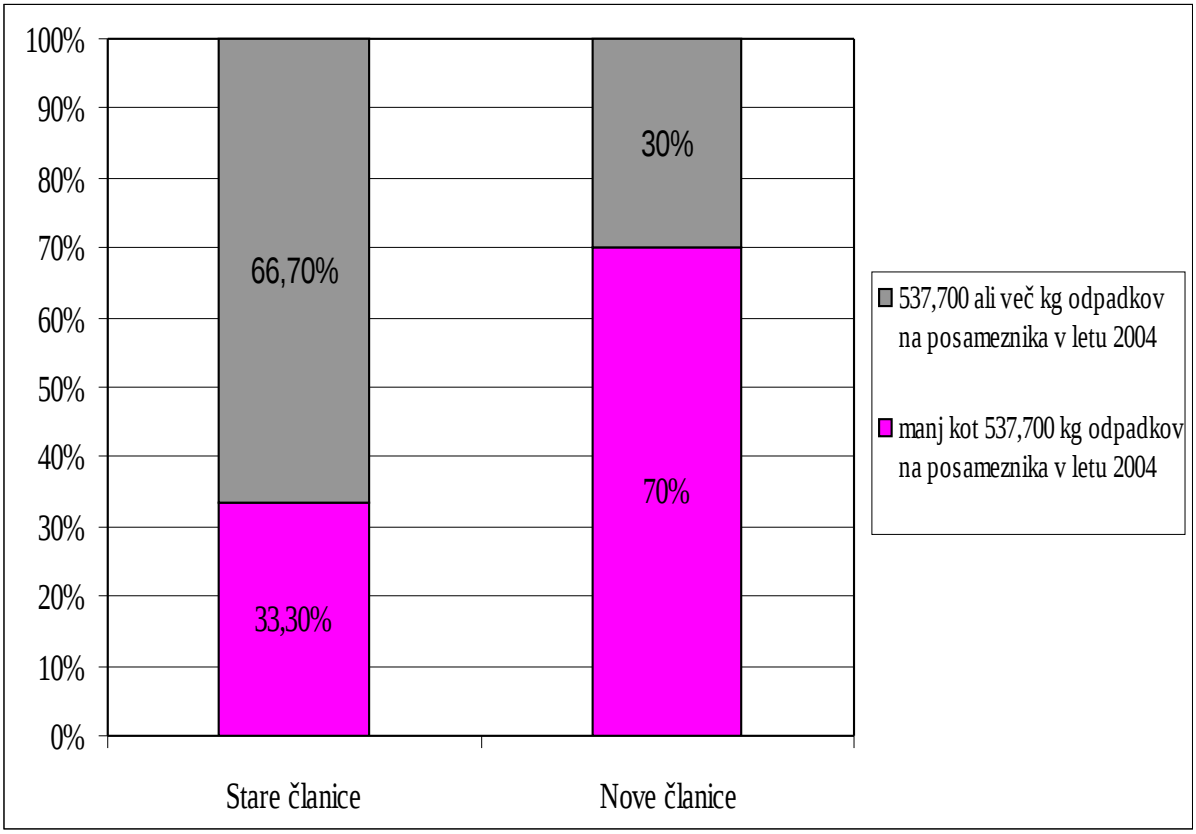
Strukturni odstotki so deleži, ki pripadajo posameznim vrednostim spremenljivke. To so relativne frekvence, ki jih izražamo v odstotkih $f_i\%$ (Brvar 1997: 40).

Tabela 3. 3: Kontingenčna tabela: Razporeditev starih in novih članih EU glede na količino odpadkov v kg na posameznika v letu

	Stare članice	Nove članice	
manjše od mediane	46,7 %	50 %	48 %
enake ali večje od mediane	53,3 %	50 %	52 %
	100 %	100 %	100 %

Strukturne odstotke grafično prikazujemo s strukturnimi stolpci. Strukturni stolpci so torej grafične predstavitve frekvenčne porazdelitve opisne spremenljivke (Ferligoj 1995: 24).

Slika 3. 1: Razporeditev starih in novih članih EU glede na količino odpadkov v kg na posameznika v letu



3. 2. 2 X^2, Cramerjev in kontingenčni koeficient

X^2 – test, Cramerjev in kontingenčni koeficient so primerni za ugotavljanje povezanosti med dvema nominalnima spremenljivkama. X^2 – test sicer kaže na povezanost med spremenljivkama, preizkus pa ne pove kolikšna je stopnja povezanosti med spremenljivkama. Prav tako pa dve X^2 – vrednosti nista primerljivi. To nam omogočajo različni kontingenčni koeficienti (Pearsonov, Cramerjev, popravljen Pearsonov). Njihova prednost je torej v izražanju stopnje in jakosti povezanosti med dvema spremenljivkama, kar omogoča primerjavo povezanosti med različnimi pari spremenljivk. Vendar pa kontingenčni koeficienti ne izražajo stopnje povezanosti. To lahko ugotovimo le s primerjavo dejanskih in teroretičnih frekvenc v kontingenčni tabeli (Brvar 1997: 208).

Najprej izračunamo X^2, ki ga bomo potrebovali dalje za izračun kontingenčnega in Cramerjevega koeficienta. X^2 namreč omogoča izračun vseh potrebnih parametrov za preverjanje hipotez (Brvar 1997: 199). X^2-test sloni na primerjavi dejanskih empiričnih frekvenc s teoretičnimi frekvencami, ki so v tem primeru frekvence, ki bi bile v kontingenčni tabeli, če spremenljivke ne bi bili povezani med seboj. Teoretične frekvence so torej tiste, ki jih pričakujemo glede na teorijo verjetnosti. Empirične pa so tiste, ki jih dejansko dobimo s preizkušanjem in pogosto odstopajo od pričakovanih. X^2 računamo po spodnji formuli, kjer je k število celic v kontingenčni tebeli.

$$X^2 = \sum_{i=1}^{k} \frac{(f_i - f_i^{'})^2}{f_i^{'}}$$

$$X^2 = \frac{(5-7)^2}{7} + \frac{(7-5)^2}{5} + \frac{(10-8)^2}{8} + \frac{(3-5)^2}{5} = 2{,}671$$

Statistika X^2 se porazdeljuje po X^2- porazdelitvi s $(s-1)(v-1)$ prostostnimi stopnjami, kjer je v število vrstic kontingenčne tabele in s število stolpcev.

Ničelna in osnovna domneva za ta test sta:

$H_0 : X^2 = 0$ (spremenljivki nista povezani)

$H_1 : X^2 > 0$ (spremenljivki sta povezani)

Iz tabele za porazdelitev X^2 lahko razberemo kritično vrednost statistike pri 5 % stopnji značilnosti:

$$X^2_{1-\alpha}[(s-1)(v-1)] = X^2_{0,95}(1) = 3{,}84$$

Sedaj lahko izračunamo Cramerjev koeficient, ki je mera povezanosti definirana na intervalu $[0,1]$. k predstavlja število vrstic, če je vrstic manj kot stolpcev, oziroma število stolpcev, če je stolpcev manj kot vrstic. Pred ostalimi koeficienti na isti osnovi ima to prednost, da je

uporaben za tabele poljubnih velikosti in da je njegova zgornja meja ni odvisna od velikosti tabel.

$$k = \min(v, s)$$
$$k = \min(2,2) = 2$$

$$\alpha = \sqrt{\frac{X^2}{n \cdot (k-1)}} = \sqrt{\frac{2,671}{25 \cdot 1}} = 0,327$$

Izračunamo še kontingenčni koeficient, ki je mera povezanosti, definirana na intervalu $[0, C_{\max}]$. Pri čemer je ponovno $k = \min(v, s)$.

$$C_{\max} = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{1}{2}} = 0,707$$

Kontingenčni koeficient je tako definiran na intervalu $[0,0'707]$.

$$C = \sqrt{\frac{X^2}{X^2 + n}} = \sqrt{\frac{2,671}{2,671 + 25}} = 0,311$$

Nato pa izračunamo tudi popravljen kontingenčni koeficient, ki je definiran na intervalu $[0,1]$.

$$C_{pop} = \frac{C}{C_{\max}} = \frac{0,327}{0,311} = 0,440$$

3. 3 Vsebinske ugotovitve

Ponovno sem obravnavala celotno populacijo držav EU, torej 25 držav članic. Analizirala sem dve dihotomni spremenljivki, pripadnost starim in novim članicam EU in vrednost prve spremenljivke glede na mediano količine odpadkov v kg na posameznika v letu 2004. Na osnovi teh dihotomnih spremenljivk sem države razporedila v kontigenčno tabelo. Starih članic je tako petnajst, novih pa deset.

Najprej sem v kontingenčno tabelo razporedila stare in nove članice EU glede na mediano z dejanskimi empiričnimi frekvencami. Vrednost mediane je $Me = 537,7$. Dvanajst držav članic ima količino odapadkov manjšo od mediane, od tega pet starih in sedem novih. Med ostalimi trinajstimi članicami, katerih vrednosti presegajo mediano, pa je deset starih in le tri nove. Iz tega lahko sklepamo, da med državami, ki pregajo mediano najdemo predvsem stare članice. Na podlagi delitve na stare in nove članice glede na mediano, pa je opazno, da ima kar 66,7 odstotkov starih držav članic 537,7 kg odpadkov ali več, medtem ko jih ima takšno količino odpadkov med novimi le 30 odstotkov. Med prevladujočim deležem starih članic najdemo večinoma gospodarsko močno razvite države, z visokim BDP-jem in izredno potrošniško družbo. Članic s količino odpadkov manjših od mediane je med starimi le 33,3 odstotkov, novih pa kar 70 odstotkov. Vse izmed teh novih so države z nizkim BDP-jem, ki so se gospodarsko počasneje razvijale. Nato sem izračunala še teoretične frekvence, ki jih pričakujemo glede na teorijo verjetnosti in jih zaokrožila na cela števila. Te kažejo, da je starih držav članic, s količino odpadkov manjših od 537,7 kg sedem, tistih, ki presegajo mediano pa osem. Med novimi pa je razmerje enakovredno, saj je članic z vrednostjo pod mediano in tistih z vrednostjo enako ali višjo od mediane obojih pet.

Z X^2 – testom sem nato preverila povezanost med dihotomima spremenljivkama in sicer s primerjavo teoretičnih frekvenc z dejanskimi. Eksperimentalna vrednost statistike X^2 je 2,671. Dobljena vrednost je manjša od kritične, ki jo razberemo iz tabele in znaša 3,84. Eksperimentalna vrednost je manjša od kritične, zato ne pade v kritično območje. To pomeni, da ničelne domneve ne moremo zavrniti in lahko rečemo, da pri 5% stopnji značilnosti

spremenljivki nista povezani. Kljub temu pa Cramerjev, kontingenčni koeficient in popravljen kontingečni koeficient kažejo na delno povezanost. Vrednost Cramerjevega koeficienta, definiranega na intervalu $[0,1]$ je 0,327, kar kaže na zmerno povezanost. Enako kaže vrednost kontingenčnega koeficienta, definiranega na intervalu $[0,0'707]$ ki je 0,311. Popravljen kontingenčni koeficient, definiran na intervalu $[0,1]$, znaša 0,440 in tudi kaže na nizko povezanost.

3. 4 Literatura

Brvar, Bogo (1997) *Osnove statistike*. Ljubljana: Visoka policijsko-varnostna šola.

Ferligoj, Anuška (1995) *Osnove statistike na prosojnicah.* Ljubljana: Samozaložba.

4. ANALIZA POVEZANOSTI RAZMERNOSTNIH SPREMENLJIVK

4.1. Uvod

V četrtem delu ponovno analiziramo obe spremenljivki (kot v prvem delu naloge) na vzorcu iz drugega dela naloge.

Najprej bomo skušali ugotoviti ali obstaja linearna povezanost med spremenljivkama na populaciji držav EU. Nato bomo analizirali linearno odvisnost prve (odvisne) spremenljivke od druge (neodvisne) spremenljivke (BDP na zaposlenega prebivalca) in to odvisnost tudi grafično ponazorili z razsevnim grafikonom. Skušali bomo ugotoviti kako in za koliko se spreminja odvisna spremenljivka, če se vrednost neodvisne spremenljivke spremeni za določeno enoto. Na koncu pa bomo analizirali kolikšen delež variabilnosti odvisne spremenljivke je mogoče pojasniti z vplivom neodvisne spremenljivke in kolikšna je variabilnost držav okoli regresijske premice.

Tabela 4. 1: Količina odpadkov v kg na posameznika v letu in BDP na zaposlenega prebivalca glede na kupno moč (v tisočih evrov) za vzorec desetih držav članic EU

Država	y_i	x_i
Belgija	469,000	67,000
Italija	559,500	61,400
Litva	355,000	25,900
Malta	537,700	42,500
Nizozemska	667,700	56,600
Poljska	256,000	29,000
Portugalska	416,600	37,100
Slovenija	456,800	37,900
Švedska	510,400	55,400
Velika Britanija	618,000	56,300
SKUPAJ	4846,700	469,100

4.2. Analiza povezanosti med spremenljivkama

Preveriti želimo domnevo o linearni povezanosti med spremenljivkama na populaciji držav EU pri 10% stopnji značilnosti. Pri tem si pomagamo s vzorčnim koeficientom korelacije.

$$r_{XY} = \frac{C_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{N} (x_i - \mu_X) \cdot (y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N} (x_i - \mu_X)^2 \cdot \sum_{i=1}^{N} (y_i - \mu_Y)^2}}$$

Pri čemer je C_{XY} kovarianca, ki meri linearno povezanost med spremenljivkama

$$C_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{X}) \cdot (y_i - \overline{Y})$$

Da bomo lahko preverili domnevo, moramo najprej postaviti ničelno in osnovno hipotezo. Kot smo že omenili ničelna hipoteza nasprotuje naši osnovni domnevi, ki jo želimo preveriti.

$$H_0 : r = 0 \quad spremenljivki \quad nista \quad linearno \quad povezani$$
$$H_1 : r \neq 0 \quad spremenljivki \quad sta \quad linearno \quad povezani$$

Sedaj s pomočjo tabele 4. 2 izračunamo vzorčni koeficient korelacije

Tabela 4. 2: Računanje vzorčnega koeficienta korelacije

Država	y_i	x_i	$y_i - \overline{Y}$	$x_i - \overline{X}$	$(y_i - \overline{Y})^2$	$(x_i - \overline{X})^2$	$(y_i - \overline{Y}) \cdot (x_i - \overline{X})$
Belgija	469,000	67,000	-15,670	20,090	245,549	403,608	-314,810
Italija	559,500	61,400	74,830	14,490	5599,529	209,960	1084,287
Litva	355,000	25,900	-129,670	-21,010	16814,309	441,420	2724,367
Malta	537,700	42,500	53,030	-4,410	2812,181	19,448	-233,862
Nizozemska	667,700	56,600	183,030	9,690	33499,981	93,896	1773,561
Poljska	256,000	29,000	-228,670	-17,910	52289,969	320,768	4095,480
Portugalska	416,600	37,100	-68,070	-9,810	4633,525	96,236	667,767
Slovenija	456,800	37,900	-27,870	-9,010	776,737	81,180	251,109
Švedska	510,400	55,400	25,730	8,490	662,033	72,080	218,448
Velika Britanija	618,000	56,300	133,330	9,390	17776,889	88,172	1251,969
SKUPAJ	4846,700	469,100	0,000	0,000	135110,701	1826,769	11518,313

$$\overline{X} = \frac{1}{n_X} \cdot \sum_{i=1}^{n_X} x_i = \frac{1}{10} \cdot 469{,}100 = 46{,}910$$

$$\overline{Y} = \frac{1}{n_Y} \cdot \sum_{i=1}^{n_Y} y_i = \frac{1}{10} \cdot 4846{,}700 = 484{,}670$$

$$r_{XY} = \frac{\sum_{i=1}^{N} (x_i - \overline{X}) \cdot (y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (x_i - \overline{X})^2 \cdot \sum_{i=1}^{N} (y_i - \overline{Y})^2}} = \frac{11518{,}313}{\sqrt{1826{,}769 \cdot 135110{,}701}} = 0{,}733$$

Koeficient korelacije lahko zavzame vrednosti v intervalu [-1,1]. Če je koeficient manjši od nič sta spremenljivki negativno linearno povezani, če je večji od nič sta pozitivno linearno povezani. Če pa je enak nič, spremenljivki nista linearno povezani. O popolni povezanosti govorimo ko je koeficient enak 1 ali -1 (Brvar 1997: 221-2).

Statistika $t = \frac{r_{XY} \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$ se porazdeljuje po t —porazdelitvi z $m = (n-2)$ prostostopnimi stopnjami. Ker gre za dvostranski test, sta kritični vrednosti enaki $\pm t_{\alpha/2}(n-2) = \pm t_{0,05}(8) = \pm 1{,}86$. Eksperimentalna vrednost statistike pa je

$$t = \frac{r_{XY} \cdot \sqrt{n-2}}{\sqrt{1-r_{XY}^2}} = \frac{0{,}733 \cdot \sqrt{8}}{\sqrt{1-0{,}733^2}} = 3{,}048$$

4.3. Analiza odvisnosti med spremenljivkama

Sedaj bomo analizirali linearno odvisnost prve (odvisne) spremenljivke od druge (neodvisne) spremenljivke s pomočjo regresijske premice $Y' = f(X)$. Ta kaže, kakšen bi bil vpliv spremenljivke X na spremenljivko Y, če razen vpliva spremenljivke X ne bi bilo drugih vplivov na spremenljivko Y (Ferligoj 1995: 186).

4. 3. 1 Regresijska premica in regresijski koeficient

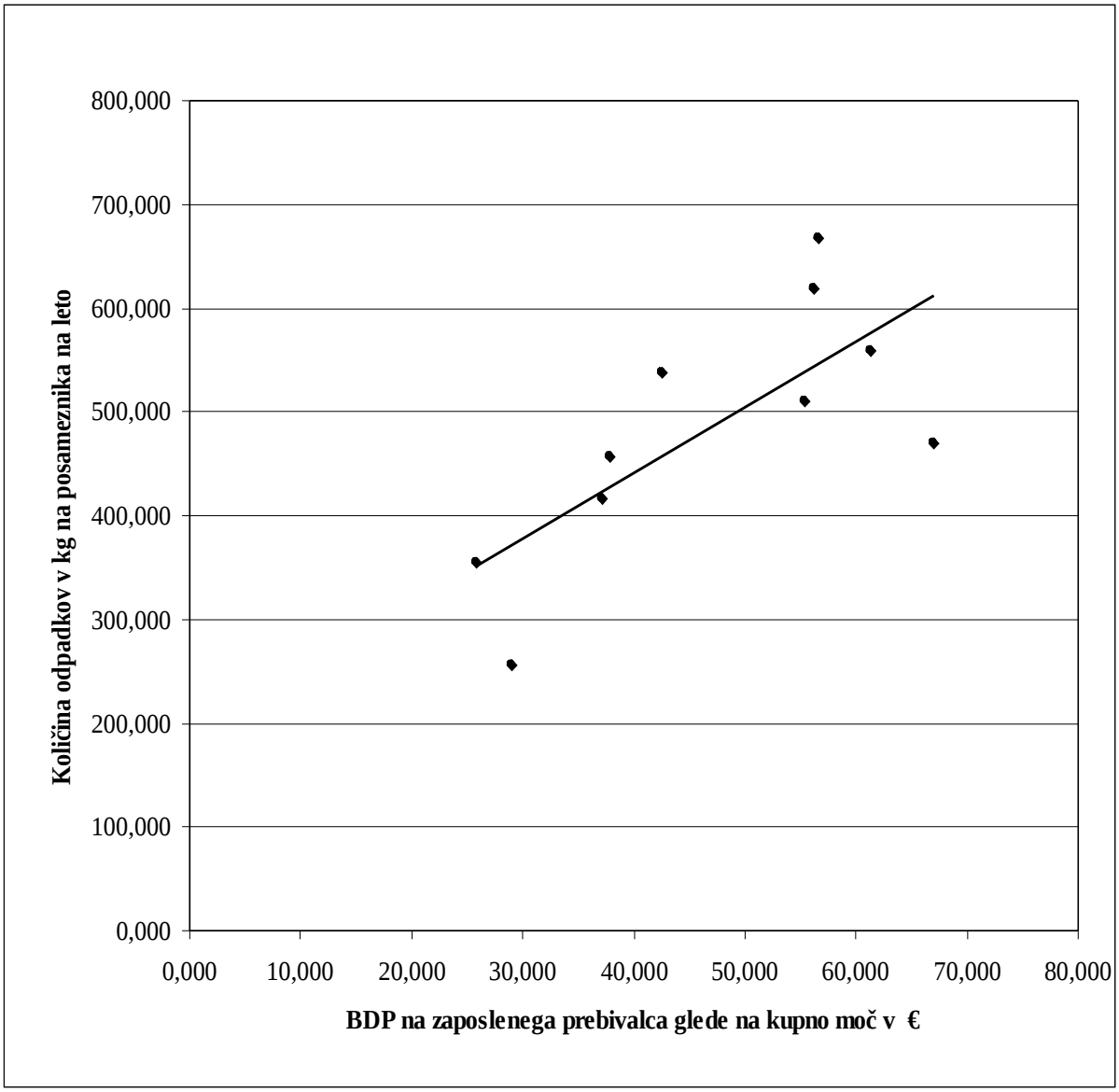
Regresijska premica je točkam najbolj prilegajoča se premica (Brvar 1997: 220).

$$Y' = \overline{Y} + \frac{\sum_{i=1}^{n}(x_i - \overline{X}) \cdot (y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \cdot (X - \overline{X}) = 484{,}670 + \frac{11518{,}313}{1826{,}769} \cdot (X - 46{,}910)$$

$$Y' = 188{,}889 + 6{,}305X$$

Premica je vrisana spodaj na Sliki 4. 1.

Slika 4. 1: Količina odpadkov v kg na posameznika v letu v odvisnosti od BDP na zaposlenega prebivalca glede na kupno moč



Razsevni grafikon je ponazoritev povezanosti med dvema številskima spremenljivkama. Enote s pari vrednosti, vrisane v koordinatni sistem, imajo za koordinati obe spremenljivki (Ferligoj 1995: 174).

Ocenjeni regresijski koeficient je enak 6,305. To pomeni, da če se BDP na zaposlenega prebivalca poveča za 1000 evrov, se količina odpadkov v kg v povprečju poveča za 6,305 kg.

4. 3. 2 Delež pojasnjene variance

Delež pojasnjene variance spremenljivke Y s spremenljivko X imenujemo tudi determinacijski koeficient in je definiran na intervalu $[0,1]$. Izračunamo ga tako, da

pojasnjeno vaianco spremenljivke Y delimo z celotno varianco spremenljivke Y. Determinacijski koeficient je pri linearni regresijski odvisnosti enak kvadratu koeficienta korelacije (Ferligoj 1995: 191).

$$R = \frac{\sigma_{Y'}^2}{\sigma_Y^2} = r_{XY}^2 = 0{,}733^2 = 0{,}537$$

4. 3. 3 Standardna napaka regresijske ocene

Standardna napaka regresijske ocene meri razpršenost okoli regresijske krivulje. Z njo merimo kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo (Ferligoj 1995: 191). Pokaže povprečno razliko med dejanskimi vrednostimi in z regresijo ocenjenimi vrednostmni. Pove nam za koliko smo se v povprečju zmotili, če uporabljamo z regresijo ocenjene vrednosti namesto dejanskih (Brvar 1997: 238). Je kvadratni koren iz nepojasnjene variance σ_e^2, v primeru linearne odvisnosti pa je standardna napaka enaka

$$\sigma_e = \sigma_Y \cdot \sqrt{1 - r_{XY}^2} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \overline{Y})^2} \cdot \sqrt{1 - R} = \sqrt{\frac{135110{,}701}{9}} \cdot \sqrt{1 - 0{,}537} = 83{,}371$$

4. 4 Vsebinske ugotovitve

V zadnjem delu sem analizirala obe spremenljivki na vzorcu desetih držav članic EU iz drugega dela naloge. Skušala sem namreč ugotoviti ali obstaja linearna povezanost med spremenljivkama na populaciji držav EU.

Preveriti sem želela domnevo o linearni povezanosti med spremenljivkama na populaciji držav EU pri 10% stopnji značilnosti. Linearno povezanost sem najprej preverila s Pearsonovim koeficientom linearne korelacije. Korelacijskegi koeficient, ki v mojem primeru znaša $r_{XY} = 0{,}733$, lahko zavzame vrednosti v intervalu $[-1,1]$. Sklepamo lahko torej, da sta spremenljivki visoko pozitivno povezani. Z večanjem BDP-ja se veča tudi količina odpadkov na posameznika, kar je lepo vidno iz regresijske premice $Y' = 188{,}889 + 6{,}305X$, vrisane v razsevnem grafikonu. Tudi ocenjeni regresijski koeficient, ki znaša 6,305 kaže na to, saj nam pove, da če se BDP na zaposlenega prebivalca poveča za 1000 evrov, se količina odpadkov v kg v povprečju poveča za 6,305 kg. Da bi lahko posplošila linearne povezanosti na populacijo, sem morala s testno statistiko *t* preveriti domnevo pri 10% stopnji značilnosti. Kritično območje je določeno s kritičnima vrednostima -1,86 in 1,86. Eksperimentalna vrednost statistike pa je 3,048 in tako pade v kritično območje. Zato lahko ničelno domnevo ovržemo in z 90% gotovostjo rečemo, da je količina odpadkov na posameznika linearno povezana z BDP-jem na zaposlenega prebivalca glede na kupno moč.

Nato sem s pomočjo determinacijskega koeficienta R skušala oceniti kolikšen del variance ene spremenljivke lahko pojasnimo z variiranjem druge. Determinacijski koeficient je enak 0,537. To pomeni, da je 53,7 % variance količine odpadkov na posameznika na leto pojasnjujemo z linearno odvisnostjo od BDP-ja na zaposlenega prebivalca glede na kupno moč.

Kot zadnje sem izmerila še kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo, tako da sem izračunala standardno napako regresijske ocene σ_e. Ocenila sem, da je standardna napaka ocene regresije, ki meri variiranje opazovanih vrednosti količine odpadkov na posameznika na leto okrog ocenjene regresijske funkcije enaka 83,371. Če bi torej uporabljali z regresijo dobljene vrednosti namesto dejanskih, bi se torej v povprečju zmotili za 83,371.

4. 5 Literatura

Brvar, Bogo (1997) *Osnove statistike*. Ljubljana: Visoka policijsko-varnostna šola.

Ferligoj, Anuška (1995) *Osnove statistike na prosojnicah.* Ljubljana: Samozaložba.