Statistinės duomenų analizės praktinės užduotys

1. **Duomenų rūšiavimas ir etiketės priskyrimas**

   Faile Prat1 KraujSp.sav pateiktos 120 įvairaus amžiaus ir lyties asmenų kraujo spaudimo reikšmės.

   (a) Priskirti etiketes kintamajam "Lytis":
   "1" - Vyras
   "0" - Moteris

   (b) Pirmame stulpelyje apibrėžti kintamąjį "Nr", nurodantį eilutės numerį lentelėje.

   (c) Surūšiuoti stebėjimus didėjančia tvarka kintamojo "Spaudimas" atžvilgiu .

   (d) Surūšiuoti stebėjimus didėjančia, po to mažėjančia tvarka kintamojo "Amžius", po to abiejų kintamųjų "Spaudimas" ir "Amžius" atžvilgiu.

   (e) Atstatyti pradinį failo pavidalą.

2. **Duomenų suklasifikavimas pagal kategorijas ir empirinių charakteristikų skaičiavimas kiekvienoje kategorijoje**

   Suklasifikuoti asmenis pagal amžių į septynias kategorijas, žymimas nuo 1 iki 7 didėjančia amžiaus tvarka. Pavadinkime kintamąjį, įgyjantį šias reikšmes, "katamž":

amžius	≤ 20	$21-30$	$31-40$	$41-50$	$51-60$	$61-70$	≥ 71
katamžius	1	2	3	4	5	6	7

   Rasti kintamojo "spaudimas" empirines charakteristikas

   (a) tarp visų asmenų;

   (b) kiekvienoje amžiaus kategorijoje;

   (c) vyrams ir moterims ;

   (d) vyrams ir moterims kiekvienoje amžiaus kategorijoje.

3. **Imčių generavimas.**

   (a) Sugeneruoti imtį $X_1, \ldots, X_{300}$, $X_i \sim N(65, 11)$ (kintamasis *norm1*).

   (b) Su tuo pačiu generatoriumi sugeneruoti imtį $Y_1, \ldots, Y_{300}$, $Y_i \sim N(65, 1)$ (kintamasis *norm2*) Palyginti empirines charakteristikas.

   (c) Sugeneruoti dar šešias tūrio $n = 300$ imtis (kintamieji *eksp; tolyg; ber; bin; puas, geom*) iš atitinkamai eksponentinio $E(65)$, tolygaus $U(54, 76)$, Bernulio $Bern(0.4)$, binominio $Bin(10, 0.4)$ bei Puasono $P(65)$ skirstinių.

   (d) Visų skirstinių atveju palyginkite teorinius ir empirinius vidurkius bei dispersijas.

4. Kritinių reikšmių ir kvantilių radimas.

   Rasti chi-kvadrato, Fišerio, Stjudento bei normalaus skirstinių kritines reikšmes

   $$\chi^2_{0,05}(8),\ \chi^2_{0.95}(8),\ F_{0,05}(100,5),\ F_{0.95}(100,5),\ t_{0,05}(26),\ t_{0.95}(26),\ z_{0,05},\ z_{0,95}.$$

5. **Pasikliovimo intervalai.**

   (a) Sukonstruoti 0.95 (0.90)-pasikliautinąjį intervalą *normalaus skirstinio* vidurkiui, naudojant imtį *norm1*.

   (b) Sukonstruoti 0.95 (0.90)-pasikliautinąjį intervalą *normalaus skirstinio* dispersijai, naudojant tą pačią imtį.

(c) Bandant sportinį lėktuvą, gautos šios jo maksimalaus greičio (m/s)reikšmės:

$$422,2;\ 418,7;425,6;\ 420,3;\ 425,8;\ 423,1;\ 431,5;\ 428,2;438,3;\ 434,0;$$

$$411,3;\ 417,2;\ 413,5;\ 441,3;\ 423,0.$$

Tarę, kad buvo stebimas *normalusis* a. d., raskite maksimalaus greičio vidurkio ir vidutinio kvadratinio nuokrypio taškinius įvertinius ir 0.95 pasikliovimo intervalus.

(d) sukonstruoti 0.95 -pasikliautinąjį intervalą eksponentinio skirstinio vidurkiui, naudojant imtį *eksp*;

(e) Laikas nuo užsakymo pateikimo iki jo gavimo (pristatymo laikas) yra pasiskirstęs pagal eksponentinį skirstinį. Lentelėje pateikiamos atsitiktinai parinktų užsakymų pristatymo laikas.

(i – eilės numeris, X_i – laikas).

i	X_i	i	X_i	i	X_i	i	X_i
1	10	6	7	11	10	16	7
2	10	7	11	12	6	17	6
3	6	8	12	13	13	18	16
4	11	9	12	14	8	19	9
5	8	10	6	15	12	20	5

Raskite vidutinio pristatymo laiko taškinį įvertinį ir $0,9$ pasikliovimo intervalą.

(f) sukonstruoti 0.95 -pasikliautinąjį intervalą Bernulio skirstinio parametrui p, naudojant imtį *ber*;

(g) Bandant kiekvieną iš 10 prietaisų, nebuvo rasta nė vieno defektinio. Raskite tikimybės, kad prietaisas yra defektinis, pasikliovimo intervalą, kai pasikliovimo lygmenys yra 0,8; 0,9; 0,99. Išspręskite tą patį uždavinį, tarę, kad buvo rasti trys defektingi gaminiai.

(h) sukonstruoti 0.95 -pasikliautinąjį intervalą Puasono skirstinio vidurkiui, naudojant imtį *puas*;

(i) Lentelėje pateikti skaičiai m_i tokių vienodo ploto (0,25 kv.km.) pietinės Londono dalies rajonų, į kuriuos Antrojo pasaulinio karo metu pataikė po i lėktuvų, sviedinių.

i	0	1	2	3	4	5	Σ
m_i	229	211	93	35	7	1	576

Tarę, kad buvo stebimas Puasono a. d., raskite parametro λ taškinį įvertinį ir $0,95$ pasikliovimo intervalą.

6. **Parametrinių hipotezių tikrinimas.**

(a) *Hipotezė apie normalaus skirstinio vidurkio reikšmę: Stjudento kriterijus vienai normaliajai imčiai* (ČM III.3.3.4).

Užkandžiais prekiaujanti firma nusprendė mėsainius su žuvimi pakeisti mėsainiais su bananais. Dvylikoje užkandinių per savaitę buvo parduota atitinkamai 530;540;510;500;520;

532;540;515;517;522;530 ir 510 naujųjų užkandžių. Anksčiau kiekviena užkandinė parduodavo vidutiniškai po 520 senųjų užkandžių per savaitę. Ar naujoji produkcija blogiau perkama? Reikšmingumo lygmuo 0,05.

(b) *Hipotezė apie proporcijos reikšmę* (ČM III.3.3.7). Naujo medikamento reklamoje teigiama, kad jis sukelia pašalines reakcijas ne daugiau kaip 1 procentui pacientų. Ištyrus 1000 vaistą vartojusių ligonių, nustatyta, kad pašalinį poveikį pajuto 32 ligoniai. Ar duomenys neprieštarauja reklaminiam teiginiui? Reikšmingumo lygmuo 0,05.

(c) *Hipotezė apie proporcijos reikšmę* (ČM III.3.10). Ekonomistas nori patikrinti, ar padaugėjo smulkių įmonių (procentais). Prieš dešimt metų jos sudarė 20 procentų visų įmonių. Šiuo metu iš 100 atsitiktinai parinktų įmonių 27 buvo smulkios. Reikšmingumo lygmuo 0,05.

(d) *Hipotezė apie koreliacijos koeficiento lygybę nuliui* (ČM III.3.13). Duomenys apie pardavėjo stažą (metais) ir jo pradinį atlyginimą (sutartiniais vienetais) pateikti lentelėje. Ar atlyginimas priklauso nuo pardavėjo stažo?

Stažas	Atlyginimas	Stažas	Atlyginimas
2	100	8	500
1,5	300	7	400
3	400	5	400
10	600	4	250
12	600	2	200
4	300	1	100
2	100	6	350

(e) *Dviejų priklausomų normaliųjų imčių vidurkių palyginimas: Stjudento kriterijus.*

Tiriamas fizinių pratimų poveikis svoriui. Parenkamos 5 moterys ir matuojamas jų svoris prieš ir po fizinių pratimų kurso. Gauti rezultatai:

$$84, 97, 77, 91, 85 \quad \text{(prieš)} \quad \text{ir} \quad 78, 95, 73, 88, 80 \quad \text{(po)}.$$

Rasti 0.9-pasikliautinąjį intervalą svorių skirtumui ir patikrinti hipotezę: vidutinis svoris nepakinta.

(f) *Dviejų priklausomų normaliųjų imčių vidurkių palyginimas: Stjudento kriterijus.*

Krakmolo kiekis bulvėse nustatomas dviem būdais. Norint palyginti tuos būdus, buvo paimta 16 bulvių ir kiekvienoje iš jų krakmolo kiekis nustatytas abiem būdais. Gauti rezultatai (krakmolingumas procentais) surašyti lentelėje.

Eil. Nr.	I būdas	II būdas	Eil. Nr.	I būdas	II būdas
1	21.7	21.5	9	14.0	13.9
2	18.7	18.7	10	17.2	17.0
3	18.3	18.3	11	21.7	21.4
4	17.5	17.4	12	18.6	18.6
5	18.5	18.3	13	17.9	18.0
6	15.6	15.4	14	17.7	17.6
7	17.0	16.7	15	18.3	18.5
8	16.6	16.9	16	15.6	15.5

Laikydami, kad nustatomas krakmolingumo procentas turi normalųjį skirstinį patikrinkite prielaidą, kad abu metodai yra ekvivalentūs.

(g) *Dviejų nepriklausomų normaliųjų imčių vidurkių palyginimas: Stjudento kriterijus.*

Matuojamas 16 detalių, pagamintų vieną dieną, atsparumas:

$$13.1; \quad 12.8; \quad 11.9; \quad 12.4; \quad 13.5; \quad 13.5; \quad 12.0; \quad 13.8;$$

$$10.6; \quad 12.4; \quad 13.5; \quad 11.7; \quad 13.9; \quad 11.5; \quad 12.5; \quad 11.9$$

Kitų 9 detalių, pagamintų kitą dieną, atsparumas:

$$13.7; \quad 13.5; \quad 14.2; \quad 15.6; \quad 14.8; \quad 14.3; \quad 15.4; \quad 14.0; \quad 15.1.$$

Patikrinti hipotezę, kad abi dienas buvo gaminamos vidutiniškai vienodo atsparumo detalės.

(h) *Dviejų nepriklausomų normaliųjų imčių vidurkių palyginimas: Stjudento kriterijus.*

Lentelėje pateikti dviejų nepriklausomų eksperimentų su musėmis rezultatai. Pirmajame bandyme tam tikrais nuodais musės buvo veikiamos 30 sekundžių, antrajame – 60 sekundžių. Paralyžiuojantį nuodų poveikį nusako vadinamasis reakcijos laikas, praėjęs nuo musės sąlyčio su nuodais iki to momento, kai musė jau nebegali stovėti. Reikia patikrinti hipotezę, kad vidutinis reakcijos laikas antrajame bandyme yra trumpesnis.

I bandymas				II bandymas			
i	X_i	i	X_i	i	Y_i	i	Y_i
1	4,9	9	17,1	1	10,8	9	30,6
2	16,2	10	17,9	2	10.9	10	36.3
3	25,4	11	26,6	3	13,3	11	26,9
4	8,6	12	33,7	4	13,40	12	22,4
5	10,9	13	33,9	5	17,1	13	51,9
6	12,5	14	28.1	6	19,2	14	23,8
7	12,9	15	15,9	7	25,0	15	26,9
8	9,8	16	66.2	8	26.0		

Nurodymas. Iš histogramų pavidalo matome, kad X ir Y skirstiniai labai asimetriški. Todėl reikėtų atlikti stebimų dydžių transformacijas, kad naujų a. d. skirstiniai būtų patenkinamai aprašomi normaliuoju skirstiniu, po to remtis Stjudento kriterijumi. Nesunku įsitikinti, kad nagrinėjamame pavyzdyje $\ln X$ ir $\ln Y$ tiksliau aprašomi normaliuoju skirstiniu, negu a. d. X ir Y. Kitaip sakant, stebimieji a. d. tiksliau aprašomi lognormaliuoju skirstiniu.

(i) *Dviejų nepriklausomų normaliųjų imčių dispersijų palyginimas: Fišerio kriterijus.*
Naudojant 6g pratimo duomenis patikrinti hipotezę, kad abi dienas pagamintų detalių atsparumai turi vienodas dispersijas.

(j) *Proporcijų palyginimas.*
Dviejose šalyse buvo pateiktas klausimas: "ar bijote vaikščioti gatvėje naktį?" Pirmoje šalyje iš $n_1 = 300$ apklaustųjų buvo gauta $S_1 = 120$ teigiamų atsakymų. Antrojoje šalyje gauta $S_2 = 148$ teigiamų atsakymų tarp $n_2 = 200$ apklaustųjų. Patikrinti hipotezę, kad abiejose šalyse žmonės vienodai baiminasi vaikščioti gatvėje naktį.

(k) *Proporcijų palyginimas.* (ČM III.4.9) Nepriklausomas ekspertas tiria, kiek kartų garantinio TV taisymo prireikė televizoriams, surinktiems Pietryčių Azijoje, ir kiek - Rytų Europoje. Iš 150 azijinių televizorių garantinio remonto prireikė 4, iš 100 europinių - 2. Ar galima teigti, kad europiniams televizoriams garantinio remonto reikia rečiau?

7. Dažnių lentelės

(a) *Hipotezės apie polinominio skirstinio parametrų reikšmes tikrinimas.* (ČM III.5.1) Atliktas tyrimas, kurio tikslas - nustatyti, kokios spalvos automobiliai populiariausi. Atsitiktinai apklausus 200 potencialių pirkėjų, gauti rezultatai, kurie pateikiami lentelėje.

Spalva	Raudona	Geltona	Mėlyna	Balia	Ruda
Dažnis	39	65	46	37	13

Patikrinkite hipotezę, kad pirkėjai nevienodai vertina automobilių spalvas.

(b) *Hipotezės apie polinominio skirstinio parametrų reikšmes tikrinimas.* (ČM III.5.2) Rinkos analitikas mano, kad A,B,C ir D rūšies dantų pastos vartotojų dalis yra atitinkamai 0,30; 0,60; 0,08 ir 0,02. Atsitiktinai apklausus 600 žmonių, kokią pastą jie vartoja, gauti rezultatai, kurie pateikiami lentelėje.

Rūšis	A	B	C	D
Dažnis	192	342	44	22

Ar šie duomenys leidžia suabejoti rinkos analitiko teiginiu?.

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

	0-1	1-2	2-3	3	Σ
0	2161	3577	2184	1635	9557
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	325	419	96	38	878
4	39	98	31	14	182
Σ	6216	10928	5173	3046	25362

(d) *Požymių nepriklausomumo tikrinimas.* (ČM III.5.4) Buvo tirta, ar užimamos pareigos ir pasitenkinimas darbu yra tarpusavyje susiję dalykai. Atsitiktinai apklausus 800 aukštųjų mokyklų dėstytojų, buvo gauti tokie rezultatai:

	Asistentas	Lektorius	Docentas	Profesorius
Patenkintas	40	60	52	63
Neturi nuomonės	78	87	82	88
Nepatenkintas	57	63	66	64

Patikrinkite hipotezę apie pareigų ir pasitenkinimo darbu priklausomybę.

(e) *Homogeniškumo hipotezės tikrinimas.* Viename sraute iš 300 stojančiųjų pažymius "nepatenkinamai", "patenkinamai", "gerai" ir "labai gerai" gavo atitinkamai 33, 43, 80 ir 144; kito srauto stojantieji atitinkamai 39, 35, 72 ir 154. Ar galima laikyti, kad abiejų srautų stojantieji pasiruošę vienodai?

Patikrinkite hipotezę apie pareigų ir pasitenkinimo darbu priklausomybę.

(f) *Homogeniškumo hipotezės tikrinimas.* (ČM III.5.6) Sveikatos apsaugos ministerija tyrė, ar įvairių profesijų žmonių alkoholio vartojimo įpročiai yra tokie pat. Atsitiktinai apklausus 200 mokytojų, 300 teisininkų ir 400 gydytojų, buvo gauti tokie rezultatai:

	Mokytojai	Teisininkai	Gydytojai
Mažai	100	50	100
Vidutiniškai	50	150	200
Daug	50	100	200

Ar galima teigti, kad šių trijų profesijų atstovų alkoholio vartojimo įpročiai tokie pat?

8. **Neparametriniai kriterijai**

(a) *Spirmeno koreliacijos koeficientas. Požymių nepriklausomumo tikrinimas* (ČM II.1.7). Kavos rūšių Atuvos rinkoje dalys (procentais) ir reklamos išlaidos pateiktos lentelėje.

Rūšis	A	B	C	D	E	F	Kita
Rinkos dalis (%)	15,7	3,9	10,6	9,6	12,3	26,2	21,7
Reklamos išlaidos (tūkst. eurų)	7,6	3,5	6,1	6,8	8,3	10,1	7,1

Apskaičiuokite Spirmeno koreliacijos koeficientą. Ar kintamieji yra priklausomi?

(b) *Dviejų nepriklausomų imčių palyginimas: Vilkoksono-Mano-Vitnio kriterijus* (ČM II.1.3). Besiruošdamas jubiliejui, verslininkas Apolonijus Drinkūnas atliko eksperimentą. Prisipirkęs pigių ir brangių šventinių žvakučių, jis fiksavo laiką, per kurį žvakutės sudega. Ar ponas Apolonijus teisus nutaręs, kad žvakučių degimo laikas nesiskiria ir jubiliejui tiks pigiosios žvakutės? Duomenys pateikti lentelėje.

Pigios	25	27	23	28	22	20	
Brangios	31	26	30	24	29	21	33

(c) *Dviejų priklausomų imčių palyginimas: Vilkoksono žymėtųjų rangų kriterijus* (ČM II.1.2). Norima patikrinti, ar tam tikri pratimai mažina sistolinį kraujo spaudimą. Atsitiktinai parinktų 15 žmonių spaudimas buvo matuotas prieš darant pratimus ir po to. Duomenys pateikti lentelėje.

Asmuo	Prieš	Po	Asmuo	Prieš	Po
1	164	162	9	138	139
2	146	144	10	130	124
3	148	146	11	175	170
4	154	156	12	146	147
5	143	145	13	160	157
6	160	159	14	160	140
7	150	145	15	153	148
8	148	150			

Ar pratimai efektyvūs?

(d) *Kelių nepriklausomų imčių palyginimas: Kruskalo-Voliso kriterijus.* Trijose gamyklose buvo testuojami kineskopai. Jų funkcionavimo trukmė (mėnesiais iki pirmo gedimo) surašyti pateikiamoje lentelėje.

1 gamyklos kineskopai	41	70	26	89	62	54	46	77	34	51		
2 gamyklos kineskopai	30	69	42	60	44	74	32	47	45	37	52	81
3 gamyklos kineskopai	23	35	29	38	21	53	31	25	36	50	61	

Ar galima tvirtinti, kad visose trijose gamyklose gaminamų kineskopų funkcionavimo trukmės turi vienodus skirstinius?

(e) *Kelių priklausomų imčių palyginimas: Frydmano kriterijus.* Penki nepriklausomi ekspertai vertino 3 rūšių (A,B ir C) alų. Duomenys pateikti lentelėje.

Ekspertas	A	B	C
Pirmas	10	7	8
Antras	5	2	4
Trečias	6	8	6
Ketvirtas	3	4	6
Penktas	9	8	10

Ar visų rūšių alus vienodai geras?

(f) *Kelių priklausomų imčių palyginimas: Frydmano kriterijus* (ČM II.1.5). Lentelėje pateikti duomenys apie 3 tiekėjų siūlomų 12 skirtingų tipų spausdintuvų kainas.

Tipas	1 tiekėjas	2 tiekėjas	3 tiekėjas	Tipas	1 tiekėjas	2 tiekėjas	3 tiekėjas
1	660	673	658	7	1980	1950	1970
2	790	799	785	8	2300	2295	2310
3	590	580	599	9	2500	2480	2490
4	950	945	960	10	2190	2190	2210
5	1290	1280	1295	11	5590	5500	5550
6	1550	1500	1499	12	6000	6100	6090

Ar tiekėjų siūlomos spausdintuvų kainos skiriasi?

(g) *Proporcijų palyginimas, kai imtys priklausomos: Maknemaros kriterijus.* Prieš rinkiminius debatus 1000 žmonių iš įvairių visuomenės sluoksnių atsakė į klausimą: "ar balsuosite už kandidatą N?" Po rinkiminių debatų tie patys žmonės atsakė į tą patį klausimą. Rezultatai pateikti lentelėje.

	Taip (po)	Ne (po)
Taip (prieš)	421	115
Ne (prieš)	78	386

Ar rinkiminiai debatai pakeitė rinkėjų nuomonę?

(h) *Serijų kriterijus* (ČM II.1.3 pvz.). Degalinėje yra A-95 (A) ir A-9ū (B) oktaninio skaičiaus benzino. Vieną dieną 50 automobilių benzino A ir B prisipylė tokia tvarka:

AABAABABBAAABBABBABBABBAABABBABBAABBBBAABABABAAABA

Ar galima teigti, kad benzino rūšies pasirinkimas yra atsitiktinis (t.y. paros metas neturi įtakos benzino rūšies pasirinkimui)?

9. **Dispersinė analizė**

(a) *Vienfaktorė dispersinė analizė: kelių normaliųjų vidurkių palyginimas.* Iš darbininkų, aptarnaujančių didelės įmonės surinkimo konvejerį, buvo atrinkti 4 darbininkai ir kiekvienam iš jų buvo užfiksuotas tam tikros detalės surinkimo laikas.

Darbininkas	Surinkimo laikas							
1	24,2	22,2	24,5	21,1	22,0			
2	19,4	21,1	16,2	21,2	21,6	17,8	19,6	
3	19,0	23,1	23,8	22,8				
4	19,9	15,7	15,2	19,8	18,9	16,1	16,2	18,5

Ar skiriasi darbininkai pagal detalės surinkimo laiką?

(b) *Vienfaktorė dispersinė analizė: kelių normaliųjų vidurkių palyginimas* (ČM II.2.2). Norėdamas rasti optimalų darbo režimą, firmos prezidentas trijuose vienodo našumo padaliniuose (kiekviename dirba po 6 darbuotojus) išbandė skirtingas 48 valandų darbo savaites: 5 dienų, 4 dienų ir 3,5 dienos. Eksperimentas vyko vienus metus. Lentelėje pateikta, kiek per vieną savaitę vidutiniškai kiekvienas darbuotojas pagamino detaliu.

5 dienos	4 dienos	3,5 dienos
360	340	371
302	350	365
333	306	303
351	337	300
329	371	336
340	365	365

Ar kuris nors savaitės darbo režimas davė statistiškai reikšmingai daugiau naudos nei likusieji?

(c) *Vienfaktorė dispersinė analizė: kontrastai* (ČM II.2.8). Kultūros savitumų tyrinėtojas skaičiavo, kiek gestų per dviejų valandų pranešimą padarė skirtingų šalių pranešėjia. Ar galima teigti, kad vidutinis kilikiečių gestikuliavimo lygis yra lygus nubų ir ilyrų vidutinių gestikuliavimo lygių vidurkiui?

Šalis	Gestai														
Nubija	41	37	40	36	28	38	44	27	28	39	41	39	42	35	44
Ilyrija	9	5	12	6	7	11	6	3	5	6	12	5	9	4	5
Kilikija	19	21	15	21	23	27	31	16	15	14	19	17	15	28	27

(d) *Dvifaktorė dispersinė analizė: hipotezių apie dviejų faktorių įtaką bei jų sąveiką tikrinimas* (ČM II. 3.1). Sąvoka "brangus" neretai asocijuojasi su sąvoka "geras". 54 tiriamieji buvo suskirstyti į tris grupes: valstiečius, verslininkus ir inteligentus. Kiekvienam tiriamajam buvo parodytas tas pats abstraktus paveikslas. Kartu buvo pasakoma viena iš trijų tariamų paveikslo kainų: maža, vidutinė arba didelė. Tiriamasis turėjo balais (iki 100) įvertinti paveikslą. Apklausos rezultatai pateikti lentelėje.

	Kaina																	
	Maža						Vidutinė						Didelė					
Valstiečiai	52	50	48	51	53	52	55	63	52	64	59	62	55	55	56	55	54	54
Verslininkai	47	45	45	44	46	43	65	56	66	57	69	60	78	72	76	75	74	77
Inteligentai	56	60	58	61	57	62	70	75	71	74	72	75	73	73	69	76	72	70

Ar skirtingų profesijų atstovai vienodai vertina abstraktųjį meną? Ar paveikslo kaina turi įtakos vertinimams? Ar kainos didėjimas vienodai paveikė visų profesijų atstovų nuomonę?

(e) (ČM II. 3.2). Chemijos laboratorija išbandė tris valiklius (I,II ir III). Kiekvienu iš jų buvo valomos keturių rūšių dėmės (purvo, riebalų, vyno, rašalo). Stebėta, per kiek sekundžių dėmė išnyksta. Bandymas buvo pakartotas keturis kartus. Duomenys pateikti lentelėje.

	Dėmės rūšis															
Val.	Purvo				Riebalų				Vyno				Rašalo			
I	12	10	9	11	13	13	14	14	18	18	19	17	25	26	17	25
II	12	11	10	12	14	15	14	15	18	17	17	16	20	21	22	22
III	8	9	9	10	16	13	15	14	17	18	18	19	24	23	23	25

Ką galima pasakyti apie valiklių efektyvumą?

(f) *Blokuotųjų duomenų vienfaktorė dispersinė analizė* (ČM II.4.2). Prekybos centro vadybininkas ieško tinkamos vietos indų plovikliui "Laumė". Ploviklis po savaitę buvo laikomas viršutinėje, vidurinėje ir apatinėje lentynose. Užfiksuota, kiek ploviklio indelių parduota. Eksperimente dalyvavo 7 parduotuvės, kuriose lentynų keitimo tvarka buvo randomizuota. Duomenys pateikti lentelėje.

	Lentyna		
Parduotuvė	Viršutinė	Vidurinė	Apatinė
Pirma	10	12	8
Antra	5	6	4
Trečia	16	18	15
Ketvirta	30	38	25
Penkta	19	19	17
Šešta	29	30	24
Septinta	39	41	36

Ar galima teigti, kad parduodamų ploviklių skaičius priklauso nuo to, kurioje lentynoje jie laikomi?

(g) *Blokuotųjų duomenų vienfaktorė dispersinė analizė* (ČM II.4.4). Eksperimento metu buvo skaičiuota, kiek važiavimo klaidų padaro kiekvienas ką tik teises gavęs vairuotojas, važiuodamas kiekvieną keturių skirtingų modelių mašina. Duomenys pateikti lentelėje.

	Mašinos modelis			
Vairuotojas	A	B	C	D
1	13	10	15	17
2	12	11	14	15
3	13	12	15	17
4	20	20	26	28
5	13	11	16	10
6	17	22	27	26
7	21	21	25	27
8	16	12	14	16
9	19	20	25	24
10	21	21	24	26

Ar visų modelių mašinomis važiuojama vienodai sėkmingai?

(h) *Blokuotųjų duomenų dvifaktorė dispersinė analizė* (ČM II.4.5). Sesijos metu studentai laiko keturis egzaminus (kalbos, politologijos, matematikos ir logikos). Duomenys pateikti lentelėje.

	Egzaminas			
Lytis	Kalbos	Politologijos	Matematikos	Logikos
Vyrai	9	10	9	7
	7	8	7	5
	5	8	7	6
	7	8	6	3
	7	9	6	5
	9	8	7	6
Moterys	6	5	5	2
	8	8	7	6
	9	10	8	8
	10	10	10	9
	10	9	8	6
	8	9	5	4

Ar visi egzaminai išlaikyti vienodai gerai? Ar studentai ir studentės egzaminus išlaikė vienodai gerai? Ar kuris nors egzaminas itin patiko studentėms?

(i) *Blokuotųjų duomenų dvifaktorė dispersinė analizė* (ČM 4.5 pvz.). Tris grupes karių apmokė skirtingi seržantai. Po apmokymo visų trijų grupių kariai šaudė į į taikinius naudodami indėniškąjį (I), mongoliškąjį (M) ir patobulintąjį (su optiniu taikikliu, P) lankus. Surinkti taškai pateikti lentelėje.

	Lankas		
Grupė	I	M	P
1	33	30	35
	32	31	34
	33	32	35
	32	30	36
	33	31	36
	27	32	37
2	31	31	35
	36	32	34
	29	30	35
	31	31	34
	31	30	35
	30	29	35
3	25	24	34
	24	23	33
	26	24	36
	25	25	35
	25	23	34
	26	24	35

Ar galima teigti, kad skirtingų grupių karių taiklumas skiriasi? Ar galima teigti, kad ne iš visų lankų vienodai gerai pataikoma? Ką dar galima pasakyti apie seržantų darbą?

10. **Tiesinė regresija**

(a) *Vieno kintamojo tiesinė regresija* (ČM II.5.2). Visuomeninis centras "Madam" kreipėsi į vyriausybę su prašymu visus valdininkus priversti lankyti jų centro organizuojamus džentelmeniškumo kursus. Kursų naudą pagrindė duomenimis, pateiktais lentelėje.

Treniruotės	254	230	254	300	320	364	312	264	274	226	274
Komplimentai	124	108	85	152	140	198	182	125	130	95	171
Treniruotės	234	274	306	234	252	340	364	324	368	286	318
Komplimentai	102	115	109	115	134	213	155	188	204	85	148
Treniruotės	216	350	216	358	222	374	222	230	360	336	
Komplimentai	106	155	73	179	118	159	79	74	180	126	

Joje užfiksuota, kiek kartų kiekvienas vyras kursuose treniravosi būti džentelmenu (x) ir kiek po to per mėnesį savo žmonai viešai pasakė komplimentų (skaičiavo uošvė). Ištirkite, ar tinka tiesinės regresijos modelis, ir padarykite prognozę, kiek komplimentų pasakys vyras, treniravęsis 267 kartus. Sudarykite 95% prognozės ir vidurkio pasikliovimo intervalus.

(b) *Vieno kintamojo tiesinė regresija* (ČM II.5.3). Paplūdimio gelbėtojų tarnyba visą mėnesį fiksavo vandens temperatūrą ir maksimalų besimaudančiųjų skaičių. Duomenys pateikti lentelėje.

Vandens temperatūra	17	18	16	18	16	18	14	15	19	12	12	14
Mauduolių skaičius	79	83	78	82	78	81	74	76	84	71	72	73
Vandens temperatūra	20	14	15	15	20	16	18	21	17	17	16	12
Mauduolių skaičius	85	76	76	77	86	79	82	88	81	82	79	120

Ištirkite, ar vandens temperatūra įtakoja mauduolių skaičių. Reikšmingumo lygmuo $\alpha = 0,05$. Ką galima pasakyti apie regresijos modelį, pašalinus išskirtį?

(c) *Kelių kintamųjų tiesinė regresija.* (ČM II.6.1). Policijos komisaras pastebėjo, kad po kiekvienos stambesnės privatizacijos kai kurie valdininkai gauna iš *nežinomos tetos* dovanų (palikimus). Duomenys apie dovanų vertę (tūkst. dolerių), privatizuotų objektų kainą (mln. eurų) ir konkurse dalyvavusių firmų skaičių pateikti lentelėje.

Vertė (x_1)	88	83	88	78	70	80	61	78
Firmų skaičius (x_2)	24	4	20	8	20	12	16	16
Dovana (Y)	106,5	74,5	93,5	80	85,5	91	80	85,5
Vertė (x_1)	87	82	87	77	69	79	60	77
Firmų skaičius (x_2)	28	4	17	9	17	12	16	20
Dovana (Y)	108,6	75,6	97,6	81,2	86,6	92,1	81,1	86,6

Kokią tetos dovaną prognozuotumėte valdininkui, jei objektas privatizuotas už 90 mln. eurų, o konkurse dalyvavo 10 firmų? Ar šiems duomenims tinka tiesinės regresijos modelis?

(d) *Kelių kintamųjų tiesinė regresija* (ČM II.6.2). Sporto apžvalgininkas tiria, kaip krepšinio komandos laimėtų rungtynių procentas Y priklauso nuo komandos biudžeto x_1 (mln.litų), vidutinio per rungtynes pelnyto taškų skaičiaus $x_@$ ir tritaškių pataikymo procento x_3. Duomenys apie 16 krepšinio komandų pateikti lentelėje.

Pergalės (Y)	91	33	73	43	53	63	43	53
Biudžetas (x_1)	6	7	6,1	8	9,6	7,6	11,4	8
Taškai (x_2)	85	74	83	77	78	81	80	79
Tritaškiai (x_3)	51,1	45,7	49,2	47	49	49	49	48,3
Pergalės (Y)	95	35	75	45	55	65	45	55
Biudžetas (x_1)	6,2	7,2	6,2	8,2	9,8	7,8	11,6	8,2
Taškai (x_2)	86	75	82	78	82	81	77	81
Tritaškiai (x_3)	51,5	46,1	49,5	47,5	49,3	49,3	49,2	49

Ar tiesinės regresijos modelis tinka? Ar visi kintamieji jame reikalingi?

11. **Logistinė regresija**

(a) (ČM II.7.1) Edukologas tiria žargono vartojimą sąlygojančius veiksnius. Po smagaus humoro vakaro, kurio metu visi leipo juokais, sociologas apklausė 22 žiūrovus. Kai kurie iš jų pasakė, kad prisijuokė, kiti pasidžiaugė, kad prisižvengė. Žiūrovo linksmumas (kintamasis Y=0 - žvengia, Y=1 - juokiasi) amžius ir lytis (0 - vyras, 1 - moteris) pateikta lentelėje. Remdamiesi logistine regresija, nustatykite, ar galima pagal amžių atskirti besijuokiančius žiūrovus nuo žvengiančiųjų. Lytį panaudoti kaip pseudokintamąjį.

Linksmumas (Y)	1	0	1	0	0	1	1	1	1	1	0
Lytis (x_1)	1	1	1	1	1	1	1	1	1	1	0
Amžius (x_2)	64	56	40	24	24	40	56	64	40	40	32
Linksmumas (Y)	1	1	0	0	0	0	1	0	0	1	0
Lytis (x_1)	0	0	0	0	0	0	0	0	0	0	0
Amžius (x_2)	64	56	40	24	24	40	56	64	40	40	32

(b) (ČM II.7.2) Ar galima pagal pajamas (PAJAMOS) ir darbo prestižiškumo indeksą (PREST) atpažinti, kad respondentas aukštąjį išsilavinimą turi (MOKSL=1) arba neturi (MOKSL=0)? Duomenys pateikti lentelėje.

PAJAMOS	3670	1923	3067	3811	3494	2012	1637	1265	2722
PREST	60	65	70	105	70	55	55	35	105
MOKSL	1	0	1	1	1	0	0	0	0
PAJAMOS	4050	1501	3340	3193	3125	4050	3458	2219	3781
PREST	135	50	65	60	95	115	65	65	90
MOKSL	1	0	1	1	0	1	0	0	1
PAJAMOS	2736	2568	3408	3298	3043	3536	3780	3798	
PREST	85	135	110	60	95	80	94	78	
MOKSL	0	0	0	1	1	1	1	1	

(c) (ČM II.7.3) Gimdymo namuose surinkti duomenys apie gimdyvių amžių, svorį (kg), rūkymą (1 - rūko, 0 - nerūko), hipertoniją (1 - serga, 0 - neserga) ir naujagimio svorį (g) (žr. lentelę).

Motinos amžius	Rūkymas	Hipertonija	Naujag. svoris (g)	Motinos svoris (kg)
24	0	0	1703	64,0
21	1	1	1792	82,5
21	0	0	1930	100,0
19	0	0	2084	51,0
24	0	0	2102	69,0
17	1	0	2227	55,0
18	0	0	2284	74,0
15	0	0	2383	57,5
17	0	0	2440	60,0
20	0	0	2452	52,5
14	1	0	2468	50,5
14	0	0	2497	50,0
21	1	1	2497	65,0
33	0	0	2553	77,5
32	0	0	2837	60,5
28	0	0	2879	83,5
29	0	0	2922	75,0
26	1	0	2922	84,0
17	0	0	2922	56,5
35	1	0	2950	60,5
33	1	0	3035	54,5
21	1	0	3044	92,5
19	0	0	3064	94,5
21	0	0	3064	80,0
19	0	0	3177	57,5
28	0	0	3236	70,0
16	1	0	3376	67,5
22	0	0	3462	65,5
32	0	0	3475	85,0
19	0	0	3574	52,5
24	0	0	3730	55,0
25	0	1	3985	60,0

Naujagimis sveria nepakankamai, jeigu jo svoris nesiekia 2500 g. Įvertinkite tikimybę, kad 38 metų būsimoji motina, kuri rūko, serga hipertonija ir sveria 85kg, pagimdys nepakankamo svoio naujagimį. Kokį naujagimio svorį prognozuotumėte taikydami tiesinę regresiją?

12. **Klasterinė analizė**

(a) (ČM II.8.4) Socialiniai bei ekonominiai 10-ties apskričių 1999 metų rodikliai pateikti lentelėje.

Apskritis	a_1	a_2	a_3	a_4	a_5	a_6	a_7
Alytaus	22,5	9,3	260433	122	24,5	171585	-1,3
Kauno	22,9	11,1	1052822	209	27,3	792675	-0,6
Klaipėdos	19,6	12,5	690000	247	28,7	850725	-0,2
Marijampolės	19,8	7,7	146914	164	24,1	20806	-0,6
Panevėžio	22,9	9,8	396002	210	23,9	334743	-1,8
Šiaulių	21,1	8,8	296607	189	24,3	147408	-1,1
Tauragės	21,6	7,0	35921	156	22,4	17674	-1,3
Telšių	21,1	10,0	503425	171	24,4	115108	-0,8
Utenos	24,5	10,3	268631	140	23,5	91212	-3,9
Vilniaus	20,6	15,8	2502666	236	26,9	3959258	-1,0

Čia a_1 - gyventojų aprūpinimas gyvenamuoju plotu (kiek vienam gyventojui vidutiniškai tenka naudingo ploto (m^2); a_2 - bendrasis vidaus produktas (BVP) vienam gyventojui (tūkst.litų); a_3 - materialinės investicijos (tūkst. litų); a_4- nusikalstamumas (kiek užregistruota nusikaltimų, tenkančių 10000 gyventojų); a_5 - gyventojų aprūpinimas telefonais butuose (100 gyventojų); a_6 - tiesiogins užsienio investyicijos (tūkst. Litų, sausio 1 d. Duomenys); a_7 - natūralus gyventojų prieauglis (1000 gyventojų). Kokius apskričių klasterius galima sudaryti pagal šiuos rodiklius? Vordo ir pilnosios jungties metodais suklasterizuokite apskritis pagal požymius a_1, a_2, a_4, a_5 ir a_6. Ar rezultatai skiriasi? Standartizavę požymius, k-vidurkių metodu suskirstykite apskritis į tris klasterius. Ar rezultatai skiriasi nuo gautų kitais metodais.

(b) (ČM II.8.5) Turime 4 alaus bendrovių akcijų kainų koreliacijos matricą (žr. Lentelę).

	(1)	(2)	(3)	(4)
(1)	1	0,51	0,39	-0,24
(2)	0,51	1	0,72	0,10
(3)	0,39	0,72	1	-0,44
(4)	-0,24	0,10	-0,44	1

Suskirstykite bendroves į du klasterius.

(c) (ČM II.8.6) Kazanova visas damas klasterizuoja pagal krūtinės ir klubų matmenis.

Dama	Krūtinė	Klubai
Rima	100	106
Irma	96	106
Mania	99	108
Ninel	98	104
Odeta	97	103
Pamela	94	111
Sigurn	100	102

Vienetinės jungties metodu pakartokite Kazanovos klasterizavimą.

Nurodymai

1. **Duomenų rūšiavimas ir etiketės priskyrimas**

   (a) SPSS data editor (duomenų editoriuje) apačioje kairėje pasirinkti "Variable view" ir spragtelėjus atidaryti kitą langą, kuriame įvedami duomenys apie kintamąjį (jo pavadinimas, tipas ir t.t.). Stulpelyje "Label" įveskite:
   1 - Vyras, 0 - Moteris

   (b) Grįžtame į langą "Data view". Kursorių ant pirmo kintamojo.
   Data-Insert variable. Vėl į langą "Variable view" ir Name: nr, Decimals: 0
   1 į pirmą langelį ir nukopijuoti į kitus langelius.
   Trnsform-Create time series
   Function: cumulative sums
   nr į langą-OK
   Su Copy-paste pirmą stulpelį pakeisti gautu ir vėl duoti pavadinimą nr

   (c) Data-sort cases
   Pasirenkame ascending
   "spaudim" į langą-ok

   (d) Analogiškai, tik į langą "Amžių" arba "Spaudimą" ir "Amžių" .

   (e) Analogiškai su nr.

2. **Duomenų suklasifikavimas pagal kategorijas ir empirinių charakteristikų skaičiavimas kiekvienoje kategorijoje**

   Transform-recode-into different variables

   Input variable-amžius, output variable-katamžius

   Change

   Old and new variables:

   Old: range: lowest through 20, new 1

   old: 21 through 30 new 2 ir t.t.

   old: 61 through 70 new 6

   old: 71 through highest 7

   Continue ok

   (a) *Analyze-Descriptive statistics-Frequences*
   Spaudim į langą
   Statistics: mean, median, std.deviation, variance, minimum, mazimum, s.e.mean
   Charts: histogram

   (b) *Data-Split file-Organize output by groups*
   Katamž į langą
   Dabar jau duomenys paruošti apdorojimui kiekvienoje amžiaus grupėje. Toliau darom kaip a)

   (c) Kaip b), Katamž pakeičiant į lytį

   (d) Kaip b), prie Katamž dar pridedant lytį

3. **Imčių generavimas.**

(a) *Transform-Random Numbers Generator*

Set starting point: pažymėkite "Fixed value" ir "Value" 2000000

OK

1 į 300 langelį (taip nurodome kokio dydžio imtį generuoti).

Transform-Compute

Lange "Function group" pasirinkite "Random numbers"ir lange "Functions and special variables" pasirinkite RV.NORMAL(mean,stddev), kurį permeskite į langą "Numeric expressio". Vietoje mean įveskite 65, vietoje stddev įveskite 11.

Target variable: norm1 ir OK.

(b) Tas pats, tik stddev=1, Target variable: norm2

Empirines charakteristikas skaičiuoti kaip 2a

(c) Tas pats, naudojam RV.EXP(1/65), RV.UNIFORM(54,76), RV.BERNOULLI(0.4), RV.BINOM(10,0.4), RV.POISSON(65).

4. **Kritinių reikšmių ir kvantilių radimas.**

Tarkime, kad atsitiktinio dydžio pasiskirstymo funkcija $F(x)$ turi atvirkštinę jo reikšmių srityje. Tada a.d. α- kvantiliu vadinamas skaičius $x(\alpha) = F^{-1}(\alpha)$, o α-kritine reikšme skaičius $x_\alpha = F^{-1}(1-\alpha)$, čia F^{-1} yra atvirkštinė funkcija. Pastebėkime, kad $F(x(\alpha)) = \mathbf{P}\{X \le x(\alpha)\} = \alpha$, $\bar{F}(x_\alpha) = \mathbf{P}\{X > x_\alpha\} = \alpha$.

Skaičiavimas.

Transform-Compute variable

Iš "Function group" pasirinkti "Inverse DF" ir įmesti į "Numeric expresion" langą atitinkamo skirstinio atvirkštinės pasiskirstymo funkcijos pažymėjimą. Skaičiuojant α-kvantilius pirmojo argumento vietoje įvedama α, o skaičiuojant α kritinę reikšmę įvedama $1-\alpha$. Taigi skaičiuojant kvantilius, naudojame

Funkcijos IDF.CHISQ(0.05,8), IDF.CHISQ(0.95,8), IDF.F(0.05,100,5), IDF.F (0.95,100,5),

IDF.T(0.05,26), IDF.T(0.95,26), IDF.NORMAL(0.05,0,1), IDF.NORMAL(0.95,0,1).

Skaičiuojant kritines reikšmes $\chi^2_{0,05}(8)$, $\chi^2_{0.95}(8)$, $F_{0,05}(100,5)$; $F_{0.95}(100,5)$, $t_{0,05}(26)$, $t_{0.95}(26)$, $z_{0,05}$, $z_{0,95}$, naudojame

Funkcijos IDF.CHISQ(0.95,8), IDF.CHISQ(0.05,8), IDF.F(0.95,100,5), IDF.F (0.05,100,5),

IDF.T(0.95,26), IDF.T(0.05,26), IDF.NORMAL(0.95,0,1), IDF.NORMAL(0.05,0,1).

5. **Pasikliovimo intervalai.**

a) *Normalaus skirstinio vidurkio pasikliovimo intervalas.*

Formulės.

Tarkime, kad $X_1, ..., X_n$ yra imtis, $X_i \sim N(\mu, \sigma^2)$. Vidurkio lygio $1-\alpha$ pasikliovimo intervalas

$$(\underline{\mu},\ \overline{\mu}) = (\bar{X} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}});$$

čia $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ir $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

Skaičiavimas.

Analyze-Compare means-One sample T test

norm1 į langą ok

Jei norime pakeisti pasikliovimo lygmenį į 0.9, tai prieš ok dar

Options-90%

Rezultatai.

Lentelėje "One sample statistics" pateiktos imties empirinės charakteristikos (imties dydis, vidurkis, vidutinis kvadratinis nuokrypis). Pagrindinė yra lentelė "One sample test". Ten ir duotas pasikliovimo intervalas $(\underline{\mu},\ \overline{\mu}) = (lower, upper)$.

b) *Normalaus skirstinio dispersijos pasikliovimo intervalas.*

Formulės.

$1-\alpha$ pasikliovimo intervalas normalaus skirstinio dispersijai yra $(\underline{\sigma^2}, \overline{\sigma^2})$;čia

$$\underline{\sigma^2} = \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \quad \overline{\sigma^2} = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}.$$

Skaičiavimas.

Analyze-Descriptive statistics-Descriptives

norm1 į langą ok

Suskaičiuojama $S = 11,0180$ ($S = std.deviation$).

Duomenų lange įvedame kintamąjį S, kurio pirmąją reikšmę apibrėžiame $S = 11,0180$ ir suskaičiuojame (žr. 4 uždav.) $chi1 = \chi^2_{0,025}(299) = 348,79$ bei $chi2 = \chi^2_{0,975}(299) = 252,99$.

Transform-compute

299*S**2/chi1 ir 299*S**2/chi2

c) kaip a) ir b), tik naudojami suvesti duomenys.

d)-e) *Eksponentinio skirstinio vidurkio pasikliovimo intervalas.*

Formulės.

$1-\alpha$ pasikliovimo intervalas eksponentinio skirstinio vidurkiui yra $\underline{\theta}, \overline{\theta}$, kur

$$\underline{\theta} = \frac{2n\bar{X}}{\chi^2_{\alpha/2}(2n)}, \quad \overline{\theta} = \frac{2n\bar{X}}{\chi^2_{1-\alpha/2}(2n)},$$

taigi skaičiavimas analogiškas normalaus skirstinio dispersijos pasikliautinojo intervalo skaičiavimui.

f)-g) *Bernulio skirstinio vidurkio (proporcijos) pasikliovimo intervalas.*

Formulės.

$1-\alpha$ pasikliovimo intervalas Bernulio skirstinio parametrui p yra $(\underline{p}, \overline{p})$, kur

$$\underline{p} = 1 - \beta_{\alpha/2}(n-T+1, T), \ (T = 1, \dots, n), \quad \overline{p} = 1 - \beta_{1-\alpha/2}(n-T, T+1), \ (T = 0, \dots, n-1);$$

čia $T = n\bar{X} = \sum_{i=1}^{n} X_i$, $\beta_\alpha(a,b)$ yra beta skirstinio kritinės reikšmės.

Jei $T = 0$, tai $\underline{p} = 0$. Jei $T = n$, tai $\overline{p} = 1$.

Taigi tereikia suskaičiuoti beta skirstinio kritines reikšmes.

Skaičiavimas: suma T surandama per *Analyze - Descriptive statistics -Frequences*, prie "Statistics" pažymėjus "sum".

Transform-Compute variable

Iš "Function group" pasirinkti "CDF and Non-central CDF" ir įmesti į "Numeric expresion" langą beta skirstinio pasiskirstymo funkcijos pažymėjimą "Cdf Beta". Lieka į pirmo argumento vietą įrašyti $\alpha/2$ arba $1-\alpha/2$, o antro ir trečio argumentų vietose atitinkamus laisvės laipsnius:

$$p_{apat} = 1 - IDF.BETA(\alpha/2, n-T+1, T), \quad p_{virs} = 1 - IDF.BETA(1-\alpha/2, n-T, T+1).$$

h)-i) *Puasono skirstinio vidurkio pasikliovimo intervalas.*

Formulės.

$1-\alpha$ pasikliovimo intervalas Puasono skirstinio parametrui λ yra $(\underline{\lambda}, \overline{\lambda})$; čia

$$\underline{\lambda} = \frac{1}{2n}\chi^2_{1-\alpha/2}(2T), \ (T = 1, 2, \dots), \quad \overline{\lambda} = \frac{1}{2n}\chi^2_{\alpha/2}(2T+2),$$

o $T = n\bar{X}$. Jei $T = 0$, tai $\underline{\lambda} = 0$.

Skaičiavimas: suma T surandam per *Analyze - Descriptive statistics -Frequences*, prie "Statistics" pažymėjus "sum". O toliau jau tereikia surasti chi kvadrato skirstinio kritines reikšmes (žr. 4 uždavinį).

6. **Parametrinių hipotezių tikrinimas.**

Sąvokos. Tarkime, kad $\mathbf{X} = (X_1, ..., X_n)^T$ yra paprastoji imtis, o X_i pasiskirstymo funkcija priklauso šeimai $\{F_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbf{R}^m\}$.

Teiginys $H : \boldsymbol{\theta} \in \boldsymbol{\Theta}_H$, $\boldsymbol{\Theta}_H \subset \boldsymbol{\Theta}$, vadinamas *parametrine hipoteze.* Hipotezė $\bar{H} : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{\bar{H}} = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_H$ vadinama *alternatyva* (hipotezei H).

Tarkime, kad

$$X_i \sim N(\mu, \sigma^2),\ \boldsymbol{\theta} = (\mu, \sigma^2)^T \in \boldsymbol{\Theta} = \mathbf{R} \times (0, \infty) \subset \mathbf{R}^2.$$

Hipotezių pavyzdžiai: $H_1 : \mu = 3, \sigma^2 = 1$, kuri užrašoma $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_1} = \{(3, 1)\} \subset \boldsymbol{\Theta}$; $H_2 : \mu = 3$, kuri užrašoma $H_2 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_2} = \{\boldsymbol{\theta} = (\mu, \sigma^2)^T : \boldsymbol{\theta} \in \{3\} \times (0, \infty)\} \subset \boldsymbol{\Theta}$; $H_3 : \mu \leq 3$, kuri užrašoma $H_3 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{H_3} = \{\boldsymbol{\theta} = (\mu, \sigma^2)^T : \boldsymbol{\theta} \in (-\infty, 3] \times (0, \infty)\} \subset \boldsymbol{\Theta}$.

Turint imties realizaciją, atsižvelgiant į jos reikšmę, priimamas vienas iš dviejų sprendimų: d_H – *hipotezė H teisinga* ir $d_{\bar{H}}$ – *hipotezė H klaidinga.* Hipotezės priėmimo ar atmetimo taisyklė vadinama *statistiniu kriterijumi.*

Jei hipotezė atmetama, kai imties realizacija $\boldsymbol{x} = (x_1, ..., x_n)^T$ patenka į sritį $\boldsymbol{K} \subset \mathbf{R}^n$, tai sritis $\boldsymbol{K}$ vadinama kritine sritimi.

Hipotezės atmetimo tikimybė

$$\beta(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\theta}}\{\mathbf{X} \in K\}$$

vadinama kriterijaus *galios funkcija*, arba trumpiau – kriterijaus galia.

Naudodami statistinį kriterijų galime padaryti tokias klaidas:

1. Galime atmesti hipotezę $H : \boldsymbol{\theta} \in \boldsymbol{\Theta}_H$ tada, kai ji teisinga. Tokia klaida vadinama *pirmosios rūšies klaida.* Jeigu tikroji parametro reikšmė yra $\boldsymbol{\theta}$ ir $\boldsymbol{\theta} \in \boldsymbol{\Theta}_H$, tai šios klaidos tikimybė yra $\beta(\boldsymbol{\theta})$.

2. Galime priimti hipotezę $H : \boldsymbol{\theta} \in \boldsymbol{\Theta}_H$ tada, kai ji klaidinga. Tokia klaida vadinama *antrosios rūšies klaida.* Jeigu tikroji parametro reikšmė yra $\boldsymbol{\theta}$ ir $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\bar{H}}$, tai šios klaidos tikimybė yra $1 - \beta(\boldsymbol{\theta})$.

Didinant kritinę sritį K pirmosios rūšies klaidos tikimybė $\beta(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\theta}}\{\mathbf{X} \in K\}, \boldsymbol{\theta} \in \boldsymbol{\Theta}_H$, didėja, bet tada antrosios rūšies klaidos tikimybė $1 - \beta(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\bar{H}}$, mažėja. Nerasime srities, kuri minimizuotų abiejų rūšių klaidas vienu metu.Todėl parenkamas artimas nuliui skaičius α (paprastai imama 0,1; 0,05; 0,01 ir pan.) ir nagrinėjami tik tokie kriterijai, kurie tenkina sąlygą $\sup_{\boldsymbol{\theta} \in \Theta_H} \beta(\boldsymbol{\theta}) \leq \alpha$. Skaičius

$$\alpha = \sup_{\boldsymbol{\theta} \in \Theta_H} \beta(\boldsymbol{\theta})$$

vadinamas *reikšmingumo lygmeniu.* Taigi reikšmingumo lygmuo yra maksimali hipotezės atmetimo tikimybė, kai hipotezė yra teisinga.

Tolydžių skirstiniu atveju dažnai egzistuoja pasirinkto reikšmingumo lygmens α kriterijus, o diskrečių skirstinių atveju ieškomas kriterijus, kurio reikšmingumo lygmuo kuo artimesnis pasirinktąjam (bet nedidesnis už jį).

Kritinė sritis dažniausiai turi vieną iš tokių trijų pavidalų:

$$1)\ K_1 = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq c_1\}; \quad 2)\ K_2 = \{\boldsymbol{x} : T(\boldsymbol{x}) \leq c_2\};$$

$$3)\ K_3 = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq d_1 \ \text{ arba } \ T(\boldsymbol{x}) \leq d_2\};$$

čia $T = T(\mathbf{X})$ yra kokia tai vienamatė empirinė charakteristika, tokia kaip empirinis vidurkis, empirinė dispersija ar jų funkcija.

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

Apibrėžkime P reikšmes tokio tipo kritinėms sritims lygybėmis:

$$1)\ pv = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_H} \mathbf{P}_{\boldsymbol{\theta}}\{T \geq t\}; \quad 2)\ pv = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_H} \mathbf{P}_{\boldsymbol{\theta}}\{T \leq t\};$$

$$3)\ pv = 2\min(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_H} \mathbf{P}_{\boldsymbol{\theta}}\{T \geq t\},\ \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_H} \mathbf{P}_{\boldsymbol{\theta}}\{T \leq t\}).$$

Įrodoma, kad eksperimente, kuriame statistika T įgijo reikšmę t, hipotezė H atmetama reikšmingumo lygmens α kriterijumi tada ir tik tada, kai $pv \leq \alpha$.

(a) *Hipotezė apie normalaus skirstinio vidurkio reikšmę.*

Formulės.

Tarkime, kad $X_1, ..., X_n$ yra imtis, $X_i \sim N(\mu, \sigma^2)$.

Hipotezė $H_1 : \mu \geq \mu_0$ (arba $H_1 : \mu > \mu_0$) atmetama, kai $\frac{\sqrt{n}(\bar{X}-\mu_0)}{s} < -z_\alpha$.

Hipotezė $H_2 : \mu \leq \mu_0$ (arba $H_2 : \mu < \mu_0$) atmetama, kai $\frac{\sqrt{n}(\bar{X}-\mu_0)}{s} > z_\alpha$.

Hipotezė $H_3 : \mu = \mu_0$ atmetama, kai $|\frac{\sqrt{n}(\bar{X}-\mu_0)}{s}| > z_{\alpha/2}$.

p reikšmės yra atitinkamai

$$pv = CDF.T(t, n-1), \quad pv = 1 - CDF.T(t, n-1),$$

$$pv = 2\min(CDF.T(t, n-1), 1 - CDF.T(t, n-1)) = 2(1 - CDF.T(|t|, n-1));$$

čia t yra stebėta $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{s}$ reikšmė.

Skaičiavimas.

Šiuo konkrečiu atveju reikia patikrinti hipotezę H_2.

Analyze - Compare Means - One - Sample T Test

Test Variable "Pardavimai". Test Value 520. OK .

Lentelėje "One sample statistics pateikiamos imties empirinės charakteristios (empirinis vidurkis, empirinis kvadratinis nuorypis ir t.t.).

Lentelėje "One sample test" pateikiama statistikos t reikšmė (0,599), jos skirstinio laisvės laipsniai $n-1$ (11), p-reikšmė atitinkanti dvipusę alternatyvą ($pv = 2\min(CDF.T(t, n-1), 1 - CDF.T(t, n-1)) = 0,561$).

Kadangi tikriname hipotezę $H_2 : \mu \leq 520$, tai p-reikšmė yra $pv = 1 - CDF.T(t, n-1)$. Atkreipkime dėmesį, kad gavome $t > 0$, todėl $CDF.T(t, n-1) > 0,5, 1 - CDF.T(t, n-1) < 0,5$, taigi $2\min(CDF.T(t, n-1), 1 - CDF.T(t, n-1)) = 2(1 - CDF.T(t, n-1))$. Tokiu būdu $pv = 1 - CDF.T(t, n-1) = 0,561/2 = 0,28$. Duomenys neprieštarauja hipotezei. Nors empirinis pardavimų vidurkis gavosi 522,17, toks pardavimų padidėjamas statistiškai nėra reikšmingas (nes imtis nedidelė).

Pastaba: prieš tikrinant hipotezę galima būtų vizualiai patikrinti normalumo prielaidoą:

Graphs - Histogram

Kintamasis "Pardavimai". Pažymėkite "Display normal curve". OK

Forma turėtų būti apytikriai varpo formos.

Jei nenusprendžiate, pabandykite Kolmogorovo - Smirnovo kriterijų.

Analyze - Nonparametric Tests - 1 - Sample K - S

Kintamasis "Pardavimai". Pažymėkite "Normal box". OK. Pažiūrite į p-reikšmę.

Jei normalumo neatmeta, pereinama prie Stjudento kriterijaus.

(b) *Hipotezė apie proporcijos reikšmę.*

Formulės.

Tarkime, kad eksperimento metu gali įvykti įvykis A arba jam priešingas įvykis. Atliekama n nepriklausomų eksperimentų. Pažymėkime X_i įvykio A indikatorių, įgyjantį reikšmę 1, kai įvyksta įvykis A ir reikšmę 0 priešingu atveju. Taigi gaunama imtis $X_1, ..., X_n$. Jei Kažkoks požymis gali įgyti tik dvi reikšmes, tai įvykį A galima apibrėžti kaip įvykį, kuris reiškia,

kad požymis įgijo pirmąją reikšmę. Tuo atveju p galima interpretuoti kaip populiacijos dalį (proporciją), kuriai požymis įgija pirmąją reikšmę. A.d. X_i turi Bernulio skirstinį: $X_i \sim B(1,p)$. R

Pažymėkime $S_n = \sum_{i=1}^n X_i \sim B(1,p)$. Kuo p didesnis, tuo didesnes reikšmes įgyja S_n. Todėl

$H_2 : p \le p_0$ atmetama, jei $S_n \ge c$, $pv = \mathbf{P}\{S_n \ge s_n\} = \sum_{k=s_n}^n C_n^k p_0^k (1-p_0)^{n-k} = I_{p_0}(s_n, n-s_n+1)$, $I_x(a,b)$ žymi beta skirstinio su a ir b laisvės laipsnių pasiskirstymo funkciją, o $s-n$ yra stebėtoji S_n reikšmė. Atkreipkite dėmesį, kad beta skirstinio atveju p.f. argumentas rašomas indekse, o ne skliaustuose.

$H_1 : p \ge p_0$ atmetama, jei $S_n \le c$, $pv = \mathbf{P}\{S_n \le s_n\} = \sum_{k=0}^{s_n} C_n^k p_0^k (1-p_0)^{n-k} = 1 - I_{p_0}(s_n+1, n-s_n)$.

$H_3 : p = p_0$ atmetama, jei $S_n \ge c_1$ arba $S_n \le c_2$, $pv = 2\min(\mathbf{P}\{S_n \ge s_n\}, \mathbf{P}\{S_n \le s_n\}) = \min(I_{p_0}(s_n, n-s_n+1), 1 - I_{p_0}(s_n+1, n-s_n))$.

Taigi p-reišmės skaičiavimui užtenka paskaičiuoti beta skirstinio pasiskirstymo funkcijos reikšmes. Pastebėkime, kad SPSS naudoja pažymėjimus: $I_x(a,b) = CDF.BETA(x,a,b)$.

(c) Tas pats.

(d) *Hipotezė apie koreliacijos koeficiento lygybę nuliui.*

Formulės.

Tarkime $(X_1,\ Y_1)^T, ..., (X_n,\ Y_n)^T$ yra imtis iš dvimačio normaliojo skirstinio

$$N_2(\boldsymbol{\mu},\ \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = (\mu_1, \mu_2)^T, \quad \boldsymbol{\Sigma} = [\sigma_{ij}]_{2\times 2},$$

$$\sigma_{11} = \sigma_1^2,\ \sigma_{22} = \sigma_2^2,\ \sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2;\ 0 < \sigma_1,\ \sigma_2 < \infty,\ -1 < \rho < 1.$$

Hipotezių apie koreliacijos koeficiento reikšmes tikrinimo kriterijai paprastai grindžiami empiriniu koreliacijos koeficientu

$$\hat{\rho} = r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}.$$

Hipotezė $H_0 : \rho = 0$ atmetama, jei

$$|T| > t_{\alpha/2}(n-2), \quad T = r\sqrt{\frac{n-2}{1-r^2}},$$

nes kai teisinga hipotezė, tai $T \sim S(n-2)$.

p-reikšmė skaičiuojama pagal formulę $pv = 2\min(F_T(t), 1 - F_T(t)) = 2(1 - F_T(|t|))$; čia t yra stebėtoji T reikšmė. Kadangi $T \sim S(n-2)$, tai pagal SPSS pažymėjimus $F_T(t) = CDF.T(t, n-2)$.

Duomenų įvedimas.

Į pirmą stulpelį "Stažas" suvedame stažo reikšmes, į gretima stulpelį "Atlyginimas" suvedame atitinkamas atlyginimo reikšmes.

Skaičiavimas.

Analyze - Correlate - Bivariate

Į langą "Variables" įmeskite kintamuosius "Stažas" ir "Atlyginimas". Prie "Correlation coefficients" pažymime "Pearson". Prie "Test of significance" pažymime "Two-tailed". OK.

Rezultatai.

Lenntelėje "Correlations" pateikta Pearsono koreliacijos koeficiento reikšmė ir p-reikšmė.

(e) *Dviejų priklausomų normaliųjų imčių vidurkių palyginimas: Stjudento kriterijus.*

Formulės.

Tarkime $(X_1,\ Y_1)^T, ..., (X_n,\ Y_n)^T$ yra imtis iš dvimačio normaliojo skirstinio

$$N_2(\boldsymbol{\mu},\ \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = (\mu_1, \mu_2)^T, \quad \boldsymbol{\Sigma} = [\sigma_{ij}]_{2\times 2},$$

$$\sigma_{11} = \sigma_1^2,\ \sigma_{22} = \sigma_2^2,\ \sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2;\ 0 < \sigma_1,\ \sigma_2 < \infty,\ -1 < \rho < 1.$$

Reikia patikrinti hipotezę $H_1 : \mu_1 \geq \mu_2$, $H_2 : \mu_1 \leq \mu_2$, $H_3 : \mu_1 = \mu_2$ atitinkamai su alternatyvomis $\bar{H}_1 : \mu_1 < \mu_2$, $\bar{H}_2 : \mu_1 > \mu_2$, $\bar{H}_3 : \mu_1 \neq \mu_2$.

Pažymėkime $Z_i = X_i - Y_i,\ i = 1, 2, ..., n$. Tada $Z_1, ..., Z_n$ yra paprastoji imtis a. d. $Z \sim N(\mu,\ \sigma^2)$, $\mu = \mu_1 - \mu_2$, $\sigma^2 = \sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2$. Vietoje hipotezių H_1, H_2, H_3 galime tikrinti analogiškas hipotezes apie normaliojo skirstinio vidurkio reikšmes μ pagal imtį $Z_1, ..., Z_n$, kai dispersija σ^2 yra nežinoma:

Hipotezė H_1 atmetama, kai $T = \frac{\sqrt{n}\bar{Z}}{s_Z} < -z_\alpha$.

Hipotezė H_2 atmetama, kai $T > z_\alpha$.

Hipotezė H_3 atmetama, kai $|T| > z_{\alpha/2}$.

p-reikšmės yra atitinkamai $CDF.T(t, n-1)$, $1 - CDF.T(t, n-1)$ ir $2(1 - CDF.T(|t|, n-1))$. Čia t yra stebėtoji statistikos T reikšmė.

Duomenų įvedimas.

Pirmame stulpelyje "Svorisprieš" pateikiamos svorio reikšmės prieš fizinių pratimų kursą, o gretimame stulpelyje "Svorispo" pateikiamos svorio reikšmės po fizinių pratimų kurso.

Skaičiavimas.

Analyze-Compare means-Paired samples test

Kintamuosius "Svorisprieš" ir "Svorispo" į langą "Paired variables".

Options 90% ok

Rezultatai.

Lentelėje "Paired samples statistics" pateikiamos abiejų imčių empirinės charakteristikos (vidurkiai, vidutniai kvadratiniai nuokrypiai, imčių dydžiai).

Lentelėje "Paired samples correlation" pateikiamas Pirsono koreliacijos koeficientas tarp svorio prieš ir svorio po.

Lentelė "Paired samples test" yra pagrindinė. Joje pateikiama $\bar{Z}$ (mean), s_Z (std.deviation), statistikos T reikšmė ($t = 5,657$), laisvės laipsniai $n-1 = 4$ ir dvipusė p-reikšmė (sig=0,005).

.Kadangi tikrinama hipotezė H_3 atmetama, tai darome išvadą, kad vidutinis svoris pakito.

Jeigu būtume tikrinę hipotezę "Svoris sumažėjo", t.y. hipotezę H_1, tai $pv = CDF.T(t, n-1) > 0,5$, nes $t > 0$. Taigi duomenys neprieštarautų hipotezei. Jei tikrintume hipotezę "Svoris padidėjo", t.y. hipotezę H_2, tai $pv = 1 - CDF.T(t, n-1) < 0,5$. Kadangi $CDF.T(t, n-1) > 0,5$, tai $0,005 = 2(1 - CDF.T(|t|, n-1)) = 2(1 - CDF.T(t, n-1))$, tai vienpusė $pv = 1 - CDF.T(t, n-1) = 0,0025$. Hipotezė, kad svoris padidėjo atmetama.

(f) Tas pats

(g) *Dviejų nepriklausomų normaliųjų imčių vidurkių palyginimas: Stjudento kriterijus.*

Formulės.

Tarkime, kad stebimos dvi nepriklausomos normaliosios imtys: $X_1, ..., X_n$ ir $Y_1, ..., Y_m$, $X_i \sim N(\mu_1, \sigma^2)$, $Y_j \sim N(\mu_2, \sigma^2)$. Reikia patikrinti hipotezę $H_1 : \mu_1 \geq \mu_2$, $H_2 : \mu_1 \leq \mu_2$, $H_3 : \mu_1 = \mu_2$ atitinkamai su alternatyvomis $\bar{H}_1 : \mu_1 < \mu_2$, $\bar{H}_2 : \mu_1 > \mu_2$, $\bar{H}_3 : \mu_1 \neq \mu_2$.

Pažymėkime

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{s_1^2(n-1) + s_2^2(m-1)}} \sqrt{\frac{mn(m+n-2)}{m+n}}.$$

Hipotezė H_1 atmetama, kai $T < -t_\alpha(m+n-2)$.

Hipotezė H_2 atmetama, kai $T > t_\alpha(m+n-2)$.

Hipotezė H_3 atmetama, kai $|T| > t_{\alpha/2}(m+n-2)$.

p-reikšmės yra atitinkamai

$CDF.T(t, m+n-2)$, $1 - CDF.T(t, m+n-2)$ ir $2(1 - CDF.T(|t|, m+n-2))$. Čia t yra stebėtoji statistikos T reikšmė.

Duomenų įvedimas.

Suvedame visus atsparumo duomenis į pirmą stulpelį. Antrame stulpelyje įvedame 1, jei matavimas atliktas pirmąją dieną, ir 0, jei antrąją.

Skaičiavimas

Analyze-Compare means-Independent samples T test

Test variable-atsparum

Grouping variable-diena

Define groups: Group 1-0, Group 2-1

Kadangi SPSS atiminėja iš Group 2 vidurkio Group 1 vidurkį, tai todėl pirmos dienos duomenis, kuriuos pažymėjome vienetuku, priskiriame Group 2.

Rezultatai

Lentelėje "Group statistics" pateikiamos abiejų imčių empirinės charakteristikos. Lentelėje "Independent samples test" pateikiama statistikos T reikšmė ($t = 5,321$), statistikos laisvės laipsnių skaičius ($m + n - 2 = 23$), p-reišmė (dvipusė) (sig=0,000). Kadangi tikrinama hipotezė H_3, tai ir reikalinga būtent dvipusė p-reikšmė. Hipotezė atmetama.

(h) Tas pats, bet reikia atsižvelgti į formuluotėje duotą pastabą, todėl pradžioje verta nubrėžti kiekvienos imties duomenų histogramą (žr. 6a nurodymus) ir pastebėjus, kad tos histogramos labai nesimetriškos. logaritmuoti reakcijos laiko duomenis ir po to vėl nubrėžti histogramas, kurios jau turi formą panašią į varpo. Taigi vietoje pradinio stulpelio naudojamas logaritmuotų duomenų stulpelis. Yra dar vienas skirtumas: tikrinama hipotezė H_2, o ne H_3, todėl naudojama vienpusė p-reikšmė.

(i) *Dviejų nepriklausomų normaliųjų imčių dispersijų palyginimas: Fišerio kriterijus.*

Formulės.

Tarkime, kad stebimos dvi nepriklausomos normaliosios imtys: $X_1, ..., X_n$ ir $Y_1, ..., Y_m$, $X_i \sim N(\mu_1, \sigma_1^2)$, $Y_j \sim N(\mu_2, \sigma_2^2)$. Reikia patikrinti hipotezę $H_1 : \sigma_1 \geq \sigma_2$, $H_2 : \sigma_1 \leq \sigma_2$, $H_3 : \sigma_1 = \sigma_2$ atitinkamai su alternatyvomis $\bar{H}_1 : \sigma_1 < \sigma_2$, $\bar{H}_2 : \sigma_1 > \sigma_2$, $\bar{H}_3 : \sigma_1 \neq \sigma_2$.

Pažymėkime

$$F = s_1^2/s_2^2.$$

Hipotezė H_1 atmetama, kai $F < f_{1-\alpha}(n-1, m-1)$.

Hipotezė H_2 atmetama, kai $F > f_\alpha(n-1, m-1)$.

Hipotezė H_3 atmetama, kai $F < f_{1-\alpha/2}(n-1, m-1)$ arba $F > f_{\alpha/2}(n-1, m-1)$;

čia $f_\alpha(n-1, m-1)$ žymi Fišerio skirstinio su $n-1$ ir $m-1$ laisvės laipsniais α kritinę reikšmę.

p-reikšmės yra atitinkamai

$CDF.F(f, m-1, n-1)$, $1 - CDF.F(f, m-1, n-1)$ ir $2\min\{(1 - CDF.F(f, m-1, n-1), 1 - CDF.F(f, m-1, n-1))$. Čia f yra stebėtoji statistikos F reikšmė.

Skaičiavimas.

Tikrinama hipotezė H_3. Pratime 6b suskaičiuotos s_1 ir s_2 reikšmės. Suskaičiuojame

$$f = s_1^2/s_2^2,$$

po to p-reikšmę (dvipusę).

(j) *Proporcijų palyginimas.*

a) Naudojant sąlyginį dažnio skirstinį.

Formulės

Proporcijos palyginimo hipotezės tikrinimas grindžiamas tuo, kad jei $S_1 \sim B(n, p_1)$, $S_2 \sim B(n, p_2)$, tai kai teisinga lygybė $p_1 = p_2$

$$\mathbf{P}\{S_1 = j | S_1 + S_2 = s\} = \frac{C_n^j C_m^{s-j}}{C_{m+n}^s},$$

taigi sąlyginis S_1 skirstinys žinant $S_1 + S_2$ yra hipergeometrinis.

Jei tikrinama hipotezė $H_1 : p_1 \le p_2$ su alternatyva $\bar{H}_1 : p_1 > p_2$, tai hipotezė atmetama, kai $S_1 \ge c_1$, todėl

$$pv = Z_1 = \mathbf{P}\{S_1 \ge s_1 | S_1 + S_2 = s\} = \sum_{j=s_1}^{\min(s,n)} \frac{C_n^j C_m^{s-j}}{C_{m+n}^s} =$$

$$CDF.HYPER(\min(S,n), N, S, n) - CDF.HYPER(S_1 - 1, N, S, n);$$

čia $S = S_1 + S_2$, $N = n + m$, o $CDF.HYPER$ yra hipergeometrinio skirstinio pasiskirstymo funkcija.

Jei tikrinama hipotezė $H_2 : p_1 \ge p_2$ su alternatyva $\bar{H}_2 : p_1 < p_2$, tai hipotezė atmetama, kai $S_1 \le c_2$, todėl

$$pv = Z_2 = \mathbf{P}\{S_1 \le s_1 | S_1 + S_2 = s\} = \sum_{j=0}^{s_1} \frac{C_n^j C_m^{s-j}}{C_{m+n}^s} =$$

$$Z_2 = CDF.HYPER(S_1, N, S, n).$$

Jei tikrinama hipotezė $H_3 : p_1 = p_2$ su alternatyva $\bar{H}_3 : p_1 \ne p_2$, tai hipotezė atmetama, kai $S_1 \ge c_3$ arba $S_1 \le c_4$, todėl $pv = 2\min(Z_1, Z_2)$.

Taigi p-reikšmių skaičiavimui užtenka suskaičiuoti hipergeometrinio skirstinio pasiskirstymo funkcijos reikšmes.

b) Hipotezę galima patikrinti ir kitų kriterijumi: chi kvadrato homogeniškumo kriterijų (teoriją žiūrėkite 7e nurodymuose).

Duomenų įvedimas

Apibrėžiame du kintamuosius: Šalis, įgyjanti reikšmes 1 ir 2, bei Atsakymas, įgyjantis reikšmes 1 (etiketė taip) ir 0 (etiketė ne).

Taigi turime 120 porų $(1,1)$, 180 porų $(1,0)$, 148 poras $(2,1)$ ir 52 poras $(2,0)$.

Analyze-Descriptive statistics-Crosstabs

Šalį į langą Row, Atsakymą į Column

Statistics: Chi-square

Paprastesnis duomenų įvedimas

Lentelės duomenis suvedame kitaip:

Dažniai	Šalis	Atsakymas
120	1	1
180	1	0
148	2	1
52	2	0

1 žingsnis: *Data-Weight cases-Weight cases by.* Į "frequency variable" įmetame "dažniai".

2 žingsnis: Analyze-Descriptive statistics-Crosstabs

Šalį į langą Row, Atsakymą į Column

Statistics: Chi-square

Rezultatai

Pateikiama pradinė duomenų lentelė ir chi kvadrato statistikos reikšmę ir pv.

(k) Tas pats.

7. **Dažnių lentelės**

(a) *Hipotezės apie polinominio skirstinio parametrų reikšmes tikrinimas.*

Formulės. Tarkime, kad atliekama n nepriklausomų eksperimentų. Kiekvieno eksperimento metu gali įvykti vienas iš nesutaikomų įvykių $A_1, A_2, \cdots, A_k$, kurių skaičius $k > 2$. Įvykio A_i tikimybė kiekviename bandyme pastovi ir lygi p_i, $\quad p_1 + p_2 + \cdots + p_k = 1$. Pažymėkime

ν_i įvykio A_i, $\quad i = 1, 2, \cdots, k$, pasirodymų skaičių visų eksperimentų metu. Tada a. v. $\nu = (\nu_1, \nu_2, \cdots, \nu_k)^T$ skirstinys vadinamas k-mačiu *polinominiu*. Žymėsime $\nu \sim M_n(p_1, ..., p_k)$:

$$\mathbf{P}\{\nu_1 = m_1, \cdots, \nu_k = m_k\} = \frac{n!}{m_1!, \cdots, m_k!} p_1^{m_1} \cdots p_k^{m_k},$$

$$0 \le m_j \le n, \quad j = 1, \cdots, k; \quad m_1 + \cdots + m_k = n.$$

Koordinatės ν_i skirstinys yra binominis: $\nu_i \sim B(n, p_i)$.
A. d. ν_i vidurkis ir dispersija yra

$$\mathbf{E}\nu_i = np_i, \quad D\nu_i = np_i(1 - p_i), \quad i = 1, \cdots, k,$$

o kovariacijos yra

$$\mathbf{Cov}\,(\nu_i, \nu_j) = -np_ip_j, \quad i \ne j, \quad i, j = 1, \cdots, k.$$

Jei $(\nu_1, ..., \nu_k) \sim M_n(p_1, ..., p_k)$, tai hipotezė $H : p_1 = p_1^0, ..., p_k = p_k^0$ atmetama su apytiksliu reikšmingumo lygmeniu α, jei

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i^0)^2}{np_i^0} > \chi_\alpha^2(k-1).$$

Ši statistika naudojama todėl, kad kai teisinga hipotezė, tai $\mathbf{E}\nu_i = np_i^0$, taigi skirtumų kvadratai $(\nu_i - np_i^0)^2$ maži. Jei hipotezė neteisinga, šie skirtumų kvadratai įgija didesnes reikšmes.

Duomenų įvedimas

Stulpelyje "Spalva" įvedame spalvas 1,2,3,4,5 ir lange "Variable vue" stulpelyje "Values" parašome 1 - Raudona 2 - Geltona ir t.t.), o stulpelyje "Dažnis" įvedame atitinkamus dažnius.

Skaičiavimas

Pradžioje *Data - Weight cases*

Įmetame "Dažnis" į "Frequency variable".

Analyze - Nonparametric Tests - Legacy Dialogs - Chi-square

Į "Test variable list" įmetame "Spalva". Paliekame pažymėtą "All categories equal".

OK.

Rezultatai.

Lentelėje "Frequences" pateikiamos visų spalvų dažnių ν_i reikšmės, jų tikėtinų vidurkių np_i^0 reikšmės ir skirtumų $\nu_i - np_i^0$ reikšmės. Pagrindinėje lentelėje "Test statistics" pateikiama statistikos χ^2 reikšmė ($\chi^2 = 35$ šiuo konkrečiu atveju), laisvės laipsnių skaičius $k - 1 = 4$ ir p-reikšmė $Asym.sig = 0.000$, tiksliau $4.645 * 10^{-7}$ Hipotezė apie vienodą spalvų vertinimą atmetama. Taigi duomenys neprieštarauja hipotezei, kad pirkėjai nevienodai vertina automobilių spalvąs.

(b) Tas pats, tik vietoje "All categories equal" pažymime "Values". Tada įvedame 0,30;0,60;-0,08;0,02. OK

Gauname $\chi^2 = 10.367$, $Asymp.sig = 0.016$. Rinkos analitiko hipotezė atmetama.

(c) *Požymių nepriklausomumo tikrinimas.*

Formulės:

Tarkime

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

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

yra dvi nesutaikomų, sudarančių pilną grupę įvykių, kuriuos galime stebėti eksperimento metu, sistemos. Dažniausiai įvykis A_i reiškia, kad koks nors požymis įgijo savo i-ją reikšmę

(šiuo atveju vaikų skaičius pateko į vieną iš duotų grupių), o įvykis B_j reiškia, kad koks nors kitas požymis įgijo savo j-ją reikšmę (šiuo atveju metinės pajamos pateko į j-jį intervalą). Pažymėkime

$$\pi_{ij} = \mathbf{P}\{A_i \cap B_j\}, \quad \pi_{i\cdot} = \sum_{j=1}^{r} \pi_{ij}, \quad \pi_{\cdot j} = \sum_{i=1}^{s} \pi_{ij}.$$

Nepriklausomumo hipotezė (dviejų atsitiktinių įvykių sistemų)

$$H_0 : \pi_{ij} = \mathbf{P}\{A_i \cap B_j\} = \mathbf{P}\{A_i\}\mathbf{P}\{B_j\} = \pi_{i\cdot}\pi_{\cdot j}, \quad i = 1, \ldots, s;\ j = 1, \ldots, r. \tag{0.1}$$

Alternatyva hipotezei H_0 yra

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

Atliekama n stebėjimų, kurių metu kiekvienam iš n objektų nustatomos požymių reikšmės. Pažymėkime U_{ij} įvykio $A_i \cap B_j$ pasirodymo skaičių, $\sum_{i=1}^{s}\sum_{j=1}^{r} U_{ij} = n$, ir

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

Tikimybių $\pi_{i\cdot}$ natūralūs įvertiniai:

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

Esant teisingai hipotezei H_0, polinominio skirstinio tikimybių $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ natūralūs įvertiniai yra

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot} \cdot \hat{\pi}_{\cdot j} = \frac{U_{i\cdot}}{n}\frac{U_{\cdot j}}{n}.$$

Naudodami šiuos įvertinius gauname statistiką

$$X_n^2 = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{(U_{ij} - n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j})^2}{n\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{r} \frac{U_{ij}^2}{U_{i\cdot}U_{\cdot j}} - 1\right). \tag{0.2}$$

Gautoji statistika asimptotiškai (kai $n \to \infty$) pasiskirsčiusi pagal chi-kvadrato skirstinį su

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

laisvės laipsnių.

Chi-kvadrato nepriklausomumo kriterijus: hipotezė H_0 atmetama asimptotiniu α lygmens kriterijumi, kai

$$X_n^2 > \chi_\alpha^2((r - 1)(s - 1)). \tag{0.3}$$

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

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

Duomenų įvedimas

Dažniai	Vaikų skaičius	Pajamos
2161	0	0-1
2755	1	0-1
...	...	...
3577	0	1-2
...	...	...
14	4	3-

Data-Weight cases-Weight cases by. Į "frequency variable" įmetame "dažniai".
Analyse-Descriptive statistics-Crosstabs
Statistics-"chi square".
Įmetame į "Rows" kintamąjį "Vaikų skaičius" ir į "Columns" kintamąjį "Pajamos".
Lentelė "vaikų skaičius*Pajamos Crosstabulation sutampa užduotyje pateiktą lentelę.
Lentelėje "Chi square tests" pirmoje eilutėje pateikiama chi kvadrato statistikos $X_n^2 = PearsonChi-Square$ reikšmę, laisvės laipsnių skaičius $(r-1)(s-1)$ ir p-reikšmė Asympt. Sig.
Kai chi kvadrato statistikos reikšmė viršija laisvės laipsnių skaičių, tai vienpusė p-reikšmė gaunama dalijant dvipusę, p-reikšmę (kuri duodama lentelėje) iš dviejų.
Pirmo žingsnio nereikia, jei įvedami pradiniai duomenys.

(d) Tas pats.

(e) *Homogeniškumo hipotezės tikrinimas.*
Formulės.
Sakykime, kad yra s nepriklausomų objektų grupių; i-tosios grupės objektų skaičių pažymėkime n_i. Tarkime, kad

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

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

$$\pi_{ij} = \mathbf{P}\{B_j | \, j\text{-asis objektas priklauso } i\text{-ajai grupei}\}. \tag{0.4}$$

Homogeniškumo hipotezė (faktoriaus B atžvilgiu):

$$H_0 : \pi_{1j} = ... = \pi_{sj} := \pi_j, \; j = 1, ..., r, \tag{0.5}$$

kuri reiškia, kad esant fiksuotam j įvykio B_j tikimybė yra ta pati visų grupių objektams. Minėto pavyzdžio atveju hipotezė reiškia, kad tikimybė gauti j balų visų profesijų atstovams vienoda, kokį bepaimtume j.

Pažymėkime U_{ij} i-osios grupės objektų skaičių, kuriems įvyko įvykis B_j, $U_{i1} + ... + U_{ir} = n_i$. Esant teisingai hipotezei H_0 tikimybių π_{ij} DT įvertiniai yra

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

Naudodami šiuos įvertinius sukonstruojama statistika

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

Chi-kvadrato homogeniškumo kriterijus: homogeniškumo hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$X_n^2 > \chi_\alpha^2((r-1)(s-1)). \tag{0.6}$$

Duomenų įvedimas : kaip c) ir d).

Dažniai	Srautas	Pažymys
33	1	Nepatenkinamai
43	1	Patenkinamai
80	1	Gerai
144	1	Labai gerai
39	2	Nepatenkinamai
35	2	Patenkinamai
72	2	Gerai
154	2	Labai gerai

Skaičiavimas: kaip d).

(f) Tas pats.

8. **Neparametriniai kriterijai.**

(a) *Spirmeno koreliacijos koeficientas.*

Formulės.

Stebėjimai $(X_i, Y_i)^T$, $i = 1, 2...., n$, išdėstomi taip, kad Y_i sudarytų didėjančią seką ir tada duomenys pakeičiami jų rangais. Gauname eilute $(1, 2, \ldots, n)$ ir $R_1, ..., R_n$. Spirmeno koreliacijos koeficientas yra

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - i)^2. \tag{0.7}$$

Spirmeno nepriklausomumo kriterijus: hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, kai

$$r_s \le c_1 \quad \text{arba} \quad r_s \ge c_2; \tag{0.8}$$

kur c_1 – minimalus, o c_2 – maksimali r_S reikšmės, tenkinančios nelygybes

$$\mathbf{P}\{r_s \le c_1\} \le \alpha/2, \quad \mathbf{P}\{r_s \ge c_2\} \le \alpha/2.$$

Kai n didelis, a. d. r_S skirstinys aproksimuojamas normaliuoju:

$$Z_n = \sqrt{n-1}\, r_S \xrightarrow{d} Z \sim N(0,\, 1). \tag{0.9}$$

Asimptotinis Spirmeno nepriklausomumo kriterijus: nepriklausomumo hipotezė atmetama asimptotiniu α lygmens kriterijumi, kai

$$|Z_n| > z_{\alpha/2}. \tag{0.10}$$

Vidutinio didumo imtims statistikos $t_n = \sqrt{n-2}\frac{r_S}{\sqrt{1-r_S^2}}$ skirstinys aproksimuojamas Stjudento skirstiniu $S(n-2)$.

Asimptotinis Spirmeno nepriklausomumo kriterijus grindžiamas Stjudento skirstiniu: nepriklausomumo hipotezė atmetama asimptotiniu α lygmens kriterijumi, kai $|t_n| > t_{\alpha/2}(n-2)$.

Skaičiavimas

Analyze - Correlate - Bivariate

Į langelį "Variables" įmeskite kintamuosius "Rinkosdalis" ir "Reklamosislaidos". Pažymėkite "Spearman" OK

Rezultatai.

Lentelėje "Correlate" pateikiama Spirmeno koreliacijos koeficiento reikšmė $r_S = 0,821$, o po lentele sakoma, kad koreliacija reikšminga, kai reikšmingumo lygmuo yra $0,05$, kas reiškia, kad p-reikšmė (dvipusė) yra mažesnė už $0,05$. Taigi hipotezė apie požymių nepriklausomumą atmetama. Taigi duomenys neprieštarauja tam, kad kintamieji priklausomi.

(b) *Dviejų nepriklausomų imčių palyginimas: Vilkoksono-Mano-Vitnio kriterijus*

Formulės.

Tegu $\mathbf{X} = (X_1, ..., X_m)^T$ ir $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ yra dvi nepriklausomos paprastosios imtys gautos stebint absoliučiai tolydžius a. d. $X \sim F(x)$ ir $Y \sim G(x)$. Reikia patikrinti hipotezę, kad pasiskirstymo funkcijos sutampa:

$$H_0 : F(x) = \mathbf{P}\{X \le x\} \equiv \mathbf{P}\{Y \le x\} = G(x). \tag{0.11}$$

Tarkime, kad alternatyva yra *poslinkio*: egzistuoja toks $\theta \ne 0$, kad su visais $x \in \mathbf{R}$ teisinga hipotezė

$$H_1 : G(x) = F(x - \theta). \tag{0.12}$$

Pažymėkime $R_1, R_2, ..., R_m$ stebėjimų $X_1, ..., X_m$ rangus jungtinėje didumo $m+n$ imtyje $(X_1, ..., X_m, Y_1, ..., Y_n)^T$. Tada Vilkoksono kriterijaus statistika W yra lygi šių rangų sumai:

$$W = \sum_{i=1}^{m} R_i.$$

Pažymėkime U skaičių tokių atvejų, kai pirmosios imties elementai viršija antrosios imties elementus:

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} h_{ij}, \quad h_{ij} = \begin{cases} 1, & \text{kai } X_i > Y_j, \\ 0, & \text{kai } X_i < Y_j. \end{cases} \tag{0.13}$$

Statistikos W ir U yra glaudžiai susiję:

$$U = W - m(m+1)/2.$$

Vilkoksono kriterijus grindžiamas statistika W, o Manio-Vitnio – statistika U. Kadangi statistikos skiriasi tik konstanta, tai abu kriterijai yra ekvivalentūs.
Kai teisinga alternatyva $\theta > 0$, tai antrosios imties elementai turės tendenciją įgyti didesnes reikšmes, negu pirmosios imties elementai, taigi rangų suma W turės tendenciją igyti mažesnes reikšmes. Atvirkščiai, kai $\theta < 0$, statistika W turės tendenciją įgyti didesnes reikšmes.

Vilkoksono kriterijus: dvipusės alternatyvos atveju hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, kai

$$W \le c_1 \quad \text{arba} \quad W \ge c_2;$$

čia c_1 yra maksimalus, o c_2 minimalus skaičiai tenkinantys nelygybes

$$\sum_{k=w_1}^{c_1} \mathbf{P}\{W = k|H_0\} \le \alpha/2. \quad \sum_{i=c_2}^{w_2} \mathbf{P}\{W = k|H_0\} \le \alpha/2.$$

Vienpusių alternatyvų atveju ($\theta > 0$ arba $\theta < 0$), kriteinė sritis yra vienpusė, t. y. turi atitinkamai pavidalą $W \ge d$ arba $W \le c$.
Jeigu m ir n yra dideli, tai statistikos W skirstinys aproksimuojamas normaliuoju. Tegu $N = m + n$. Rangų sumos W vidurkis ir dispersija yra

$$\mathbf{E}(W) = \frac{m(N+1)}{2}, \quad \mathbf{D}(W) = \frac{mn(N+1)}{12}.$$

Pažymėkime

$$Z_{m,n} = \frac{W - \mathbf{E}(W)}{\sqrt{\boldsymbol{D}(W)}} = \frac{U - \mathbf{E}(U)}{\sqrt{\boldsymbol{D}(U)}}.$$

Jeigu stebimų a. d. X ir Y skirstiniai absoliučiai tolydūs, $N \to \infty$, $m/N \to p \in (0, 1)$, tai esant teisingai hipotezei H_0 $Z_{m,n} \xrightarrow{d} Z \sim N(0, 1)$. Nustatyta, kad konvergavimas į normalųjį skirstinį gana greitas.

Asimptotinis Vilkoksono kriterijus: jeigu m ir n nėra maži, tai hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$|Z_{m,n}| > z_{\alpha/2}.$$

Jei egzistuoja sutampančių reikšmių, tai

$$\mathbf{D}(W) = \frac{mn(N+1)}{12}(1 - \frac{\mathbf{E}T}{N^3 - N});$$

čia

$$T = \sum_{i=1}^{k} T_i, \quad T_i = (t_i^3 - t_i),$$

k yra skaičius sutampančių elementų grupių apjungtoje imtyje, o t_i yra i-osios grupės didumas.

Taigi, kai m ir n nėra maži ir yra sutampančių reikšmių, tai statistika $Z_{m,n}$ modifikuojama:

$$Z^*_{m,n} = \frac{Z_{m,n}}{\sqrt{1 - T/(N^3 - N)}}.$$

Modifikuotas asimptotinis Vilkoksono kriterijus: hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$|Z^*_{m,n}| > z_{\alpha/2}.$$

Duomenų įvedimas: visas degimo laikų reikšmes suvedame į vieną stulpelį "Laikas". Antrame stulpelyje įvedame 1 ("Pigios"), 2 ("Brangios") greta atitinkamų laikų.

Skaičiavimas:

Analyze - Nonparametric Tests - 2 Independent Samples

Test variable "Laikas" .

"Define Groups": Group 1 įvedama "1" (Pigios), Group 2 įvedama 2 (Brangios).

"Test TypeŤ - Mann-Whitney U, OK.

Lentelėje "Ranks" pateiktos abiejų imčių dydžiai m ir n, atitinkamos rangų sumos $R^{(1)}$ ir $R^{(1)}$ apjungtoje imtyje, rangų aritmetiniai vidurkiai.

Lentelėje "Test statistics" pateiktos Vilkoksono ($W = 31$) ir Mano-Vitnio ($U = 10$) statistikų bei asimptotinės ($Z = -1,571$) statistikos reikšmės, p-reikšmė, gauta naudojant tikslų kriterijų ($pv = 0,138$) bei asimptotinį kriterijų ($pv_a = 0,116$). Duomenys neprieštarauja hipotezei, kad pigių ir brangių žvakučių degimo laiko skirstiniai nesiskiria.

(c) *Dviejų priklausomų imčių palyginimas: Vilkoksono žymėtųjų rangų kriterijus*

Tarkime

$$(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T,$$

yra paprastoji imtis, gauta stebint absoliučiai tolydų a. v. $(X, Y)^T$ su pasiskirstymo funkcja $F(x, y)$. Pažymėkime $F_1(x)$ ir $F_2(y)$ marginaliąsias pasiskirstymo funkcijas.

Dviejų priklausomų imčių homogeniškumo hipotezė:

$$H_0 : \ F_1(x) \equiv F_2(x).$$

Hipotezei tikrinti naudojama imtis $D_1, \ldots, D_n$ sudarytą iš skirtumų $D_i = X_i - Y_i$. Tegu R_i yra elemento $|D_i|$ rangas sekoje $|D_1|, \ldots, |D_n|$, o T^+ ir T^- – rangų, atitinkančių teigiamus ir neigiamus skirtumus D_i, sumos:

$$T^+ = \sum_{i:D_i>0} R_i, \quad T^- = \sum_{i:D_i<0} R_i. \tag{0.14}$$

Pavyzdžiui, jeigu $M_0 = 10$ ir imties realizacija yra 7, 16, 5, 8, 14, tai skirtumų $D_1, \ldots, D_5$ realizacija yra -3, 6, -5, -2, 4, o $|D_1|, \ldots, |D_5|$ – 3, 6, 5, 2, 4. Taigi rangai $R_1, \ldots, R_5$ įgijo reikšmes 2, 5, 4, 1, 3. Rangai atitinkantys teigiamus ir neigiamus skirtumus D_i yra 3, 5 ir 1, 2, 4, atitinkamai. Taigi, $T^+ = 3 + 5 = 8$, $T^- = 1 + 2 + 4 = 7$.

Praktiškai T^+ ir T^- yra patogu skaičiuoti taip: skirtumų $D_1, \ldots, D_5$ realizaciją -3, 6, -5, -2, 4 išrikiuojame jų absoliutinių didumų didėjimo tvarka ir surašome palikdami jų ženklus: -2, -3, 4, -5, 6. Priskiriame rangus irgi palikdami ženklus: -1, -2, 3, -4, 5. Tada $T^+ = 3 + 5 = 8$, $T^- = 1 + 2 + 4 = 7$.

Pakanka nagrinėti vieną iš statistikų T^+ arba T^-, kadangi jų suma $T^+ + T^- = R_1 + \cdots + R_n = n(n+1)/2$. Galimos statistikos T^+ reikšmės yra $0, 1, \ldots, n(n+1)/2$.

Esant teisingai hipotezei H_0, skirtumai D_i yra simetriškai pasiskirstę nulio atžvilgiu, todėl a. d. T^+ ir T^- skirstiniai sutampa ir $\mathbf{E}T^+ = \mathbf{E}T^-$.

Jeigu a. d. D_i turi tendenciją dažniau įgyti teigiamas reikšmes negu neigiamas, tai T^+ turi tendenciją įgyti didesnes reikšmes už T^- ir $\mathbf{E}T^+ > \mathbf{E}T^-$.

Analogiškai, jeigu a. d. D_i turi tendenciją dažniau įgyti neigiamas reikšmes negu teigiamas, tai T^+ turi tendenciją įgyti mažesnes reikšmes už T^- ir $\mathbf{E}T^+ < \mathbf{E}T^-$.
Kadangi suma $T^+ + T^- = n(n+1)/2$ yra pastovi, tai sąlygos

$$\mathbf{E}T^+ = \mathbf{E}T^-, \quad \mathbf{E}T^+ > \mathbf{E}T^-, \quad \mathbf{E}T^+ < \mathbf{E}T^-$$

yra ekvivalenčios sąlygoms

$$\mathbf{E}T^+ = n(n+1)/4 =: N, \quad \mathbf{E}T^+ > N, \quad \mathbf{E}T^+ < N.$$

Vilkoksono ranginis ženklų kriterijus: kai alternatyva dvipusė, tai hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, kai

$$T^+ \geq T^+_{\alpha/2}(n) \quad \text{or} \quad T^+ \leq T^+_{1-\alpha/2}; \tag{0.15}$$

čia $T^+_{\alpha/2}(n)$ yra mažiausias skaičius tenkinatis nelygybę $\mathbf{P}\{T^+ \geq T^+_{\alpha/2}(n)\} \leq \alpha/2$, o and $T^+_{1-\alpha/2}$ yra didžiausias skaičius, tenkinantis nelygybę $\mathbf{P}\{T^+ \leq T^+_{1-\alpha/2}\} \leq \alpha/2$. Vienpusių alternatyvų H_1 arba H_2 atveju hipotezė H_0 atmetama, kai

$$T^+ \geq T^+_{\alpha}(n) \quad \text{arba} \quad T^+ \leq T^+_{1-\alpha}. \tag{0.16}$$

Didelių imčių atveju asimptotinį kriterijų konstruojame naudodami ribinį statistikos T^+ skirstinį. Jei hipotezė H_0 teisinga, tai

$$Z_n = \frac{T^+ - \mathbf{E}(T^+)}{\sqrt{\boldsymbol{V}(T_+)}} \xrightarrow{d} Z \sim N(0,\,1),$$

$$\mathbf{E}(T^+) = \frac{n(n+1)}{4}, \quad \boldsymbol{V}(T_+) = \frac{n(n+1)(2n+1)}{24}.$$

Jeigu n yra didelis ir hipotezė H_0 teisinga, tai ji atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$|Z_n| \geq z_{\alpha/2}.$$

p-reikšmės yra $pv = 1 - CDF.Normal(z, 0, 1)$ (hipotezė, jog X_i turi tendenciją įgyti mažesnes reikšmes negu Y_i, $pv = CDF.Normal(z, 0, 1)$ (hipotezė, jog X_i turi tendenciją įgyti didesnes reikšmes negu Y_i ir $pv = 2(1 - CDF.Normal(|z|, 0, 1))$ (hipotezė $F(x) = G(x)$). Jeigu yra sutampančių reikšmių, tai statistika modifikuojama:

$$Z^*_n = \frac{Z_n}{\sqrt{1 - T/(2n(n+1)(2n+1))}};$$

čia $T = \sum_{l=1}^{k}(t_l^3 - t_l)$; k yra sutampančių grupių skaičius, o t_l – yra l-osios grupės didumas. Hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$|Z^*_n| \geq z_{\alpha/2}.$$

Duomenų įvedimas: "Prieš" ir "Po" reikšmės suvedamos į du stulpelius.
Skaičiavimas:
Analyze - Nonparametric Tests - 2 Related Samples
Objective: "Automatically compare ..." Fields: įmetame "Prieš" ir "Po".
Settings: "Automatically choose..."
Run
Rezultatai: lentelėje "Hypothesis test summary" pateikiama dvipusė p-reikšmė: 0.037. Kliktelėjus į nurodytą išvadą (šiuo atveju "reject the null hypothesis") dar pasirodo lentelė, iš kurios matosi, kad buvo skaičiuojami skirtumai "Po-Prieš" (taigi X_i=Po, Y_i =Prieš, pateikta teigiamų rangų suma $T^+ = 23.5$ bei statistikos Z reikšmė -2.091.
Kadangi hipotezė yra, jog Po turi tendenciją įgyti mažesnes reikšmes negu Prieš, tai $pv = 1 - CDF.Normal(-2.091, 0, 1) = 1 - (1 - CDF.Normal(|-2.091|, 0, 1) = 1 - 0.037/2$.
Išvada: duomenys neprieštarauja hipotezei, kad spaudimas sumažėja.

(d) *Kelių nepriklausomų imčių palyginimas: Kruskalo-Voliso kriterijus*
Tarkime, kad

$$\mathbf{X}_1 = (X_{11}, ..., X_{1n_1})^T, \quad \dots \quad , \quad \mathbf{X}_k = (X_{k1}, ..., X_{kn_k})^T$$

yra k paprastųjų imčių, gautų stebint n. a. d. $X_1, ..., X_k$ su absoliučiai tolydžiomis pasiskirstymo funkcijomis $F_1(x), ..., F_k(x)$.

Kelių nepriklausomų imčių homogeniškumo hipotezė:

$$H_0 : F_1(x) = F_2(x) = \cdots = F_k(x) =: F(x), \ \forall x \in \mathbf{R}. \tag{0.17}$$

Tarkime, kad alternatyva hipotezei H_0 yra poslinkio

$$H_1 : F_j(x) = F(x - \theta_j) \ \text{ for all } x \in \mathbf{R}, \quad j = 1, \dots, k, \quad \sum_{j=1}^{k} \theta_j^2 > 0. \tag{0.18}$$

Pažymėkime $n = \sum_{i=1}^{k} n_i$.
Stebėjimų $X_{i1}, ..., X_{in_i}$ rangus jungtinėje visų stebėjimų variacinėje eilutėje pažymėkime $R_{i1}, ..., R_{in_i}$, šių rangų sumą $R_{i\cdot} = \sum_{j=1}^{n_i} R_{ij}$, o aritmetinį vidurkį $\bar{R}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$.
Jei teisinga hipotezė, tai

$$\mathbf{E}R_{i\cdot} = \frac{n_i(n+1)}{2}, \quad \mathbf{E}\bar{R}_{i\cdot} = \frac{n+1}{2}.$$

Taigi kriterijus grindžiamas skirtumais $R_{i\cdot} - \frac{n+1}{2}$, kurie turėtų būti maži, kai teisinga hipotezė: *Kruskalo-Voliso statistika* yra

$$F_{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_{i\cdot} - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \bar{R}_{i\cdot}^2 - 3(n+1) =$$

$$= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_{i\cdot}^2}{n_i} - 3(n+1).$$

Proporcingumo koeficientas parenkamas taip, kad esant teisingai hipotezei H_0, statistikos F_{KW} skirstinys kai imčių didumai auga, artėtų į chi kvadrato skirstinį.
Statistika F_{KW} turi tendenciją igyti didesnes reikšmes, kai teisinga alternatyva.

Kruskalo-Voliso kriterijus: hipotezė H_0 atmetama reikšmingumo lygmens α kriterijumi, kai $F_{KW} > F_{KW}(\alpha)$; čia $F_{KW}(\alpha)$ yra minimalus skaičius c tenkinantis nelygybę $\mathbf{P}\{F_{KW} > c | H_0\} \le \alpha$.

Jeigu visos imtys yra didelės, tai statistikos F_{KW} skirstinys aproksimuojamas chi-kvadrato skirstiniu su $k-1$ laisvės laipsniu.

Asimptotinis Kruskalo-Voliso kriterijus: jeigu n_i nėra maži, tai hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai $F_{KW} > \chi_\alpha^2(k-1)$.
Jei duomenys apvalinami, tai galimos vienodos kai kurių stebėjimų reikšmės jungtinėje imtyje netgi ir tuo atveju, kai skirstiniai absoliučiai tolydūs. Jei yra sutampančių reikšmių, tai statistika F_{KW} modifikuojama analogiškai Vilkoksono statistikai:

$$F_{KW}^* = F_{KW} / (1 - T/(n^3 - n));$$

čia $T = \sum_{i=1}^{s} T_i, \quad T_i = (t_i^3 - t_i)$, s yra sutampančių narių grupių skaičius jungtinėje imtyje; t_i yra i-osios grupės didumas.

Modifikuotasis asimptotinis Kruskalo-Voliso kriterijus: hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai $F_{KW}^* > \chi_\alpha^2(k-1)$.

Duomenų įvedimas: "Trukmės įvedamos į vieną stulpelį, atitinkami gamyklos numeriai - į antrą stulpelį.

Skaičiavimas:

Analyze - Nonparametric Tests - K Independent Samples

Test variable "Trukmė", Grouping factor: "Gamykla". Pažymime "Kruskal - Wallis H" "Define Range" ir įvedame minimalią (1) ir maksimalią (3) grupuojančio faktoriaus reikšmę. "Continue", OK.

Rezultatai

Lentelėje "Ranks" pateikiamos visų trijų imčių dydžiai ir vidutinės rangų sumos apjungtoje imtyje. Pagrindinėje lentelėje "Test statistics" pateikiama statistikos F^*_{KW} reikšmė $(6,549)$, laisvės laipsniai $k-1=2$ ir p-reikšmė $Asymp.sig = 0,038$. Hipotezė apie vidutinių funkionavimo trukmių lygybę atmetama, bet nelabai reikšmingai.

(e) *Kelių priklausomų imčių palyginimas: Frydmano kriterijus*

Formulės

Tarkime, kad stebima n nepriklausomų vektorių

$$(X_{11},...,X_{1k})^T,\ldots,(X_{n1},...,X_{nk})^T,\quad k>2.$$

Duomenis galima interpretuoti ir kaip k priklausomų (ar nepriklausomų, jei pradinių vektorių koordinatės nepriklausomos) imčių

$$(X_{11},...,X_{n1})^T,\ldots,(X_{1k},...,X_{nk})^T.$$

Visus turimus stebėjimus galime surašyti į tokią matricą:

$$\mathbf{X}=\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k}\\ X_{21} & X_{22} & \cdots & X_{2k}\\ \cdots & \cdots & \cdots & \cdots\\ X_{n1} & X_{n2} & \cdots & X_{nk}\end{pmatrix}$$

Pažymėkime $F_{i1},\ldots,F_{ik}$ a. d. $X_{i1},...,X_{ik}$ pasikirstymo funkcijas.

k priklausomų imčių homogeniškumo hipotezė:

$$H_0: F_{i1}(x)=...=F_{ik}(x),\quad \forall\, x\in\mathbf{R},\quad i=1,...,n.$$

Homogeniškumo hipotezės H_0 alternatyva H_1 turi tokį pavidalą:

$$H_1:\ F_{ij}(x)=F_{i1}(x-\theta_{ij}),\quad i=1,\ldots,n;\quad j=2,\ldots,k;\quad \sum_{i=1}^{n}\sum_{j=2}^{k}\theta_{ij}^2>0.$$

t. y. marginaliosios pasiskirstymo funkcijos skiriasi tik poslinkio parametrais.

Frydmano kriterijaus statistikos konstravimas. Su kiekvienu fiksuotu i randame a. v. $(X_{i1},...,X_{ik})^T$ rangų vektorių $(R_{i1},...,R_{ik})^T$. Tada vietoje pradinių duomenų matricos $\mathbf{X}$ gauname rangų matricą

$$\boldsymbol{R}=\begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1k}\\ R_{21} & R_{22} & \cdots & R_{2k}\\ \cdots & \cdots & \cdots & \cdots\\ R_{n1} & R_{n2} & \cdots & R_{nk}\end{pmatrix}$$

Rangų suma kiekvienoje eilutėje ta pati:

$$R_{i\cdot}=R_{i1}+...+R_{ik}=k(k+1)/2,\quad i=1,2....,n.$$

Pažymėkime

$$\bar{R}_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n}R_{ij}$$

j-ojo stulpelio rangų aritmetinį vidurkį. Visų rangų aritmetinis vidurkis yra

$$\bar{R}_{\cdot\cdot} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} R_{ij} = \frac{1}{nk} n\frac{k(k+1)}{2} = \frac{k+1}{2}.$$

Esant teisingai hipotezei H_0 atsitiktiniai dydžiai $\bar{R}_{\cdot 1}, ..., \bar{R}_{\cdot k}$ yra vienodai pasiskirstę ir kiekvieno vidurkis yra

$$\mathbf{E}\bar{R}_{\cdot j} = \mathbf{E}R_{1j} = \frac{k+1}{2} = \mathbf{VarR}_{\cdot\cdot}.$$

Frydmano kriterijaus statistika grindžiama skirtumais $\bar{R}_{\cdot j} - \bar{R}_{\cdot\cdot} = \bar{R}_{\cdot j} - (k+1)/2$:

$$S_F = \frac{12n}{k(k+1)}\sum_{j=1}^{k}(\bar{R}_{\cdot j} - \frac{k+1}{2})^2 = \frac{12n}{k(k+1)}\sum_{j=1}^{k}\bar{R}_{\cdot j}^2 - 3n(k+1) =$$

$$= \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_{\cdot j}^2 - 3n(k+1); \tag{0.19}$$

čia $R_{\cdot j} = \sum_{i=1}^{n} R_{ij}$. Normuojantis daugiklis $12n/k(k+1)$ parenkamas taip, kad esant teisingai hipotezei H_0, asimptotiškai (kai $n \to \infty$) statistikos S_F skirstinys artėtų prie chi-kvadrato skirstinio.

Frydmano kriterijus: hipotezė H_0 atmetama kriterijumi su reikšmingumo lygmeniu α, kai $S_F \geq S_{F,\alpha}$; čia $S_{F,\alpha}$ yra mažiausia S_F reikšmė c tenkinanti nelygybę $\mathbf{P}\{S_F \geq c|H_0\} \leq \alpha$.

P-reikšmė yra $pv = \mathbf{P}\{S_F \geq s\}$; čia s yra gautoji statistikos S_F realizacija.

Kai hipotezė H_0 teisinga, tai

$$S_F \xrightarrow{d} S \sim \chi^2(k-1), \quad n \to \infty.$$

Asimptotinis Frydmano kriterijus: jei n yra didelis, tai hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai $S_F > \chi^2_\alpha(k-1)$.

Tada P-reikšmės aproksimacija $pv_a = 1 - F_{\chi^2_{k-1}}(s)$; kur s yra stebėtoji statistikos S_F reikšmė.

Jeigu yra sutampančių stebėjimų, tai Frydmano statistika modifikuojama. Pažymėkime

$$S_F^* = \frac{S_F}{1 - \sum_{i=1}^{n} T_i/(n(k^3-k))};$$

čia

$$T_i = \sum_{j=1}^{k_i}(t_{ij}^3 - t_{ij}),$$

k_i yra sutampančių reikšmių grupių skaičius i-ajam objektui (t. y. i-ojoje matricos $\boldsymbol{R}$ eilutėje), t_{ij} yra j-osios grupės elementų skaičius. Kai hipotezė H_0 teisinga, tai $S_F^* \xrightarrow{d} S \sim \chi^2(k-1), \quad n \to \infty$.

Modifikuotasis asimptotinis Frydmano kriterijus: jei n yra didelis, tai hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai $S_F^* > \chi^2_\alpha(k-1)$.

Duomenų įvedimas.

Tas pats ekspertas vertina trijų rūšių alų, taigi jo įvertiniai susiję. Taigi turime tris priklausomas dydžio 5 imtis. Jas suvedame į tris stulpelius (A, B ir C).

Skaičiavimas.

Analyze - Nonparametric Tests - K Related Samples

Naujesnėse versijose

Analyze - Nonparametric Tests - Legacy Dialogs - K Related Samples

Test variables: A,B ir C.

Pažymėkite "Friedman" kaip reikalingą kriterijų. OK.

Options (nebūtina): Descriptive and Quartiles within the Statistics area.

Continue. OK.

Rezultatai

Lentelėje "Descriptive statistics" pateikiamos kiekvienos iš trijų imčių empirines charakteristikos.

Lentelėje "Ranks" pateikiami visoms trims alaus rūšims vidutiniai rangai $(R_{\cdot 1}, R_{\cdot 2}, R_{\cdot 3}) = (2,10; 1,60; 2,30)$.

Lentelėje "Test statistics" pateikiama priklausomų imčių dydžiai $n = 5$, chi kvadrato statistikos reikšmė $(S_F = 1,368)$, statistikos laisvės laipsnių skaičius $(k - 1 = 2)$, p-reikšmė $(pv = 0,582)$, asimptotinė p-reikšmė $(pv_a = 0,504)$. Duomenys neprieštarauja hipotezei.

(f) Skaičiavimas toks pat. Tik reikia pastebėti, jog šiuo atveju imtys nepriklausomos, bet nėra paprastosios, nes skirtingų spausdintuvų kainos turi skirtingus skirstinius. Taigi čia negalima taikyti Kruskalo-Voliso kriterijaus.

pats.

(g) *Proporcijų palyginimas. kai imtys priklausomos: Maknemaros kriterijus*

Formulės.

Tarkime marginalieji a. v. $(X_i, Y_i)^T$, $i = 1, \ldots, n$, skirstiniai yra *Bernulio*, t. .y.

$$X_i \sim B(1, p_{i1}), \quad Y_i \sim B(1, p_{i2}), \quad p_{i1} = \mathbf{P}\{X_i = 1\}, \quad p_{i2} = \mathbf{P}\{Y_i = 1\};$$

čia a. d. X_i ir Y_i įgyja reikšmę 1, kai tam tikras įvykis A įvyksta, ir reikšmę 0 priešingu atveju. Šio konkretaus uždavinio atveju $X_{i1} = 1$ $(X_{i1} = 0)$, jei i-asis apklaustasis prieš debatus atsakė, kad balsuos už (prieš) kandidatą N, o $Y_{i1} = 1$ $(Y_{i1} = 0)$, jei po debatų jis atsakė, kad balsuos už (prieš) kandidatą N.

Homogeniškumo hipotezė:

$$H_0 : \; p_{i1} = p_{i2} \quad \text{su visais} \quad i = 1, \ldots, n.$$

Nagrinėsime alternatyvą

$$H_3 = H_1 \cup H_2;$$

čia

$$H_1 : \; p_{i1} \le p_{i2} \quad \text{su visais} \quad i = 1, \ldots, n, \text{egzistuoja} \quad i_0, \text{kad} \quad p_{i_0 1} < p_{i_0 2},$$

$$H_2 : \; p_{i1} \ge p_{i2} \quad \text{su visais} \quad i = 1, \ldots, n, \text{egzistuoja} \quad i_0, \text{kad} \quad p_{i_0 1} > p_{i_0 2}.$$

Hipotezė H_0 ekvivalenti tvirtinimui

$$\mathbf{P}\{X_i = 1, Y_i = 0\} = \mathbf{P}\{X_i = 0, Y_i = 1\} \quad \text{su visais} \quad i = 1, \ldots, n,$$

nes

$$\mathbf{P}\{X_i = 1\} = \mathbf{P}\{X_i = 1, Y_i = 1\} + \mathbf{P}\{X_i = 1, Y_i = 0\},$$

$$\mathbf{P}\{Y_i = 1\} = \mathbf{P}\{X_i = 1, Y_i = 1\} + \mathbf{P}\{X_i = 0, Y_i = 1\},$$

o $\mathbf{P}\{X_i = 1\} = \mathbf{P}\{Y_i = 1\}$. Alternatyvos H_1 ir H_2 ekvivalenčios analogiškam tvirtinimui, pakeičiant lygybę atitinkamomis nelygybėmis.

Remiantis teorema pakanka nagrinėti tik tuos objektus, kuriems pirmojo ir antrojo bandymo rezultatai yra skirtingi, t. y. įvyko įvykis

$$\{X_i = 1, Y_i = 0\} \cup \{X_i = 0, Y_i = 1\}.$$

Esant teisingai hipotezei

$$\mathbf{P}\{X_i = 1, Y_i = 0 | \{X_i = 1, Y_i = 0\} \cup \{X_i = 0, Y_i = 1\}\} =$$

$$\mathbf{P}\{X_i = 0, Y_i = 1 | \{X_i = 1, Y_i = 0\} \cup \{X_i = 0, Y_i = 1\}\} = 0,5.$$

Todėl kriterijus konstruojamas tokiu būdu: pažymėkime U_{kl} skaičių tokių objektų, kuriems

$$(X_i, Y_i) = (k, l),\ k,\ l = 0,\ 1; \quad U_{00} + U_{01} + U_{10} + U_{11} = n.$$

Stebėjimo rezultatus galime surašyti į 5.3.1 lentelę.

k/l	0	1	
0	U_{00}	U_{01}	$U_{0.}$
1	U_{10}	U_{11}	$U_{1.}$
	$U_{.0}$	$U_{.1}$	n

Sudarydami kriterijų naudojame tik $m = U_{10} + U_{01}$ objektų stebėjimus.
Esant teisingai hipotezei H_0 sąlyginis statistikos U_{10} skirstinys, kai $m = U_{10}+U_{01}$ fiksuotas, yra binominis $B(m,\ 1/2)$.

Maknemaros kriterijus: homogeniškumo hipotezė, kai alternatyva dvipusė, yra atmetama ne didesnio už α reikšmingumo lygmens kriterijumi, kai

$$U_{10} \le c_1 \quad \text{arba} \quad U_{10} \ge c_2, \tag{0.20}$$

čia c_1 yra didžiausias sveikas skaičius tenkinantis nelygybę

$$\mathbf{P}\{U_{10} \le c_1\} = \sum_{k=0}^{c_1} C_m^k (1/2)^m = 1 - I_{0,5}(c_1 + 1, m - c_1) = I_{0,5}(m - c_1, c_1 + 1) \le \alpha/2,$$

o c_2 minimalus sveikas skaičius tenkinantis nelygybę

$$\mathbf{P}\{U_{10} \ge c_2\} = \sum_{k=c_2}^{m} C_m^k (1/2)^m = I_{0,5}(c_2,\ m - c_2 + 1) \le \alpha/2.$$

Jei u yra stebėtoji statistikos U_{10} reikšmė, tai P-reikšmė (žr. 1.4 skyrelį) yra

$$pv = 2\min(F_{U_{10}}(u), 1 - F_{U_{10}}(u)).$$

Įrodoma, kad

$$Q_2 = \frac{(U_{10} - U_{01})^2}{U_{10} + U_{01}} \xrightarrow{d} \chi^2(1), \quad \text{kai} \quad m \to \infty.$$

Asimptotinis Maknemaros kriterijus: jei m yra didelis,tai hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$Q_2 > \chi^2_\alpha(1). \tag{0.21}$$

Vidutinėms m reikšmėms chi kvadrato skirstiniu geriau aproksimuojama modifikuota statistika, gaunama atsižvelgiant į tolydumo pataisą:

$$Q_2^* = \frac{(|U_{10} - U_{01}| - 1)^2}{U_{10} + U_{01}}.$$

Asimptotinis Maknemaros kriterijus su tolydumo pataisa: jei m yra didelis, tai hipotezė H_0 atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai $Q_2^* > \chi^2_\alpha(1)$.
Duomenų įvedimas

Pirmame stulpelyje "Dazniai" įvedame dažnius, antrame stulpelyje "Pries" įvedame 1, jei prieš debatus atsakymas buvo "taip", įvedame 2, jei prieš debatus atskymas buvo "ne". Trečiame stulpelyje "Po" įvedame 1 ar 2 priklausomai nuo atsakymo po debatų.

Skaičiavimas

Data - Weight cases

Pažymime "Weight cases by" ir įmetame "Dažniai".

Analyze - Descriptive statistics - Crosstabs

"Pries" ir "Po" įmeskite atitinkamai į "Rows" ir "Columns".

"Statistics" - pažymėkite "McNemar" ir "Chi square". Continue, OK.

Rezultatai.

Lentelė "Pries*Po crosstabulation" sutampa su sąlygoje pateikta lentele. Lentelėje "Chi square tests" pirmos eilutėje pateikta statistikos Q_2 reikšmė $379,140$ ir p-reikšmė 0,000, o antroje eilutėje tas pats, kai naudojama statistika Q_2^*. Abi statistikos atmeta hipotezę, kad debatai nepakeitė rinkėjų nuomonės. Taigi duomenys neprieštarauja tam, kad rinkėjai pakeitė nuomonę.

(h) *Serijų kriterijus*

Formulės

Serija vadinama vieno tipo įvykių seka, prieš kurią ir po kurios įvyksta kitokio tipo arba joks įvykis.

Nagrinėsime ilgio $N = m + n$ seką, sudarytą iš m įvykių A ir n jam priešingų įvykių $\bar{A}$. Skirtingų tokio tipo sekų skaičius yra $C_N^m = C_N^n$.

Žymėsime V serijų skaičių minėtoje sekoje. Pavyzdžiui, sekoje

$$A\,A\,\bar{A}\,A\,A\,\bar{A}\,\bar{A}\,A\,\bar{A}$$

yra $V = 6$ serijos; $m = 5$, $n = 4$, $N = 9$.

Dviejų įvykių atsitiktinio išsidėstymo hipotezė: jeigu m ir n fiksuoti, tai kiekvienos iš galimų C_N^m sekos pasirodymas yra vienodai galimas.

Kai teisinga įvykių atsitiktinio išsidėstymo hipotezė, tai serijų skaičiaus skirstinys turi tokį pavidalą:

$$\mathbf{P}\{V = 2i\} = \frac{2C_{m-1}^{i-1}C_{n-1}^{i-1}}{C_N^m}, \quad i = 1, ..., \min(m, n),$$

$$\mathbf{P}\{V = 2i+1\} = \frac{C_{m-1}^{i-1}C_{n-1}^{i} + C_{m-1}^{i}C_{n-1}^{i-1}}{C_N^m}, \quad i = 1, ..., \min(m, n).$$

Serijų skaičiaus vidurkis ir dispersija yra

$$\mathbf{E}V = \frac{2mn}{N} + 1 =, \quad DV = \frac{2mn(2mn - N)}{N^2(N-1)}.$$

Kai $m, n \to \infty$, $m/n \to p \in (0,\, 1)$,

$$Z_{m,n} = \frac{V - \mathbf{E}V}{\sqrt{D}} \xrightarrow{d} Z \sim N(0,\, 1).$$

Serijų skaičiaus statistiką naudosime suformuluotos atsitiktinio įvykių išsidėstymo hipotezei tikrinti.

Tarkime, kad hipotezė neteisinga ir įvykiai A ir B išsidėsto nearsitiktinai. Apie tai liudytų, pavyzdžiui tokio tipo sekų $AAAAAAABBBBBBB$, $ABBBBBBAAAAAAB$, kuriose sekų skaičius mažas pasirodymas. Iš kitos pusės tokio tipo sekos $ABABABABABABAB$ pasirodymas irgi liudija, kad įvykiai kaitaliojasi determinuota tvarka, o ne išsidėsto atsitiktinai.

Taigi hipotezę dėl atsitiktinio įvykių išsidėstymo reikėtų atmesti, kai serijų skaičius yra per daug didelis arba per daug mažas.

Serijų kriterijus: hipotezė atmetama ne didesnio reikšmingumo lygmens kaip α kriterijumi, kai $V \leq c_1$ arba $V \geq c_2$; čia c_1 yra maksimalus svekasis skaičius tenkinantis nelygybę $\mathbf{P}\{V \leq c_1 | H_0\} \leq \alpha/2$, ir c_2 yra minimalus sveikasis skaičius tenkinantis nelygybę $\mathbf{P}\{V \geq c_2 | H_0\} \leq \alpha/2$.
Kai m ir n yra dideli, kriterijų konstruojame remdamiesi normaliąja aproksimacija.

Asimptotinis serijų kriterijus: atsitiktinio įvykių išsidėstymo hipotezė atmetama asimptotiniu reikšmingumo lygmens α serijų kriterijumi, kai $|Z_{k_1,k_2}| \geq z_{\alpha/2}$.

Jeigu k nėra labai didelis rekomenduojama naudoti tolydumo pataisą.

Asimptotinis serijų kriterijus su tolydumo pataisa: hipotezė atmetama asimptotiniu reikšmingumo lygmens α kriterijumi, kai

$$|Z^*_{k_1,k_2}| = |\frac{|V - \mathbf{E}V| - 0,5}{\sqrt{\boldsymbol{V}\boldsymbol{V}}}| \geq z_{\alpha/2}.$$

Duomenų įvedimas: Viename stulpelyje įvedame 1 (benzinas A) ir 2 (benzinas B).
Skaičiavimas:
Analyze - Nonparametric tests - Runs
Test variable: Benzinas
Cut point: Mean
Lentelėje "Runs test" pateikia: A ir B skaičių ($m = n = 25$), serijų skaičių ($V = 31$), asimptotinės statistikos reikšmę ($Z = 1,429$), tikslią ($pv = 0,197$) ir asimptotinę ($pv_a = 0,153$) p-reikšmes. Duomenys neprieštarauja hipotezei, kad paros metas neturi įtakos benzino rūšies pasirinkimui.

9. Dispersinė analizė

(a) *Vienfaktorė dispersinė analizė: kelių normaliųjų vidurkių palyginimas*

Tarkime, kad turime k nepriklausomų normaliųjų imčių; i-ją imtį sudaro n_i stebėjimų

$$Y_{i1}, ..., Y_{in_i}, \quad Y_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, ..., k.$$

Bendrą stebėjimų skaičių pažymėsime $n = n_1 + \cdots + n_k$.

Stebėjimus $Y_{ij} \sim N(\mu_i, \sigma^2)$ galima užrašyti tokiu pavidalu:

$$Y_{ij} = \mu_i + e_{ij} = \mu + \alpha_i + e_{ij},$$

čia

$$\mu = \frac{1}{n}\sum_{i=1}^{k} n_i \mu_i, \quad \alpha_i = \mu_i - \mu;, \quad i = 1, ..., k, \quad j = 1, ..., n_i;$$

čia $e_{ij} \sim N(0, \sigma^2)$ – nepriklausomi a. d., μ yra bendrasis vidurkis, gautas iš visų populiacijų, o α_i yra i-osios populiacijos vidurkio ir bendrojo vidurkio skirtumas, dažnai vadinamas i-osios populiacijos *efektu*.
Vienas iš svarbiausių vienfaktorinės dispersinės analizės uždavinių yra patikrinti vidurkių lygybės hipotezę

$$H : \mu_1 = \cdots = \mu_k.$$

Alternatyva $\bar{H} : \mu_i \neq \mu_j$ bent vienai porai $i \neq j$.
Vidurkių μ_i įvertiniai yra empiriniai vidurkiai

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_i, \quad i = 1, 2, ..., k,$$

Visų duomenų sklaida

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$$

apie bendrą empirinį vidurkį $\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}$ vadinama *pilnąją kvadratų suma* (angl. "total sum of squares") ir gali būti išskaidyta į du dėmenis

$$SS_T = SS_W + SS_B = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$

Kvadratų suma

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2.$$

yra stebėjimų $Y_{i1}, ..., Y_{in_i}$ sklaidų apie savo empirinius vidurkius $\bar{Y}_i$ kiekvienoje imtyje suma ir vadinama *vidine arba paklaidų kvadratų suma* (angl. "within or error sum of squares"). Šios sklaidos priežastis – atsitiktinės paklaidos e_{ij}. Kartais ši sklaida žymima SS_E.

$$SS_B = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$

yra grupių empirinių vidurkių $\bar{Y}_i$ sklaida apie bendrą empirinį vidurkį $\bar{Y}$ ir vadinama *tarpgrupine kvadratų suma* (angl. "between sum of squares").

Jei hipotezė apie vidurkių lygybę yra teisinga, tai sumos SS_B, charakterizuojančios sklaidą tarp grupių, dalis visoje sumoje SS_T turėtų būti maža, tuo pačiu sumos SS_W, charakterizuojančios sklaidą grupių viduje, turėtų būti didelė. Atvirkščiai, kuo daugiau vidurkiai skiriasi, tuo SS_B dalis turėtų būti didesnė. Taigi natūralu atmesti hipotezę, kai santykis SS_B/SS_W didelis.

Atsitiktinis dydis SS_W/σ^2 turi chi kvadrato skirstinį su $n-k$ laisvės laipsnių, o kai teisinga hipotezė H, a.d. SS_B/σ^2 turi chi kvadrato skirstinį su $k-1$ laisvės laipsnių. Be to šie atsitiktiniai dydžiai yra nepriklausomi.

Pažymėkime

$$MS_B = \frac{SS_B}{k-1}, \quad MS_W = \frac{SS_W}{n-k}.$$

Tada

$$F = \frac{MS_B}{MS_B} \sim F(k-1, n-k),$$

t.y. santykis F turi Fišerio skirstinį su $k-1$ ir $n-k$ laisvės laipsnių, jei teisinga hipotezė.

Kriterijus hipotezei H tikrinti. Hipotezė H atmetama reikšmingumo lygmens α kriterijumi, kai

$$F > F_\alpha(k-1, n-k);$$

čia $F_\alpha(k-1, n-k)$ – Fišerio skirstinio α kritinė reikšmė.

Jeigu stebėjimai neprieštarauja prielaidai H apie vidurkių $\mu_1, ..., \mu_k$ lygybę, tai analizę galima tuo ir užbaigti. Tokiu atveju visus stebėjimus galime apjungti į vieną dydžio n imtį, gautą stebint normalųjį a. d. su vidurkiu μ ir dispersija σ^2.

Jei hipotezė atmetama, tai reikia išsiaiškinti kurios grupės atsakingos už hipotezės atmetimą.

Porų lyginimui dažniausiai naudojamas Tjukio kriterijus. Tikrinama ta pati hipotezė H. Lyginant porą (i, j), skaičiuojamas statistika

$$Q(i,j) = \frac{\sqrt{n}(\bar{X}_i - \bar{X}_j)}{\sqrt{MS_W}}.$$

Vidurkiai reikšmingai skiriasi, jei $Q(i,j) > Q_\alpha(nk-k, k)$. Čia α yra reikšmingumo lygmuo, kuris gaunamas tikrinant hipotezę H.

Dar naudojamas Bonferonio kriterijus: vidurkiai reikšmingai skiriasi, jei

$$|\bar{X}_i - \bar{X}_j| > t_{\alpha/(k(k-1))}(N-k)\sqrt{MS_W(1/n_i + 1/n_j)}, \quad N = n_1 + ... + n_k.$$

Duomenų įvedimas

Pirmame stulpelyje "Laikas" (tai ir yra matuojamas dydis) suvedamos surinkimo laiko reišmės. Antrame stulpelyje "Darbininkas" (j-asis darbininkas ir formuoja j-ąją imtį) įvedamas jo numeris (1 greta pirmojo stulpelio reikšmių atitinkančių pirmąjį darbininką, 2 -greta pirmo stulpelio reikšmių atitinkančių antrąjį darbininką, ir t.t.).

Skaičiavimas.

Analyze - Compare Means - One-Way ANOVA

Priklausomą kintamąjį "Laikas" įmeskite į Dependent List langelį, nepriklausomą kintamąjį "Darbininkas" į langelį Factor.

Nuspauskite Post Hoc Button mygtuką. Pažymėkite Tukey.

Nuspauskite Continue ir Options. Pažymėkite Descriptive.

Continue ir OK.

Rezultatai.

"Descriptives" lentelė

Pateikiamos priklausomo kintamojo "Laikas" vidurkiai, vidutiniai kvadratiniai nuokrypiai, 0,95 pasikliovimo intervalai kiekvienoje grupėje (šiuo atveju kiekvieno darbininko surinkimo laikams) ir visų keturių darbininkų surinkimo laikams.

ANOVA lentelė

Pateikiamos SS_B, SS_W, SS_T reikšmės, laisvės laipsniai $k-1$ ir n, vidurkiai MS_B ir MS_W, statistikos $F = 9,928$ reikšmė bei p-reikšmė $sig = 0,000$. Taigi hipotezė apie detalių surinkimo laikų lygybę atmetama. Taigi darbininkai skiriasi pagal detalės surinkimo laiką.

Norint išsiaiškinti kurių darbininkų rezultatai skiriasi naudojame Multiple Comparisons lentele.

"Multiple Comparisons" lentelė

Pateikiami grupių vidurkių skirtumai ir p-reikšmės. 1 grupė reikšmingai nesiskiria nuo 3-ios (pirma homogeniška grupė), 2 grupė reikšmingai nesiskiria nuo 3-ios (antra homogeniška grupė), 2 grupė reikšmingai nesiskiria nuo 4-os (trečia homogeniška grupė).

(b) Tas pats

(c) *Vienfaktorė dispersinė analizė: kontrastai.*

Formulės

Kontrastu vadinama tiesinė funkcija $C_i\mu_1 + ... + C_k\mu_k$, $C_i \in \mathbf{R}$, $C_1 + ... + C_k = 0$. Dažniausiai parenkama $C_i \in \mathbf{Z}$.

Tikrinsime hipotezę $H : C_i\mu_1 + ... + C_k\mu_k = 0$.

Kriterijaus statistika

$$T = \sqrt{n}\frac{C_i\bar{X}_1 + ... + C_k\bar{X}_k}{(MS_W(C_i^2 + ... + C_k^2))^{1/2}}.$$

Hipotezė atmetama su reišmingumo lygmeniu α, jei $|T| > t_{\alpha/2}(n-k)$.

Skaičiavimas

Tas pats, tikti dar prieš OK paspauskite Contrasts. Pažymėkite Polynomial. Šiame uždavinyje tikrinama hipotezė $\mu_3 = (\mu_1 + \mu_2)/2$ ekvivalenti hipotezei $\mu_1 = \mu_2 - \mu_3 = 0$, todėl įvedame Coefficients 1,1-2. Tada Continue, OK.

Lentelėje Contrast coefficients pateikiamos įvestų koeficientų reikšmės (1,1,-2), o lentelėje Contrast tests pateikiama statistikos T reikšmė 1,045 bei p-reikšmė 0,302. Duomenys neprieštarauja hipotezei, kad vidutinis kilikiečių gestikuliavimo lygis yra lygus nubų ir ilyrų vidutinių gestikuliavimo lygių vidurkiui.

(d) *Dvifaktorė dispersinė analizė: hipotezių apie kelių faktorių įtaką bei jų sąveika tikrinimas*

Tarkime a. d. Y skirstinys gali priklausyti nuo faktoriaus A, kurio lygmenys yra $A_1, ..., A_I$, ir nuo faktoriaus B, kurio lygmenys yra $B_1, ..., B_J$, reikšmių. Pavyzdžiui, Y gali reikšti kviečių derlingumą, faktorius A – kviečių veislę, o faktorius B – jų auginimo metodiką.

Tarkime, kad stebėjimai atliekami imant visas galimas skirtingas faktorių lygmenų kombinacijas, be to, kiekvienu atveju matavimai kartojami vienodą skaičių kartų $K > 1$. Toks eksperimentų planas vadinamas *visiškai subalansuotu planu*. Jo analizė žymiai paprastesnė negu tuo atveju, kai stebėjimų skaičiai K_{ij}, gauti esant faktorių lygmenų kombinacijai $(A_i,\ B_j)$, yra skirtingi.

Stebėjimo rezultatus žymėsime $Y_{ijk},\ i = 1, ..., I,\ j = 1, ..., J,\ k = 1, ..., K$; i – faktoriaus A lygmens numeris; j – faktoriaus B lygmens numeris; k – kartotinumo numeris; visų stebėjimų skaičius $n = IJK$.

Dvifaktorės dispersinės analizės modelis: stebėjimai Y_{ijk} yra nepriklausomi a. d., pasiskirstę pagal normalųjį dėsnį $N(\mu_{ij}, \sigma^2)$.

Stebėjimus galima aprašyti taip:

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \quad i = 1, ..., I, \quad j = 1, ..., J, \quad k = 1, ..., K,$$

čia μ_{ij} – nežinomi parametrai, e_{ijk} – n. a. d., pasiskirstę pagal normalųjį dėsnį $N(0,\ \sigma^2)$.

Dispersinės analizės tikslas – ištirti stebimojo a. d. Y priklausomybę nuo faktorių A ir B.

Kad vaizdžiau suformuluotume dispersinės analizės hipotezes, įveskime kitus parametrus.

Pradžioje įvesime parametrą, charakterizuojantį "tiesioginę" faktoriaus A lygmens A_i įtaką, kai "eliminuojama" faktoriaus B įtaka. Stebėjimų, kuriems faktoriaus A reikšmė lygi A_i, ir stebėjimų, kuriems faktoriaus B reikšmė lygi B_j, vidurkių aritmetinius vidurkius žymėsime atitinkamai

$$\bar{\mu}_{i.} = \frac{1}{J}\sum_{j=1}^{J}\mu_{ij}, \quad \bar{\mu}_{.j} = \frac{1}{I}\sum_{i=1}^{I}\mu_{ij},$$

o visų stebėjimų vidurkių aritmetinį vidurkį (bendrąjį vidurkį)

$$\mu = \frac{1}{I}\sum_{i=1}^{I}\mu_{i.} = \frac{1}{J}\sum_{j=1}^{I}\mu_{.j} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}.$$

Faktoriaus A įtaką nuokrypiui nuo bendro vidurkio, kai eliminuojama faktoriaus B įtaką, charakterizuoja skirtumai $\alpha_i = \bar{\mu}_{i.} - \mu$, o Faktoriaus B įtaką nuokrypiui nuo bendro vidurkio, kai eliminuojama faktoriaus A įtaką, charakterizuoja skirtumai $\beta_j = \bar{\mu}_{.j} - \mu$.

Gauname tokį vidurkio μ_{ij} skaidinį į komponentes:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \quad \gamma_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \mu.$$

Naujai įvesti parametrai tenkina tokias papildomas sąlygas: $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$.

Jeigu $\gamma_{ij} = 0$ su visais $i = 1, ..., I$ ir $j = 1, ..., J$, tai galioja lygybės

$$\mu_{ij} = \mu + \alpha_i + \beta_j.$$

Sakome, kad turime *adityvųjį modelį*. Šiame modelyje $\mu_{ij} - \mu_{i'j} = \alpha_i - \alpha_{i'}$, taigi faktoriaus A įtaką vidurkiui nepriklauso nuo faktoriaus B reikšmės. Ir atvirkščiai, faktoriaus B įtaką vidurkiui nepriklauso nuo faktoriaus A reikšmės.

Kai egzistuoja pora i, j, tokia, kad $\gamma_{ij} \neq 0$, tai egzistuoja i': $\gamma_{ij} \neq \gamma_{i'j}$ (nes $\sum_i \gamma_{ij} = 0$), todėl $\mu_{ij} - \mu_{i'j} = \alpha_i - \alpha_{i'} + \gamma_{ij} - \gamma_{i'j}$, taigi faktoriaus A įtaką vidurkiui priklauso nuo faktoriaus B reikšmės. Ir atvirkščiai, faktoriaus B įtaką vidurkiui priklauso nuo faktoriaus A reikšmės. Taigi komponentės γ charakterizuoja faktorių A ir B *sąveiką*.

Parametrų μ_{ij} mažiausiųjų kvadratų įvertiniai $\hat{\mu}_{ij} = \bar{Y}_{ij.}$ randami minimizuojant kvadratinę formą $\sum_i \sum_j \sum_k (Y_{ijk} - \mu_{ij})^2$. Tuo pačiu gauname parametrų $\mu = \bar{\mu}_{..}$, α_i, β_j, γ_{ij} įvertinius

$$\hat{\mu} = \bar{Y}_{...}, \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \hat{\gamma}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...},$$

Dispersijos σ^2 nepaslinktas įvertinys

$$s^2 = MS_E = SS_E/(IJ(K-1)), \quad SS_E = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2.$$

Pagrindinės dvifaktorės dispersinės analizės hipotezės yra:

$$H_A : \alpha_1 = \ldots = \alpha_I, \;\; H_B : \beta_1 = \ldots = \beta_J, \;\; H_{AB} : \gamma_{11} = \ldots = \gamma_{IJ}.$$

Pažymėkime

$$SS_A = JK\sum_i \hat{\alpha}_i^2, \quad SS_B = IK\sum_j \hat{\beta}_j^2, \quad SS_{AB} = K\sum_i \sum_j \hat{\gamma}_{ij}^2.$$

Šių kvadratų sumų skirstiniai nepriklauso nuo kvadratų sumos SS_E skirstinio.

Jeigu teisingos hipotezės H_A, H_B arba H_{AB}, tai atitinkami $SS_A/\sigma^2 \sim \chi^2(I-1)$, $SS_B/\sigma^2 \sim \chi^2(J-1)$ ir $SS_{AB}/\sigma^2 \sim \chi^2((I-1)(J-1))$.

Hipotezės H_A, H_B, H_{AB} atmetamos reikšmingumo lygmens α kriterijais, jeigu atitinkamai tenkinamos nelygybės:

$$F_A = \frac{MS_A}{MS_E} > F_\alpha(I-1, IJ(K-1)), \quad F_B = \frac{MS_B}{MS_E} > F_\alpha(J-1, IJ(K-1)),$$

$$F_{AB} = \frac{MS_{AB}}{MS_E} > F_\alpha((I-1)(J-1, IJ(K-1)));$$

čia

$$MS_A = \frac{SS_A}{I-1}, \quad MS_B = \frac{SS_B}{J-1}, \quad MS_{AB} = \frac{SS_{AB}}{(I-1)(J-1)}.$$

Statistikos F_A, F_B ir F_{AB} turi tendenciją įgyti didesnes reikšmes, kai atitinkamai hipotezės H_A, H_B ir H_{AB} yra neteisingos, negu tuo atveju, kai jos teisingos.

Pažymėkime

$$SS_T = \sum_i \sum_j \sum_k (Y_{ijk}^2 - \bar{Y}_{...})^2.$$

Teisinga lygybė

$$SS_A + SS_B + SS_{AB} + SS_E = SS_T.$$

Skaičiavimo rezultatus galime surašyti į lentelę.

Faktorius	SS	$\nu(laisv.laipsn.)$	$MS = SS/\nu$
A	SS_A	$I-1$	MS_A
B	SS_B	$J-1$	MS_B
$A \times B$	SS_{AB}	$(I-1)(J-1)$	MS_{AB}
E	SS_E	I J(K-1)	MS_E
T	SS_T	I J K -1	-

Duomenų įvedimas

Duomenis suvedame į tris stulpelius: pirmame stulpelyje "Grupė" įvedame 1,2, 3 atitinkamai valstiečims, verslininkams ir inteligentams. "Variable" lange stulpelyje "Values" priskiriame 1,2 ir 3 atitinkamas etiketes "Valstiečiai", "Verslininkai" ir "Inteligentai". Antrame stulpelyje "Kaina" įvedame 1,2 3 atitinkamai mažai, vidutinei ir didelei kainai, o po to priskiriame etiketes. Trečiame stulpelyje "Balas" įvedame balų reikšmes atitinkančias duotas faktorių "Grupė" ir "Kaina" reikšmes.

Skaičiavimas.

Analyze - General Linear Model - Univariate

Atsidaro langas Univariate

Priklausomą kintamąjį "Balas" į langelį Dependent Variable, o abu fktorius "Grupė" ir "Kaina" į langelį Fixed Factor(s).

Pastaba: šiai analizei nereikia kreipti dėmesio į Random Factor(s), Covariate(s) ir WLS Weight langelius.

Nuspauskite "Plots" mygtuką. Pasirodys langas Univariate: Profile Plots.

Imeskite nepriklausomą kintamąjį "Grupė" iš Factors langelio į Horizontal Axis langelį, o kitą nepriklausomą kintamajį "Kainą" į Separate Lines langelį.

Pastaba: į Horizontal Axis langelį patartina įmesti faktorių įgijantį daugiau reikšmių.

Nuspauskite "Add". Tada "Kaina" persikels į Plots langelį.

Nuspaudus "Continue" grįžtame į Univariate dialogue box.

Nuspauskite "post-hoc" mygtuką. Atsidaro Univariate: Post Hoc Multiple Comparisons for Observed Means langas.

Perkelkite "Grupė" ir "Kaina" iš Factor(s) langelio į Post Hoc Tests for langelį. Taip suaktyvuojama Equal Variances Assumed zona. Pasirenkame Tukey, kuris yra vienas geriausių post hoc kriterijų.

Pastaba: jei kuris faktorius įgija tik dvi reikšmes, tai galime jo nekelti į Post Hoc Tests for langelį.

Nuspauskite "Continue" mygtuką ir grįžtam į Univariate dialogue box.

Nuspauskite "Options". Pasirodo Univariate: Options dialogue langas.

Perkelkite "Grupė", "Kaina" ir "Grupė*kaina" iš Factor(s) and Factor Interactions lentelės į Display Means for lentelę. Pažymėkite Descriptive Statistics.

Nuspauskite Continue ir grįšite į Univariate dialogue langą.

OK .

Rezultatai

"Descriptive Statistics" lentelėje kiekvienai faktorių kombinacijai pateikiami vidurkiai ir standartiniai nuokrypiai. Be to pateikiami eilučių "Total", t.y. kiekvienai kainai pateikiami bendras vidurkjis ir vidutinis kvadratinis nuokrypis nepriklausomai nuo grupės.

"Estimated marginal means" lentelėje pateikiami empiriniai vidurkiai ir kvadratiniai nuogrypiai tiek kiekvienam iš faktorių lygmenų tiek ir bet kuriai tų faktorių kombinacijai. Pavyzdžiui, valstiečių balų vidurkis yra $55,000$, mažos kainos paveikslų balų vidurkis yra $51,667$, o valstiečiai mažos kainos paveikslams davė vidutinį balą $51,000$.

"Profile plots" grafikas

Duoda balų vidurkių priklausomai nuo grupės grafiką.

Sąveikos buvimas charakterizuojamas tuo, kad laužtės nėra lygiagrečios.

"Tests of Between-Subjects Effects" lentelėje svarbiausios yra eilutės "Grupė", "Kaina", "Grupė*Kaina" Jose sužinome ar faktoriai "Grupė", "Kaina" ir jų sąveika "Grupė*Kaina" yra reikšmingi. Pateikiamos SS_A, SS_B, SS_{AB}, laisvės laipsnių, MS_A, MS_B, MS_{AB}, statistikų F_A, F_B, F_{AB} reikšmės bei p-reikšmės. Visos p-reikšmes yra labai mažos, tai balas priklauso ir nuo grupės ir nuo nurodytos kainos, be to kainų didėjimas nevienodai paveikė skirtingų profesijų atstovų nuomonę (t.y. yra sąveika tarp faktorių).

"Multiple comparisons" lentelėje" lyginami Valstiečiai, Verslininkai ir Inteligentai poromis. Visų skirtumai statistiškai reikšmingi.

(e) Tas pats.

(f) *Blokuotųjų duomenų vienfaktorė dispersinė analizė*

Formulės

Nagrinėsime faktoriaus B įtaką požymiui Y. Tarkime, kad faktorius B įgija b reikšmių B_1,...,B_b.

Stebima n objektų. i-am objektui atliekama b tiriamo požymio matavimų $Y_{i1}, ..., Y_{ib}$. Stebėjimas Y_{ij} gaunamas kai požymio B reikšmė lygi B_j. Dažniausios situacijos, kai tie patys objektai matuojami skirtingais laiko momentais ir norima įsitikinti ar įvyksta vidurkio pokyčiai (pavyzdžiui, tas pats vairuotojas gali važiuoti su kelių modelių automobiliais ir norima įsitikinti ar vidutinis vairavimo klaidų skaičius priklauso nuo automobilio modelio, pacientui kraujo spaudimas gali būti matuojamas skirtingais laiko momentais (prieš gydymą, po gydymo ir praėjus pusei metų po gydymo ir norima patikrinti ar kraujo spaudimo vidurkis pakito), bet nebūtinai duomenys gaunami tokiu būdu. Pavyzdžiui, duotoje parduotuvėje fiksuojamas skaičius parduotų prekių, laikytų viršutinėje, apatinėje arba vidurinėje lentynose, ir norima patikrinti ar vidutiniai pardavimai skirtingose lentynose skiriasi.

Matavimai $Y_{i1}, ..., Y_{ib}$ yra priklausomi, nes susiję su tuo pačiu objektu, bet atsitiktiniai vektoriai $(Y_{11}, ..., Y_{1b})$,..., $(Y_{n1}, ..., Y_{nb})$ (vadinami *blokais*) yra nepriklausomi, nes atitinka skirtingus objektus, bet gali būti skirtingai pasiskirstę, nes dažniausiai objektai parenkami atsitiktinai iš didelės objektų populiacijos, kuri gali būti labai nevienalytė. Pavyzdžiui, parduotuvės gali būti nevienodo dydžio ir populiarumo ir jose pardavimai tiek apatinėje, tiek vidurinėje, tiek ir viršutinėje lentynoje gali būti labai skirtingi. Taigi egzistuoja dar ir "" trukdantysis"" faktorius - blokas.

Vienfaktorėje blokuotųjų duomenų analizėje tiriama vieno faktoriaus (B) įtaka tiriamam požymiui, taigi tariama, kad nėra sąveikos tarp faktorių A ir B. Tuo atveju modelį galima užrašyti pavidalu

$$Y_{ij} = \mu_j + a_i + \varepsilon_{ij} = \mu + a_i + \beta_j + \varepsilon_{ij}, \quad i = 1, ..., n; j = 1, ..., b;$$

čia $a_i \sim N(0, \sigma_A^2)$, $\sum_{j=1}^{b} \beta_j = 0$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, $a_i$ir ε_{ij} nepriklausomi a.d.

Taigi blokas įneša papildomą duomenų išsibarstymą σ_A^2 į dispersiją $DY_{ij} = \sigma_A^2 + \sigma^2$, taigi

$$Y_{ij} \sim N(0, \sigma_A^2 + \sigma^2).$$

Pastebėkime, kad jei $j \neq j'$, tai a.d. Y_{ij} ir $Y_{ij'}$ iš tikrųjų yra priklausomi: $\mathbf{Cov}\,(Y_{ij}, Y_{ij'}) = \sigma_A^2$. Taigi patenkinta *sferiškumo sąlyga*:

$$D(Y_{ij} - Y_{ij'}) = 2\sigma^2 = const.$$

Pagrindinis vienfaktorės blokuotųjų duomenų analizės uždavinys yra patikrinti hipotezę

$$H_W : \beta_1 = ... = \beta_b = 0 \quad \sim \quad H_W : \mu_1 = ... = \mu_b$$

apie faktoriaus B įtakos nebuvimą priklausomo kintamojo vidurkiui.

Pažymėkime

$$SS_T = \sum_{i=1}^{n} \sum_{j=1}^{b} (\bar{Y}_{ij} - \bar{Y})^2$$

pilnąją kvadratų sumą,

$$SS_W = n\sum_{j=1}^{b}(\bar{Y}_{\cdot j} - \bar{Y})^2$$

vidinę kvadratų sumą,

$$SS_B = b\sum_{i=1}^{n}(\bar{Y}_{i\cdot} - \bar{Y})^2$$

tarpgrupinę kvadratų sumą, ir $SS_E = SS_T - SS_W - SS_B$ *paklaidos kvadratų sumą.*
Žymėsime

$$MS_W = \frac{SS_W}{b-1}, \quad MS_B = \frac{SS_B}{n-1}, \quad MS_E = \frac{SS_E}{(b-1)(n-1)}.$$

Jei teisinga hipotezė H_W, tai a.d. MS_W turėtų įgyti nedideles reikšmes, nes tada $\bar{Y}_{\cdot j}$ reikšmės artimos $\bar{Y}$ reikšmėms.
Hipotezė H_W atmetama su reikšmingumo lygmeniu α, jei

$$F_W = \frac{MS_W}{MS_E} > F_\alpha(b-1,(b-1)(n-1)).$$

Aišku, kad šis kriterijus naudojamas tiktai tada, kai modelis teisingas, taigi patenkinta sferiškumo sąlyga.
Galima patikrinti ir hipotezę $H_B : \sigma_A^2 = 0$, kuri reiškia, kad blokai neįneša papildomo išsibarstymo. Ši hipotezė nėra svarbi atliekant vienfaktorę dispersinę analizę, kurios pagrindinis tikslas yra hipotezės H_W patikrinimas. Hipotezė $H_B : \sigma_A^2 = 0$ atmetama su reikšmingumo lygmeniu α, jei

$$F_B = \frac{MS_B}{MS_E} > F_\alpha(n-1,(b-1)(n-1)).$$

Jei nepatenkinta sferiškumo prielaida, t.y. yra sąveika tarp faktoriaus B ir bloko, tai Grynhauzas ir Geiseris (Greenhous-Geiser) pasiūlė aproksimuoti statistikos F_W skirstinį Fišerio skirstiniu ne su $b-1$ ir $(b-1)(n-1)$, o su $(b-1)\hat{\varepsilon}$ ir $(b-1)(n-1)\hat{\varepsilon}$ laisvės laipsnių; čia $\hat{\varepsilon} \in (0,1)$ yra parenkamas tokiu būdu: pažymėkime

$$\tilde{S}_{jj'} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{\cdot j})(Y_{ij'} - \bar{Y}_{\cdot j'}), \quad \tilde{S}_{j\cdot} = \frac{1}{b}\sum_{j'=1}^{b}\tilde{S}_{jj'}, \quad S_{\cdot\cdot} = \frac{1}{b}\sum_{j=1}^{b}\tilde{S}_{j\cdot},$$

$$S_{jj'} = \tilde{S}_{jj'} - \tilde{S}_{j\cdot} - \tilde{S}_{j'\cdot} + \tilde{S}_{\cdot\cdot}.$$

Tada

$$\hat{\varepsilon} = \frac{\sum_{j=1}^{b} S_{jj}}{(b-1)\sum_{j=1}^{b}\sum_{j'=1}^{b} S_{jj'}^2} = \frac{\sum_{j=1}^{b} S_{jj}}{(b-1)\sum_{j=1}^{b-1}\lambda_j^2},$$

čia $\lambda_1, ..., \lambda_{b-1}$ yra matricos $\boldsymbol{S} = [S_{jj'}]$ tikrinės reikšmės (viena tikrinė reikšmė yra 0).
Grynhauzo-Geiserio kriterijus: hipotezė H_W atmetama su reikšmingumo lygmeniu α, jei

$$F_W = \frac{MS_W}{MS_E} > F_\alpha((b-1)\hat{\varepsilon},(b-1)(n-1)\hat{\varepsilon}).$$

Naudojama ir kita aproksimacija: Hjuinas-Feldtas apibrėžė dydį

$$\tilde{\varepsilon} = \frac{n(b-1)\hat{\varepsilon} - 2}{(b-1)[n1 - (b-1)\hat{\varepsilon}]}.$$

Hjuino-Feldto kriterijus: hipotezė H_W atmetama su reikšmingumo lygmeniu α, jei

$$F_W = \frac{MS_W}{MS_E} > F_\alpha((b-1)\tilde{\varepsilon},(b-1)(n-1)\tilde{\varepsilon}).$$

Močlis pasiūlė kriterijų sferiškumo hipotezei tikrinti. Jis apibrėžė statistiką

$$W = \frac{\prod_{j=1}^{b=1} \lambda_j}{(\frac{1}{b-1}\sum_{j=1}^{b-1}\lambda_j)^{b-1}}.$$

Ši statistika įgyja reikšmes intervale $(0, 1)$ ir jos reikšmės artimos 1, kai teisinga sferiškumo hipotezė. Sferiškumo hipotezė atmetama su apytiksliu reikšmingumo lygmeniu α, kai statistikos W reikšmė yra mažesnė už jos $1 - \alpha$ kritinę reikšmę,

Duomenų įvedimas

Pirmame stylpelyje "Parduotuve" įveskite parduotuvės numerius nuo 1 iki 7. Į kitus tris stulpelius "Virsutine", "Vidurine", "Apatine" suveskite parduotų valiklių skaičius atitinkamai viršutinėje, vidurinėje ir apatinėje lentynose.

Skaičiavimas

Analyze-General linear-Repeated measures

Factor 1: pakeičiame į "Lentyna" (kaip factor 1 paprastai įvedamas faktoriaus, pagal kurio reikšmes lyginame priklausomo kintamojo vidurkius, pavadinimas, šiuo atveju "Lentyna").

Number of levels - 3 (lentyna gali būti viršutinė, vidurinė ir apatinė).

Add-Define

Pasirodo langas

"Virsutine" į langelį *within subject variable*

"Vidurine" į langelį *within subject variable*

"Apatine" į langelį *within subject variable*

Plots

Įmeskite "Lentyna" į "Horizontal Axis" langelį.

Add, Continue.

Options

Atsidaro langas "Repeated Measures: Options"

Perkelkite "Lentyna" iš Factor(s) and Factor Interactions langelio į Display Means for langelį.

Pažymėkite "Compare main effects" ir pasirinkite "Bonferroni" langelyje "Confidence interval adjustment".

Pažymėkite "Descriptive statistics" *Display* zonoje.

Continue, OK.

Rezultatai

Within-Subjects Factors lentelėje primenama kokius lygmenis turi nepriklausomas kintamasis "Lentyna".

"Descriptive Statistics" lentelėje pateiktos pardavimų skirtingose lentynose empirinės charakteristikos (vidurkis, standartinis nuokrypis, objektų (šiuo atveju parduotuvių, skaičiaus reikšmės).

Lentelėje "Mauchly's test of sphericity" atsakoma į klausimą ar ptenkinta sferiškumo sąlyga. Duota Močlio statistikos W reikšmė 0,284 ir p-reikšmė 0,043. Nors ir nelabai reikšmingai, bet sferiškumo hipotezė atmetama. Lentelėje pateiktos ir $\hat{\varepsilon}$ bei $\tilde{\varepsilon}$ reikšmės Greenhouse-Geisser bei Huynh-Feldt statistikų laisvės laipsnių skaičiaus nustatymui.

"Tests of Within-Subjects Effects" lentelėje atsakoma į klausimą ar skiriasi pardavimų vidurkiai viršutinėje, vidurinėje ir apatinėje lentynose, tiksliau tikrinama hipotezė apie pardavimų lygybę. Statistikos F reikšmė lygi 10,751. Kadangi sferiškumo sąlyga neišpildyta, tai žiūrime į Greenhouse-Geisser kriterijaus p-reikšmę 0,012. Hipotezė atmetama, taigi pardavimai bent dviejose iš apatinės, vidurinės ir apatinės lentynų skiriasi.

"Pairwise Comparisons" lentelėje atsakoma į klausimą kurios lentynos vietos skiriasi pagal pardavimus. Gauta, kad viršutinė ir vidurinė lentynos reikšmingai nesiskiria pagal

pardavimus (sig=0,183), viršutinė gana reikšmingai skiriasi nuo apatinės (sig=0,017), o vidurinė reikšmingai skiriasi nuo apatinės, bet p-reikšmė 0,041 rodo, kad tas reikšmingumas nėra stiprus.

Diagramoje "Estimated marginal Means" vaizdžiai pavaizduotas pardavimų vidurkio kitimas priklausomai nuo lentynos vietos.

(g) Tas pats.

(h) *Blokuotųjų duomenų dvifaktorė dispersinė analizė*

Duomenų įvedimas

Stulpeliuose Kalba, Politologija, Matematika, Logika įvedame atitinkamų dalykų pažymius. Stulpelyje Lytis nurodome lytį (1, jei vyras, 2, jei moteris)

Skaičiavimas

Analyze - General Linear Model - Repeated Measures

Pasirodo Repeated Measures Define Factor(s) langas.

Within-Subject Factor Name langelyje pakeiskite "factor1" į "Egzaminas".

Number of Levels langelyje įrašykite 4.

Add-Define.

Kalba, Politologija, Matematika, Logika į langelį *within subject variable*

Lytis į langelį *between subjects factors*

Plots

Egzaminas į *Horizontal Axis*

Lytis į *Separate lines* (jei norime vienoje koordinačių sistemoje dviejų grafikų) arba į *Separate Plots* (du atskiri grafikai).

Add-Continue.

Options

Lytis, Egzaminas ir Lytis*Egzaminas perkelkite į *Display means for* langelį.

Pažymėkite "Compare means effect" ir pažymėkite "Bonferroni" langelyje "Cnfidence interval adjustement". Display zonoje pažymėkite Descriptive statistics.

Continue, OK.

Rezultatai

Lentelėje *within subjects factors* primenama, į kokias kategorijas buvo suskirstyti egzaminai (Kalba, politologija, matematika, logika).

Lentelėje *Between subjects factors* primenama, kokios lytys egzistuoja (vyrai ir moterys).

Lentelėje *Descriptive statistics* pateikti pažymių vidurkiai kiekvienai faktorių reišmių porai (vyrų - kalbos, moterų - kalbos, vyrų - politologijos, moterų - politologijos ir t.t.) bei kiekvienam egzaminui apjungiant lytis (kalbos, politologijos ir t.t).

Lentelėje *Mauchly's test of sphericity* pateikta Močlio statistikos W reikšmė 0,552 ir p-reikšmė 0,396. Taigi duomenys neprieštarauja sferiškumo hipotezei.

Lentelėje *Tests of Within subjects Effects* pradžioje žiūrime ar yra sąveika tarp dalyko ir lyties. Žiūrime į eilutes *Sphericity assumed*, nes sferiškumo hipotezė nebuvo atmesta. Eilutėje, kurios source yra Egzaminas*Lytis matome, kad Fišerio statistikos F reikšmė yra 1,036, o p-reikšmė yra 0,391. Taigi duomenys neprieštarauja hipotezei, kad sąveikos nėra. Toliau žiūrime ar pažymys priklauso nuo egzamino dalyko. p-reikšmė yra 0,000, taigi pažymis reikšmingai priklauso nuo dalyko.

Lentelėje *Tests of Within Subjects Effects* žiūrime ar egzamino rezultatas priklauso nuo lyties. p-reikšmė yra 0,604, taigi duomenys neprieštarauja hipotezei, kad egzamin rezultatas nepriklauso nuo lyties.

Lentelės *Pairwise comparisons*: pirmoje lentelėje lyginamos lytys ir patvirtinama, kad lytis neįtakoja egzamino rezultato; iš antros lentelės matome, kad kalbos egzamino rezultas reikšmingai nesiskiria nuo politologijos ir matematikos egzaminų rezultatų, bet reikšmingai skiriasi nuo logikos egzamino rezultatų. Politologijos egzamino rezultatas reikšmingai

skirėsi nuo matematikos ir logikos egzaminų rezultatų. Matematikos ir logikos egzaminų rezultatai reikšmingaiu nesiskirė.

Estimated marginal means diagramoje vaizdžiai pavaizduotas ezaminų vidurkių kitinas priklausomai nuo dalyko ir lyties.

Išvada: duomenys prieštarauja hipotezei, kad egzaminai išlaikyti vienodai gerai, duomenys neprieštarauja hipotezi, kad studentai ir studentės egzaminus išlaikė vienodai gerai, studentėms labiausiai sekėsi kalba ir politologija.

(i) Tas pats.

10. Tiesinė regresija

Tarkime, kad norime prognozuoti a. d. Y remdamiesi m kovariantėmis $X_1, \ldots, X_m$.

Žymėkime $\mathbf{X} = (X_0,\ X_1, ..., X_m)^T$ kovariančių vektorių papildytą koordinate $X_0 \equiv 1$. Fiksavus kovariantės $\mathbf{X}$ reikšmes $\mathbf{X}^{(i)} = \boldsymbol{x}^{(i)} = (x_{0i}, x_{1i}, ..., x_{mi})^T$ gauti nepriklausomi a. d. Y stebėjimai $Y_i,\ \ i = 1, ..., n$.

Darome prielaidą, kad dydžiai $\boldsymbol{x}^{(i)}$ yra neatsitiktiniai arba yra nepriklausomų vienodai pasiskirsčiusių a. v. $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ realizacijos. Pastaruoju atveju analizė yra sąlyginė, naudojamasi tik šiomis realizacijomis, bet ne a. v. $\mathbf{X}^{(i)}$ skirstiniais.

Tiesinės regresijos modelis:

$$Y_i = \mu(\boldsymbol{x}^{(i)}) + e_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + e_i, \quad i = 1, ..., n;$$

čia e_i vienodai pasiskirstę nepriklausomi a. d. su nuliniais vidurkiais ir vienodomis dispersijomis $De_i = \sigma^2$.

Taigi tariama, kad sąlyginis vidurkis $\mu(\boldsymbol{x}^{(i)}) = \mathbf{E}(Y_i|\boldsymbol{x}^{(i)}) = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}$ yra tiesinė $m+1$ kovariantės funkcija su nežinomais koeficientais. Iš kitos pusės, šis vidurkis yra tiesinė nežinomų parametrų funkcija su žinomais koeficientais.

Pažymėkime

$$\boldsymbol{Y} = (Y_1, ..., Y_n)^T, \quad \boldsymbol{\beta} = (\beta_0,\ \beta_1, ..., \beta_m)^T, \quad \boldsymbol{e} = (e_i, ..., e_n)^T,$$

$$\boldsymbol{A} = \begin{pmatrix} 1\ x_{1i}\ ...\ x_{1m} \\ 1\ x_{21}\ ...\ x_{2m} \\ ...\ ...\ ...\ ... \\ 1\ x_{n1}\ ...\ x_{nm} \end{pmatrix}.$$

Tada stebėjimus galima užrašyti matriciniu pavidalu

$$\boldsymbol{Y} = \boldsymbol{A\beta} + \boldsymbol{e}, \quad \mathbf{E}(\boldsymbol{Y}) = \boldsymbol{A\beta}, \quad \boldsymbol{D}(\boldsymbol{Y}) = \boldsymbol{D}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}.$$

Tarkime, kad j-oji kovariantė yra tolydi. Nagrinėkime du kovariančių vektorius $\boldsymbol{x}^{(1)}$ ir $\boldsymbol{x}^{(2)}$, kurių visos koordinatės, išskyrus j-ąją, yra vienodos, o $x_j^{(2)} = x_j^{(1)} + 1$. Tada

$$\mu(\boldsymbol{x}^{(2)}) - \mu(\boldsymbol{x}^{(1)}) = \beta_j.$$

Taigi *parametras β_j yra lygus priklausomo kintamojo Y vidurkio pokyčiui, kai j-oji kovariantė padidėja vienetu, kitoms kovariantėms nepakitus.*

Jei j-oji kovariantė nominali, tai norint, kad modelio parametrai turėtų prasmę, kovariantę reikia koduoti. Tarkime, kad j-ji kovariantė x_j nominali, pavyzdžiui, ligos stadija, lytis, rasė, ir pan., ir įgyja k skirtingų reikšmių (sakykime reikšmes 1,2,...,k). Tada vietoje $\beta_j x_j$ tiesinės regresijos modelyje imamas narys $\beta_{j1} z_{j1} + \beta_{j2} z_{j2} + \cdots + \beta_{j,k-1} z_{j,k-1}$, kur

$$z_{jl} = \begin{cases} 1, & \text{jei } x_j = l+1 \quad (l = 1, \ldots, k-1); \\ 0, & \text{jei } x_j = 1. \end{cases}$$

Jei, pavyzdžiui, x_j yra sėklos rūšis, įgyjanti 3 reikšmes, tai modelyje vietoje nario $\beta_j x_j$ imamas narys $\beta_{j1} z_{j1} + \beta_{j2} z_{j2}$, kur z_{j1} įgyja reikšmę 1 antrajai javų rūšiai, o z_{j2} įgyja reikšmę 1 trečiajai javų rūšiai.

Parametras β_{jl} parodo priklausomo kintamojo Y vidurkio pakitimą, kai j-osios kovariantės reikšmė pakinta nuo pirmosios iki $l+1$-osios, kitoms kovariantėms nepakitus. Pavyzdžiui, jei x_j yra sėklos rūšis, įgyjanti 3 reikšmes, tai β_{j1} parodo derlingumo vidurkio pokytį, pereinant nuo 1-sios prie 2-ios rūšies, o β_{j2} parodo derlingumo vidurkio pokytį, pereinant nuo 1-sios prie 3-ios rūšies.

Tarkime, kad matrica $\boldsymbol{A}^T\boldsymbol{A}$ neišsigimusi. Tada parametro $\boldsymbol{\beta}$ mažiausiųjų kvadratų įvertinys yra

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y},$$

$$\mathbf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}.$$

Šis įvertinys gaunamas minimizuojant pagal $\boldsymbol{\beta}$ kvadratų sumą

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_m x_{im})^2 = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})$$

pagal $\boldsymbol{\beta}$.

Analogiškai, jeigu $\theta = \boldsymbol{L}^T\boldsymbol{\beta} = L_0\beta_0 + L_1\beta_1 + \cdots + L_m\beta_m$ yra tiesinė regresinių parametrų funkcija, tai $\hat{\theta} = \boldsymbol{L}^T\hat{\boldsymbol{\beta}}$ yra minimalios dispersijos įvertinys visų nepaslinktųjų tiesinių parametro θ įvertinių klasėje, ir

$$\mathbf{E}(\hat{\theta}) = \boldsymbol{\theta}, \quad \boldsymbol{D}(\hat{\theta}) = \sigma^2 b^2, \quad b^2 = \boldsymbol{L}^T\boldsymbol{C}\boldsymbol{L}, \quad \boldsymbol{C} = (\boldsymbol{A}^T\boldsymbol{A})^{-1} = [c_{ij}]_{(m+1)\times(m+1)}.$$

Atskiru atveju gauname vidurkio $\mu(\boldsymbol{x}) = \mathbf{E}(Y|\boldsymbol{x})$ įvertinį:

$$\hat{\mu}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}^T\boldsymbol{x} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_m x_m.$$

Nepaslinktasis dispersijos σ^2 įvertinys yra

$$\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-m-1}, \quad \mathbf{E}s^2 = \sigma^2, \quad SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

$$\hat{Y}_i = \hat{\mu}(\boldsymbol{x}^{(i)}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im}.$$

Atsitiktiniai dydžiai $\hat{Y}_i$ ir Y_i vadinami atitinkamai *prognozuojamomis* ir *stebėtomis* priklausomo kintamojo Y reikšmėmis, o a. d. $\hat{e}_i = Y_i - \hat{Y}_i$ vadinami *liekamosiomis arba prognozės paklaidomis.*

Taigi

$$\hat{\boldsymbol{\beta}} \sim N_{m+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}), \quad \frac{s^2(n-m-1)}{\sigma^2} \sim \chi^2(n-m-1),$$

be to atsitiktiniai dydžiai $\hat{\boldsymbol{\beta}}$ ir s^2 yra nepriklausomi; čia $\boldsymbol{C} = (\boldsymbol{A}^T\boldsymbol{A})^{-1} = [c_{ij}]_{(m+1)\times(m+1)}$.

$\theta = \boldsymbol{L}^T\boldsymbol{\beta}$ įvertinys $\hat{\theta} = \boldsymbol{L}^T\hat{\boldsymbol{\beta}}$ turi savybę

$$\frac{\hat{\theta} - \theta}{s\, b(\boldsymbol{L})} \sim S(n-m-1), \quad b^2(\boldsymbol{L}) = \boldsymbol{L}^T\boldsymbol{C}\boldsymbol{L}.$$

Atskirais atvejais $\theta = \beta_i$ ir $\theta = \mu(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ gauname

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \sim S(n-m-1), \quad \frac{\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})}{s b(\boldsymbol{x})} \sim S(n-m-1),\ b^2(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x}. \tag{0.22}$$

Parametro θ pasikliovimo intervalas su pasikliovimo lygmeniu $Q = 1-\alpha$ yra

$$\hat{\theta} \pm b\, s\, t_{\alpha/2}(n-m-1). \tag{0.23}$$

Atskirais atvejais, parametro β_i pasikliovimo intervalas yra

$$\hat{\beta}_i \pm c_{ii} s\, t_{\alpha/2}(n-m-1), \tag{0.24}$$

o parametro $\mu(\boldsymbol{x})$ yra

$$\hat{\mu}(\boldsymbol{x}) \pm b(\boldsymbol{x})\, s\, t_{\alpha/2}(n-m-1). \tag{0.25}$$

Lygmens $1-\alpha$ prognozės intervalas reikšmei Y_{n+1} yra

$$\hat{\mu}(\boldsymbol{x}) \pm s\sqrt{(1 + b^2(\boldsymbol{x})}\, t_{\alpha/2}(n-m-1).$$

Jis platesnis už pasikliautinąji intervalą vidurkiui $\mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$.

Nagrinėkime hipotezę

$$H_{j_1\ldots j_k} : \beta_{j_1} = \cdots = \beta_{j_k} = 0,$$

kur $1 \le j_1 \le \cdots \le j_k \le m$, k fiksuotas skaičius, $k = 1, ..., m$. Jei ši hipotezė teisinga, tai kovariantės $x_{j_1}, ..., x_{j_k}$ nėra reikšmingos priklausomo kintamojo prognozei ir jas galima išmesti iš modelio.

Atskiru atveju $k = 1$, $j_1 = j$ turime hipotezę

$$H_j : \beta_j = 0, \tag{0.26}$$

kuri reiškia, kad kovariantė x_j nėra reikšminga priklausomo kintamojo prognozei.

Atveju $k = m$ turime hipotezę

$$H_{1\ldots m} : \beta_1 = ... = \beta_m = 0. \tag{0.27}$$

Ši hipotezė reiškia, kad regresijos aplamai nėra. Kovariančių reikšmių žinojimas neduoda jokios papildomos informacijos apie Y reikšmes.

Prognozuojamų ir stebėtųjų reikšmių sklaidą charakterizuoja kvadratų sumos:

$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ – *liekamųjų paklaidų kvadratų suma*,

$SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ – *pilnoji kvadratų suma*,

$SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ – *regresijos kvadratų suma.*

Pilnoji kvadratų suma SS_T charakterizuoja stebėjimų Y_i sklaidą apie jų aritmetinį vidurkį $\bar{Y}$; regresijos kvadratų suma SS_R – regresijos modeliu prognozuojamų reikšmių $\hat{Y}_i$ sklaidą apie $\bar{Y}$; liekamųjų kvadratų suma SS_E – atstumą tarp stebėtų ir prognozuojamų reikšmių. Teisinga lygybė $SS_E = SS_T - SS_R$, taigi ši kvadratų suma dar parodo, kurią Y_i sklaidos dalis lieka nepaaiškinta regresiniu modeliu.

Hipotezė $H_{j_1...j_k}$ atmetama su reikšmingumo lygmeniu α, jei

$$F_{j_1,,,j_k} = \frac{(SS_E^{(m-k)} - SS_E)/k}{SS_E/(n-m-1)} > F_\alpha(k, n-m-1).$$

Čia $SS_E^{(m-k)}$ yra SS_E analogas modeliui be kovariančių $x_{j_1}, ..., x_{j_k}$:

$$SS_E^{(m-k)} = \sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2, \quad \tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_{s_1}x_{s_1} + ... + \tilde{\beta}_{s_{m-k}}x_{s_{m-k}};$$

$s_1, ..., s_{m-k}$ papildo $j_1, ..., j_k$ iki $1, 2, ..., m$, o $\tilde{\beta}_0, \tilde{\beta}_{s_1}, ..., \tilde{\beta}_{s_{m-k}}$ yra regresijos parametrų įvertiniai modelyje be kovariančių $x_{j_1}, ..., x_{j_k}$.

Hipotezės $H_j : \beta_j = 0$ atveju kriterijaus statistika yra

$$F_j = \frac{SS_E^{(m-1)} - SS_E}{SS_E/(n-m-1)}.$$

Hipotezė atmetama reikšmingumo lygmens α kriterijumi, kai

$$F_j > F_\alpha(1, n-m-1).$$

Remiantis (3.3.16) šią hipotezę galima tikrinti ir naudodami Stjudento kriterijų. Hipotezė atmetama α lygmens kriterijumi, kai

$$|t| = \frac{|\hat{\beta}_j|}{s\sqrt{c_{jj}}} > t_{\alpha/2}(n-m-1).$$

Abu pastarieji kriterijai yra ekvivalentūs.

Hipotezė $H_{1...m}$ apie regresijos nebuvimą yra atmetama reikšmingumo lygmeniu α kriterijumi, kai

$$F_{1...m} = \frac{SS_R/m}{SS_E/(n-m-1)} = \frac{MS_R}{MS_E} > F_\alpha(m, n-m-1).$$

Atsitiktinis dydis

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T}$$

vadinamas *determinacijos koeficientu*.

Determinacijos koeficientas R^2 įgyja reikšmes iš intervalo $[0, 1]$. Jis parodo santykinę Y_i sklaidos dalį, paaiškinamą regresiniu modeliu, taigi charakterizuoja prognozavimo kokybę.

Jei prognozavimas idealus, t.y. $\hat{Y}_i = Y_i$, tai $SS_E = 0$ ir $R^2 = 1$. Jei nėra regresijos, t.y. su visais $\boldsymbol{x}^{(i)}$ vidurkio $\mu(\boldsymbol{x}^{(i)})$ prognozė nepriklauso nuo $\boldsymbol{x}_i$, tai $\hat{Y}_i = \bar{Y}$, taigi $SS_E = SS_T$ ir $R^2 = 0$.

(a) *Vieno kintamojo tiesinė regresija*

Duomenų įvedimas

Duomenys įvedami į du stulpelius "Trenir" (tai nepriklausomas kintamasis x) ir "Komplim" (tai priklausomas kintamasis Y).

Skaičiavimas

Pradžioje norima išsiaiškinti, ar pagal treniruočių skaičių galima prognozuoti komplimentų skaičių.

Rasime sklaidos diagramą (scatter diagram):

Graphs-Legacy dialogs-Scatter-Simple scatter-Define

Y axis: Komplim

X axis: Trenir

OK

Du kartus spragtelėkite ant grafiko. Pasirodys naujas langas.

Virš grafiko spragtelėkite penktąją diagramą "Add fit line at total". Tada papildomai bus nubrėžta regresijos tiesė $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Close. Prie grafiko pateikiama ir R^2 reikšmė. Ďvilgtelėjus į grafiką matosi, kad didėjant trniruočių skaičiui stebima komplimentų skaičiaus padidėjimo tendencija.

Suskaičiuokite regresinėje analizėje naudojamas statistikas.

Analyze-Regression-Linear

Dependent: Komplim

Independent: Trenir

Statistics: Estimates, Confidence intervals, Covariance matrix, Model fit, R squared change, Descriptives.

Continue, OK.

Rezultatai

Lentelėje *Descriptive statistics* pateikiamos komplimentų skaičiaus ir treniručių skaičiaus empiriniai vidurkiai, kvadratiniai nuokrypiai bei valdininkų skaičius.

Lentelėje *Correlations* pateiktas koreliacijos koeficientas tarp treniruočių skaičiaus ir komplimentų skaičiaus.

Lentelėje *ANOVA* pateiktos paaiškintoji regresijos (explained by regression), liekamųjų paklaidų arba liekamoji (residual) ir pilnoji (total) kvadratų sumos (sums of squares):

$$SS_E = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \quad SS_R = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \quad SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

jas atitinkančių chi kvadrato statistikų laisvės laipsnių (degrees of freedom) skaičiai 1 , $n-2$ ir $n-1$, be to vidutinės kvadratinės paklaidos (mean square errors) $MSE = SSE/1$, $MSR = SSR/(n-2)$, Fišerio statistika $F = MSR/MSE$ bei jos reikšmingumas, t.y. p-reikšmė hipotezei $H_1 : \beta_1 = 0$ tikrinti.

Lentelėje *Model summary* pateikiamas determinacijos koeficientasR^2, kuris paprastosios liesinės regresijos atveju (t.y., kai $m = 1$) skaičiuojasi labai paprastai:

$$R^2 = SS_R/SS_T = 1 - SS_E/SS_T = \frac{[\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2};$$

pataisytasis determinacijos koeficientas (adjusted coefficient of determination) $R_a^2 = 1 - \frac{SS_E/(n-2)}{SS_T/(n-1)}$ (kai turima m kovariančių, tai $R_a^2 = 1 - \frac{SS_E/(n-m-1)}{SS_T/(n-1)}$); modelio vidutinio kvadratinio nuokrypio σ įvertis $\sqrt{MSE}$, kur $MSE = \hat{\sigma}^2 = SSE/(n-2)$; Fišerio statistika $F_{1\ldots m} = F_1 = MS_R/MS_E$ ir jos laisvės laipsnių skaičiai 1 ir $n-2$; P-reikšmė tikrinant hipotezę $H : \beta_1 = 0$ apie regresijos nebuvimą. Šiuo konkrečiu atveju gauta, kad $sig = 0.000$, taigi hipotezė atmetama. Hipotezė, kad komplimentų skaičius nepriklauso nuo treniruočių skaičiaus, atmetama.

Lentelėje *Coefficients* pateikti koeficientų β_0 ir β_1 įverčiai $b_0 = \hat{\beta}_0$ ir $b_1 = \hat{\beta}_1$ bei jų standartinių nuokrypių įverčiai s_0 ir s_1; standartizuotas koeficientas $\hat{\beta} = \hat{\beta}_1 s_X/s_Y$ (jų pagalba galima lyginti kovariančių įtaką priklausomam kintamajam, eliminavus matavimo vienetus; vienos kovariantės atveju neįdomus); Stjudento statistikų $t_0 = \hat{\beta}_0/S_0$ ir $t_1 = \hat{\beta}_1/S_1$ reikšmės; pasikliautinieji intervalai koeficientams β_0 bei β_1.

Lentelėje *Coefficient correlations* pateiktas tiktai įverčio $\hat{\beta}_1$ dispersijos įvertis, nes yra tik vienas nepriklausomas kintamasis.

Fiksuokime kovariančių vektoriaus reikšmę x_0. Priklausomo kintamojo prognozė yra $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Šiuo konkrečiu atveju fiksuojama treniruočių skaičiaus reikšmė $x_0 = 267$ ir ieškoma komplimentų skaičiaus prognozė.

Duomenų faile stulpelyje "Trenir" pridėkime dar vieną reikšmę, o stulpelyje "Komplim" nieko nepridėkime.

Analyze-Regression-Linear

Nuspauskite *Save* ir prie "Predicted values" pažymėkite "Unstandardized".

Duomenų faile atsiras trečias stulpelis PRE_1, kuriame pateiktos Y prognozuojamos reikšmės ne tik kiekvienam x_i, bet ir x_0-iam. Gauname, kad valdininui, kuris treniravosi 267 kartus komplimentų skaičiaus prognozė yra 122 komplimentai.

Norint suskaičiuoti pasikliovimo intervalus vidurkiui ir prognozės intervalus eiliniam matavimui atliekame šiuos veiksmus:

Analyze-Regression-Linear-Save

Prie "Prediction intervals" pažymime Mean ir Individual.

Continue, OK.

Duomenų faile atsiras papildomi keturi stulpeliai $LMCI_1$, $UMCI_1$ (lower and upper confidence intervals for mean) bei $LICI_1$, $UICI_1$ (lower and upper confidence intervals for individual, iš tikrųjų prognozės intervalų apatinis ir viršutinis rėžiai). Pavyzdžiui, valdininkams, kurie treniravosi 267 kartus 0.95 pasikliovimo intervalas vidutiniam komplimentų skaičiui yra $(112,47; 131,64)$. Jei dar vienam valdiniūnkui būtų numatyta pravesti 267 treniruotes, tai jo komplimentų skaičiaus prognozės intervalas būtų $(71,19; 172,91)$.

Prieš darant statistines išvadas būtina patikrinti, ar modelis gerai atitinka realiems duomenims. Pagrindinės tiesinės regresijos modelio prielaidos yra :

a) a.d. $e_i = Y_i - \boldsymbol{\beta}^T \boldsymbol{x}$ dispersijų lygybė;

b) regresinės funkcijos $M(\boldsymbol{x}) = \mathbf{E}(Y(\boldsymbol{x}))$ tiesiškumas;

c) a.d. e_i normalumas;

d) e_i nepriklausomumas.

Prielaidų tikrinimas.

Skirtumus tarp stebėtų ir prognozuojamų reikšmių $\hat{e}_i = Y_i - \hat{Y}_i$ vadiname liekanomis.

Atsitiktiniai dydžiai

$$\tilde{e}_i = \frac{\hat{e}_i}{\sqrt{MS_R h_{ii}}},$$

vadinami *stjudentizuotomis liekanomis (studentized residuals)*, $\mathbf{E}(\tilde{e}_i) \approx 0, \quad \mathbf{Var}(\tilde{e}_i) \approx 1$.

Prognozuojamos reikšmės $\hat{Y}_i$ (pre_1), standartizuotos prognozuojamos reikšmės $\hat{Y}_i/\hat{\sigma}_{Y_i}$ (zpr), likučiai $\hat{e}_i$ (res), stjudentizuoti likučiai $\tilde{e}_i$ (sdr) suskaičiuojami ir išsaugomi, naudojant

Analyze-Regression-Linear-Save

Prie "Predicted values" pažymima Unstandardized ir Standardized.

Prie "Residuals" pažymima Unstandardized ir Studentized.

a) *Dispersijų lygybė.*

Jei a.d. e_i dispersijos nelygios, tai sakoma, kad turime *heterodeskatyvumą.*

Nagrinėkime plokštumą su abscisių ašimi Y ir ordinačių ašimi e. Taškai $(\hat{Y}_i, \hat{e}_i)$, $(i = 1, ..., n)$ yra išsibarstę horizontalioje juostoje apie horizontalią simetrijos ašį $e = 0$. Jei modelis vidurkiui M(x) gerai parinktas, bet yra heterodeskatyvumas, tai taškai $(\hat{Y}_i, \hat{e}_i)$ išsibarstę taip pat apie tiesę $e = 0$, bet juostos plotis nevienodas. Pavyzdžiui, jei dispersija didėja didėjant $\hat{Y}_i$, tai juosta platėja.

Patikrinti dispersijų lygybę:

Analyse-Regression-Linear-Plots

Y: ZRESID X: ZPRED

Continue, OK.

Gauname $(\hat{Y}_i, \hat{e}_i)$ diagramą. Taškai iš tikrųjų išsibarstę horizontalioje juostoje, kurios centrinė ašis yra $e = 0$.

b) *Regresinės funkcijos $M(\boldsymbol{x}) = \mathbf{E}(Y(\boldsymbol{x}))$ tiesiškumas.*

Jei taškai $(\hat{Y}_i, \hat{e}_i)$ yra išsibarstę apie kitą kreivę, nesutampančią su $e = 0$, tai $M(\boldsymbol{x})$ modelis nėra gerai parinktas.

Šiuo atveju duomenys išsibarstę apie horizontalią ašį.

Vietoje taškų $(\hat{Y}_i, \hat{e}_i)$ galima nagrinėti taškus $(x_{ij}, \hat{e}_i)$, $(i = 1, ..., n)$ su fiksuotais j. Jei modelis gerai parinktas, šie taškai turėtų būti horizontalioje juostoje su simetrijos ašimi $e = 0$. Jei taip nėra, galima tvirtinti, kad j-oji kovariantė nedaro įtakos vidurkiui $M(\boldsymbol{x})$ arba reikia įjungti daugiau kovariančių į modelį.

Graphs-Scatter-Simple-Define

Y axis Studentized residuals X axis Trenir

Paprastosios tiesinės regresijos atveju galima nagrinėti taškų (x_i, Y_i) diagramą. Jei šie taškai yra išsibarstę ne apie tiesę, o apie kokią nors kreivę, galima tarti kad modelis nėra gerai parinktas.

c) *A.d. e_i normalumas*

Dažnai koreliacijos tarp $\tilde{e}_i$ yra mažos ir į a.d. $\tilde{e}_1, ..., \tilde{e}_n$ žiūrima kaip į n.v.p. $N(0,1)$ dydžius. Norint grubiai patikrinti normalumą, galima nubrėžti a.d. $\tilde{e}_i$ histogramą.

Analyze-Regression-Plots-Histogram

Y ZRESID X ZPRED

Šiuo atveju histograma panaši į normalaus skirstinio histogramą.

Galima panaudoti kitą metodą : atidėti plokštumoje $(0eq)$ taškus $(\tilde{e}_{(i)}, q_{(i)})$, kur $q_{(i)} = \Phi^{-1}\left(\frac{i-1/2}{n}\right)$ yra $N(0,1)$ skirstinio $\left(\frac{i-1/2}{n}\right)$-kvantiliai. Tada šie taškai turėtų būti išsibarstę apie tiesę $e = q$.

Analyze-Regression-Plots-Normal probability plot

Y ZRESID X ZPRED

Šiuo atveju taip ir yra.

d) *A.d. e_i nepriklausomumas.*

Statistika

$$d = \frac{\sum_{i=2}^{n}(\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n}\hat{e}_i^2} \approx 2(1 - r_1)$$

vadinamas *Durbino Votsono (Durbin-Watson) statistika* . Ši statistika įgyja reikšmes intervale $[0, 4]$. Statistika d artima 2, jei atsitiktiniai dydžiai $\hat{e}_i$ nepriklausomi. Jei $d < 2$, teigiama koreliacija, jei $d > 2$, neigiama koreliacija. Statistikos d dėsnis nepriklauso nuo nežinomų parametrų ir kritinės reikšmės d_i ir d_s duotos lentelėse. Nepriklausomumo hipotezė atmetama, jei $d < d_i$ arba $d > d_s$.

Analyze-Regression-Linear-Statistics-Durbin-Watson

Durbino-Votsono statistikos reikšmė duota lentelėje Model summary.

Šiuo atveju $d = 2,030$, $n = 32$. V.Čekanavičiaus ir G.Murausko knygos Staistika II 13 lentelėje duota, kad šiuo atveju $d_L = 1,27, d_U = 1,40$. .Kadangi $d_U < d < 4 - d_U$, t.y. $1,40 < 2,030 < 2,60$, tai duomenys neprieštarauja hipotezei, kad stebėjimai nepriklausomi.

Įvedant duomenis galimos klaidos ir kai kurie stebėjimai gali nesiderinti su kitais ir tuo iškreipti statistines išvadas. Tokius stebėjimus vadiname *nukrypusiomis reikšmėmis (outliers)*.

Net jei modelis gerai parinktas, vertinimas gali būti blogas, jei tarp taškų (x_i, Y_i) yra nukrypusios reikšmės. Jas atitiks taškai su didelėmis liekanomis $\hat{e}_i = Y_i - \hat{Y}_i$. Nukrypusi reikšmė yra įtakinga, jei jos išmetimas smarkiai keičia įverčio $\hat{\beta}$ reikšmę. Taško (x_i, Y_i) įtaka gali būti matuojama *Kuko atstumo (Cook's distance)* pagalba.

Norint suskaičiuoti šį atstumą, suskaičiuojamos *modifikuotoji prognozuojamoji reikšmė (adjusted predicted value)* $\hat{Y}_{ia}$ kuri apibrėžiama kaip $\hat{Y}_i$, suskaičiuota naudojant tiktai taškus

$$(x_1, Y_1), ..., (x_{i-1}, Y_{i-1}), (x_{i+1}, Y_{i+1}), ..., (x_n, Y_n).$$

Kuko atstumas

$$C_i^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_{ia} - \hat{Y}_i)^2}{(m+1)MS_R}.$$

Praktinė taisyklė : jei $C_i^2 > 1$, taškas įtakingas.

Norint pažiūrėti stebėjimus, kurių reikšmės nukrypę daugiau, negu per tris standartinius nuokrypius nuo prognozuojamo vidurkio duotai kovariantės reikšmei:

Analyse-Regression-Linear-Statistics-Residuals-Casewise diagnostics-Outliers outside

Save pažymėti Cook's distance

Duodami nukrypusių stebėjimo numerį, standartizuoti likučiai, priklausomojo kintamojo reikšmės, prognozuojamos reikšmės, likučiai.

Su

Analyse-Regression-Linear-Statistics-Residuals-Casewise diagnostics-All cases

galima pažiūrėti tą patį visiems stebėjimams.

(b) *Vieno kintamojo tiesinė regresija*

Tas pats.

(c) *Kelių kintamųjų tiesinė regresija*

Duomenų įvedimas

Duomenys įvedami į tris stulpelius "Dovana" (tai priklausomas kintamasis Y), "Verte" ir "Firmusk" (tai nepriklausomas kintamieji x_1 ir x_2).

Skaičiavimas

Suskaičiuokite regresinėje analizėje naudojamas statistikas.

Analyze-Regression-Linear

Dependent: Dovana

Independent: Verte Firmusk

Statistics: Estimates, Confidence intervals, Covariance matrix, Model fit, R squared change, Descriptives.

Continue, OK.

Rezultatai

Lentelėje *Descriptive statistics* pateikiamos priklausomo ir nepriklausomų kintamųjų empiriniai vidurkiai, kvadratiniai nuokrypiai bei imčių dydžiai.

Lentelėje *Correlations* pateikta atsitiktinio vektoriaus $(Y, x_1, x_2) =$

$(Dovana, Verte, Firmusk)$ koreliacinės matricos įvertis.Koreliacijos koeficientas tarp verės ir dovanos dydžio yra 0,535, taro firmų skaičiaus ir dovanos dydžio yra 0,797.

Lentelėje *Model summary* matome, kad determinacijos koeficientas $R^2 = 0,855$, taigi regresinis modelis paaiškina $85,5\%$ duomenų sklaidos. Pataisytasis determinacijos koeficientas (adjusted coefficient of determination)panašus: $R_a^2 = 0,833$;

Lentelėje *ANOVA* pateiktos regresijos, liekamųjų paklaidų ir pilnoji kvadratų sumos $SS_R = 1273,003, SS_E = 215,831, SS_T = 1488,834$, atitinkami laisvės laipsnių skaičiai $m = 2$, $n - m - 1 = n - 3 = 13$ ir $n - 1 = 15$, be to $MS_R = SS_R/m = 636,502$, $MS_E = SS_E/(n - m - 1) = 16,602$, Fišerio statistikos reikšmė $F_{1...m} = MS_R/MS_E = 38,338$ ir P-reikšmė $sig = 0,000$, gauta tikrinant hipotezę $H : \beta_1 = \beta_2 = 0$ apie regresijos nebuvimą. Hipotezė, kad dovanos dydis nepriklauso nuo objektų vertės ir konkurse dalyvavusių firmų skaičiaus, atmetama.

Lentelėje *Coefficients* pateikti koeficientų β_0, β_1 ir β_2 įverčiai $b_0 = \hat{\beta}_0 = 29,527$, $b_1 = \hat{\beta}_1 = 0,532$, $b_2 = \hat{\beta}_2 = 1,120$ bei jų standartiniųi nuokrypių įverčiai s_0, s_1 ir s_1; standartizuoti koeficientai $\tilde{\beta}_1 = \hat{\beta}_1 s_X/s_Y = 0,471$ ir $\tilde{\beta}_2 = \hat{\beta}_2 s_X/s_Y = 0,757$ (jų pagalba lyginama kovariančių įtaką priklausomam kintamajam, eliminavus matavimo vienetus; šiuo atveju gauname, kad dalyvaujančių firmų skaičiui ir privatizuotų objektų kainai padidėjus tuo pačiu procentu didesnį efektą dovanos dydžiui turi firmų skaičiaus padidėjimas); Stjudento statistikų $t_0 = \hat{\beta}_0/S_0$, $t_1 = \hat{\beta}_1/S_1$ ir $t_1 = \hat{\beta}_1/S_1$ reikšmės (atitinkamai 3,120;4441; 7142) ir p-reikšmės hipotezėms $H_j : \beta_j = 0$ tikrinti. p-reikšmė hipotezei $H_1 : \beta_1 = 0$ gaunama 0,001, taigi hipotezė apie privatizuojamų objektų kainos įtakos nebuvimą dovanos dydžiui atmetama. p-reikšmė hipotezei $H_2 : \beta_2 = 0$ gaunama 0,000, taigi hipotezė apie privatizavime dalyvaujančių firmų skaičiaus įtakos nebuvimą dovanos dydžiui taip par atmetama.

Lentelėje dar duotas dispersijos mažėjimo daugiklis: $VIF_j = 1,007$ abiejų nepriklausomų kintamųjų atveju, taigi multikolinearumo nėra (multikolinearumas yra labai nepageidautina savybė. Jis atsiranda, jei nepriklausomi kintamieji labai stipriai koreliuoti. Tuo atveju regresijos koeficientų įvertinių dispersijos labai didelės. Tarkime, kad R_j^2 yra determinacijos

koeficientas modelyje, kuriame x_j yra priklausomas kintamasis, o kiti pradinio modelio kintramieji $x_1, ..., x_{j-1}$,

$x_{j+1}, ..., x_m$ yra nepriklausomi kintamieji. Kintamasi x_j yra multikolinearus, jei R_j^2 artimas 1. Tuomet dispersijos mažėjimo daugiklis $VIF_j = \frac{1}{1-R_j^2}$ didelis. Kintamasis yra "perdaug multikolinearus", jei $VIF_j > 4$. Tada šis nepriklausomas kintamasis pašalinamas iš pradinio modelio. Šiuo konkrečiu atveju abu nepriklausomi kintamieji nėra perdaug multikolinearūs, nes $VIF_1 = VIF_2 = 1,007$).

Lentelėje pateikti ir pasikliovimo intervalai koeficientams β_j.

Lentelėje *Coefficient correlations* pateiktas įverčio $\hat{\beta}_1$ ir $\hat{\beta}_2$ koreliacinės matricos įvertis (t.y. dispersijų ir koreliacijos koeficientų įverčiai).

Duomenų faile stulpelyje "Verte" pridėkime dar vieną reikšmę 90, o stulpelyje "Firmusk" pridėkime reikšmę 10.

Analyze-Regression-Linear

Nuspauskite *Save* ir prie "Predicted values" pažymėkite "Unstandardized".

Duomenų faile atsiras trečias stulpelis PRE_1, kuriame pateiktos Y prognozuojamos reikšmės ne tik kiekvienai porai $(x_{i1}, x_{i2}$, bet ir porai $(90, 10)$. Gauname, kad jei privatizuojamo objekto kaina yra 90 mln, eurų, o konkurse dalyvavo 10 firmų, tai dovanos prognozė yra 88,57 tūkst. dolerių.

Pasikliovimo intervalai vidurkiui ir prognozės intervalai eiliniam matavimui bei prielaidų tikrinimas atliekami kaip paaiškinta uždavinyje a).

(d) *Kelių kintamųjų tiesinė regresija* Tas pats.

11. Logistinė regresija

Tarkime, kad atsitiktinio įvykio A tikimybė gali priklausyti nuo nepriklausomų kintamųjų (kovariančių) $x_1, ..., x_m$.

Apibrėžkime atsitiktinį dydį Y, kuris įgyja reikšmę 1, kai įvykis A įvyksta ir įgyja reikšmę 0, kai įvykis A neįvyksta. Taigi galime sakyti, kad eksperimento metu stebimos a. d. Y reikšmės. Pažymėkime $\boldsymbol{x} = (x_0, x_1, ..., x_m)^T$ kovariančių vektorių papildytą koordinate $x_0 = 1$.

Pažymėkime

$$\pi(\boldsymbol{x}) = \mathbf{E}(Y|\boldsymbol{x}) = \mathbf{P}\{Y = 1|\boldsymbol{x}\} = \mathbf{P}\{A|\boldsymbol{x}\}.$$

Logistinės regresijos modelis:

$$logit(\boldsymbol{x}; \boldsymbol{\beta}) = \ln \frac{\pi(\boldsymbol{x}; \boldsymbol{\beta})}{1 - \pi(\boldsymbol{x}; \boldsymbol{\beta})} = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m = \boldsymbol{\beta}^T \boldsymbol{x}.$$

Iš modelio išplaukia, kad įvykio A sąlyginė tikimybė žinant $\boldsymbol{x}$ apibrėžiama formule

$$\pi(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m}} = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}.$$

Tarkime, kad j-oji kovariantė yra tolydi. Imkime du kovariančių vektorius $x^{(1)}$ ir $x^{(2)}$, kurių visos koordinatės, išskyrus j-ąją, yra vienodos, o $x_j^{(2)} = x_j^{(1)} + 1$. Tada

$$\frac{\pi(x^{(2)})/(1 - \pi(x^{(2)})}{\pi(x^{(1)})/(1 - \pi(x^{(1)})} = e^{\beta_j}.$$

Bet kokio įvykio B pasirodymo ir nepasirodymo tikimybių santykį $\mathbf{P}(B)/(1 - \mathbf{P}(B)$ pavadinkime *šansu*. Taigi *parametras e^{β_j} parodo kiek kartų pasikeičia įvykio A šansas, kai j-oji kovariantė padidėja vienetu, kitoms kovariantėms nepakitus; parametras e^{β_j} yra šansų santykis.*

Jei j-oji kovariantė nominali, tai norint, kad modelio parametrai turėtų prasmę, ši kovariantė koduojama lygiai taip pat, kaip tiesinės regresijos atveju.

Tarkime, kad j-ji kovariantė nominali (pavyzdžiui ligos stadija, lytis) ir įgyja k skirtingų reikšmių. Tada vietoje $\beta_j x_j$ modelyje imamas narys

$$\boldsymbol{\beta}_j^T \boldsymbol{x}_j = \beta_{j1} x_{j1} + \beta_{j2} x_{j2} + \cdots + \beta_{j,k-1} x_{j,k-1};$$

čia

$$\boldsymbol{x}_j = (x_{j1}, \ldots, x_{j,k-1})^T, \quad \boldsymbol{\beta} = (\beta_{j1}, \ldots, \beta_{j,k-1})^T,$$

$$x_{jl} = \begin{cases} 1, & \text{jei } x_j \text{ įgyja } l\text{-ją reikšmę} \quad (l = 2, \ldots, k); \\ 0, & \text{kitais atvejais,} \end{cases}$$

Panagrinėkime koeficientų ir modelio interpretaciją po kodavimo. Imkime du kovariančių vektorius $x^{(1)}$ ir $x^{(l+1)}$, kuriems visos kovariantės, išskyrus j-ąją nominalią kovariantę, yra vienodos, o j-josios kovariantės reikšmė pirmajam vektoriui yra pirmoji, o antrajam $(l+1)$-oji. Tada

$$\frac{\pi(x^{(l+1)})/(1-\pi(x^{(l+1)})}{\pi(x^{(1)})/(1-\pi(x^{(1)})} = e^{\beta_{jl}}.$$

Taigi *parametras $e^{\beta_{jl}}$ parodo šansų santykį tarp objektų, kurių j-oji kovariantė įgyja l-ją reikšmę bei objektų, kurių j-oji kovariantė įgyja nulinę reikšmę, kitoms kovariantėms nepakitus.*

Tarkime, kad nežinomo regresijos parametro $\boldsymbol{\beta}$ (tuo pačiu ir tikimybės $\pi(\boldsymbol{x};\boldsymbol{\beta})$) vertinimui atliekama n nepriklausomų eksperimentų; i-sis eksperimentas atliekamas prie kovariantės $\boldsymbol{x}$ reikšmės $\boldsymbol{x}^{(i)}$, $i = 1, ..., n$.

Kiekvieno eksperimento metu stebimas atsitiktinis dydis

$$Y_i = \begin{cases} 1, & \text{jei } i\text{-jo eksperimento metu įvyksta } A; \\ 0, & \text{priešingu atveju.} \end{cases}$$

Taigi turime imtį

$$(Y_1, \boldsymbol{x}^{(1)}), ..., (Y_n, \boldsymbol{x}^{(n)}).$$

Atsitiktiniai dydžiai Y_i turi sąlyginius Bernulio skirstinius:

$$(Y_i|\boldsymbol{x}^{(i)};\boldsymbol{\beta}) \sim B(1, \pi(\boldsymbol{x}^{(i)})), \quad i = 1, ..., n.$$

Tikėtinumo funkcija yra

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi(\boldsymbol{x}^{(i)};\boldsymbol{\beta})]^{Y_i} [1 - \pi(\boldsymbol{x}^{(i)};\boldsymbol{\beta})]^{1-Y_i},$$

jos logaritmas

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}) - \ln(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im}})].$$

Didžiausiojo tikėtinumo įvertinys $\hat{\boldsymbol{\beta}}$ maksimizuoja logtikėtinumo funkciją.

Suradus įvertinį $\hat{\boldsymbol{\beta}}$, galima įvertinti įvykio A sąlyginę tikimybę $\pi(\boldsymbol{x})$, šansų santykius e^{β_i}:

$$\hat{\pi}(\boldsymbol{x}) = \frac{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}}}{1 + e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}}}, \quad e^{\hat{\beta}_i}, \quad i = 1, ..., m,$$

ar kitas parametro $\boldsymbol{\beta}$ funkcijas.

Pažymėkime

$$\hat{Y}_i = \hat{\pi}(\boldsymbol{x}^{(i)}) = \frac{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}^{(i)}}}{1 + e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}^{(i)}}}$$

stebėtos reikšmės Y_i prognozę.

1) Kai funkcija $\pi(\boldsymbol{x})$ nežinoma, tai modelyje yra n nežinomų parametrų $p_i = \pi(\boldsymbol{x}^{(i)})$, DT įvertiniai yra $\hat{p}_i = Y_i$ ir tikėtinumo funkcijos maksimumas yra

$$L_0(\hat{\boldsymbol{p}}) = 1.$$

Siauresnio logistinės regresijos modelyje yra $(m+1)$ nežinomas parametras $\beta_0, ..., \beta_m$. Didžiausiojo tikėtinumo funkcija maksimizuojama taške $\hat{\boldsymbol{\beta}}$ ir jos maksimumas yra

$$L(\hat{\boldsymbol{\beta}}) = \prod_{i=1}^{n} \hat{Y}_i^{Y_i} (1 - \hat{Y}_i)^{1-Y_i} \le L_0(\hat{\boldsymbol{p}}).$$

Dar siauresnis modelis gaunamas, kai apskritai nėra regresijos. Šiuo atveju $\beta_1 = ... = \beta_m = 0$ ir yra tik vienas nežinomas parametras π. Tikėtinumo funkcija maksimizuojama taške $\hat{\pi} = \bar{Y} = \frac{1}{n}\sum Y_i$ ir

$$L_1(\hat{\pi}) = \prod_{i=1}^{n} \bar{Y}^{Y_i} (1 - \bar{Y})^{1-Y_i} \le L(\hat{\boldsymbol{\beta}}) \le L_0(\hat{\boldsymbol{p}}).$$

Jei n didelis ir galioja logistinės regresijos modelis, tai

$$D_E = -2 \ln \frac{L(\hat{\boldsymbol{\beta}})}{L_0(\hat{\boldsymbol{p}})} = -2 \ln L(\hat{\boldsymbol{\beta}})$$

skirstinys aproksimuojamas chi kvadrato skirstiniu su $n - m - 1$ laisvės laipsnių.

Imame pirmąjį ir trečiąjį modelius, turinčius n ir 1 nežinomus parametrus. Atsitiktinio dydžio

$$D_T = -2 \ln \frac{L_1(\hat{\pi})}{L_0(\hat{\boldsymbol{p}})} = -2 \ln L_1(\hat{\pi})$$

dėsnis artimas chi kvadrato dėsniui su $n - 1$ laisvės laipsnių, jei teisingas trečias modelis (t.y. nėra regresijos: $\beta_1 = ... = \beta_m = 0$) ir kai n didelis.

Imame antrąjį ir trečiąjį modelius, turinčius $m+1$ ir 1 nežinomus parametrus. Atsitiktinio dydžio

$$D_R = -2\ln\frac{L_1(\hat{\pi})}{L(\hat{\boldsymbol{\beta}})}$$

skirstinys artimas chi kvadrato skirstiniui su m laisvės laipsnių, jei $\beta_1 = ... = \beta_m = 0$ ir n didelis.

Teisinga lygybė $D_T = D_E + D_R$.

Determinacijos koeficientu logistinės regresijos atveju vadinamas a. d.

$$R^2 = 1 - \frac{D_E}{D_T} = \frac{D_R}{D_T}.$$

Nagrinėkime hipotezę

$$H_0 : \beta_1 = ... = \beta_m = 0.$$

Ši hipotezė reiškia, kad regresijos nėra ir reikšmės $\boldsymbol{x}$ žinojimas nepagerina tikimybės $\pi(\boldsymbol{x})$ prognozės.

Hipotezė H_0 gali būti užrašyta ekvivalenčia forma $H_0 : \pi(\boldsymbol{x}) = \pi = const.$

Kai teisinga hipotezė H_0, a. d. D_E skirstinys aproksimuojamas chi kvadrato skirstiniu su m laisvės laipsnių. Hipotezė H_0 atmetama su reikšmingumo lygmeniu α, jei

$$D_R > \chi^2_\alpha(m).$$

Nagrinėkime hipotezę

$$H_0 : \beta_{j_1} = ... = \beta_{j_l} = 0 \quad (1 \le j_1 < ... < j_l \le m,\ l < m).$$

Pažymėkime $D_R^{(m)}$ ir $D_R^{(m-l)}$ statistiką D_R atitinkamai logistinės regresijos modeliams su visais $\beta_0, ..., \beta_m$ ir be $\beta_{j_1}, ..., \beta_{j_l}$. Kai teisinga hipotezė H_0 atsitiktinio dydžio $D_R^{(m)} - D_R^{(m-k)}$ skirstinys aproksimuojamae chi kvadrato skirstiniu su $k = m - (m-k)$ laisvės laipsnių.

Hipotezė H_0 atmetama reikšmingumo lygmenens α kriterijumi, jei

$$D_R^{(m)} - D_R^{(m-k)} > \chi^2_\alpha(k).$$

Hipotezė

$$H_j : \beta_j = 0\ (j = 1, ..., m)$$

gali taip pat būti tikrinama panaudojant Fišerio informacinės matricos įvertinį.

Jei n yra didelis, tai a. d. $\sqrt{n}(\hat{\beta}_j - \beta_j)$ dėsnis aproksimuojamas normaliuoju dėsniu $N(0, \sigma_{jj})$.

Statistikos

$$W_j = \sqrt{n}\frac{\hat{\beta}_j}{\hat{\sigma}_{jj}}$$

skirstinys aproksimuojamas $N(0,1)$ dėsniu, kai n didelis. Hipotezė $H_0 : \beta_j = 0$ atmetama asimptotiniu reikšmingumo lygmenens α kriterijumi, jei $|W_j| > z_{\alpha/2}$.

(a) *Duomenų įvedimas*

Priklausomo kintąmojo "Linksmumas" ir dviejų nepriklausomų kintamųjų "Lytis" ir "Amzius" atitinkamas reikšmes įvedame į tris stulpelius.

Skaičiavimas

Analyze - Regression - Binary Logistic

Priklausomą kintamąjį "Linksmumas" į Dependent langelį, o nepriklausomus kintamuosius "Lytis" ir "Amzius" į Covariates langelį.

Stanndartinei analizei naudojamas "Enter" metodas.

Nuspauskite Categorical mygtuką. Naujame lange Define Categorical Variables kategorinį kintamąjį "Lytis" į Categorical Covariates langelį,

"Change Contrast" zonoje galima palikti Last opciją arba pakeisti į First opciją. Jei kategorinio kintamojo reikšmių įtaka priklausomam kintamajam lyginama su pirmosios reikšmės įtaka (t.y. "1" (moters)įtaka lyginama su "0" (vyro") įtaka, tai pasirenkame First opciją, jei kategorinio kintamojo reikšmių įtaka priklausomam kintamajam lyginama su paskutinės reikšmės įtaka (t.y. vyro "0" įtaka lyginama su moters "1" įtaka), tai pasirenkama opciją Last. Jei, pavyzdziui, kategorinis kintamasis įgytų tris reikšmes, tai pasirinkus First antrosios ir trečiosios reikšmių įtaka būtų lyginama su pirmosos rikšmės įtaka, o pasirinkus Last pirmosios ir antrsios reikšmių įtaka būtų lyginamos su trečiosios reikšmės įtaka.

Continue.

Nuspauskite Options mygtuką. Zonoje "Statistics and Plots" pažymėkite Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals ir CI for exp(B), o zonoje "Display" pažymėkite At last step.

Continue, OK

Rezultatai

"Model Summary" lentelėje pateikiamos Kokso-Snelo $R^2 = 0,402$ ir Nagelkerkės $R^2 = 0,536$ reikšmės, abu metodai vertina nepriklausomais kintamaisiais paaiškintą priklausomo kintamojo sklaidos dalį.

"Variables in the Equation" lentelėje pateikiami koeficientų $\beta_1, \beta_2, \beta_0$ įverčiai $\hat{\beta}_1 = -0,232$, $\hat{\beta}_2 = 0,123$, $\hat{\beta}_0 = -4,179$, atitinkamų Voldo statistikų reikšmės $W_1 = 0,034$, $W_2 = 4,613$, $W_0 = 3,178$, p-reikšmės hipotezėms $H_j : \beta_j = 0$ tikrinti: $pv_1 = 0,854$, $pv_2 = 0,032$, $pv_0 = 0,075$.

Pateiktos ir šansų santykių $e^{\hat{\beta}_1} = 0,793, e^{\hat{\beta}_2} = 1,131$ reikšmės bei parametrų β_1, β_2 pasikliovimo intervalai.

Duomenys neprieštarauja hipotezei H_1, kuri reiškia, kad linksmumas nepriklauso nuo lyties. Hipotezė H_2 atmetama, bet nelabai stipriai. Taigi linksmumas priklauso nuo amžiaus. Pastebėkime, kad $Y = 1$ reiškė juokimasi, todėl nagrinėjami juokimosi šansai (atžvilgiu žvengimo). Be to, nagrinėjami santykiai tarp moterų šansų ir vyrų šansų (nes 1 koduotos moterys ir buvome pasirinkę opciją First). Nors šansų santykis $e^{\hat{\beta}_1} = 0,793$ yra mažesnis už 1, kas reikštų, kad šansas, jog moteris juokiasi (o ne žvengia) sudaro 0,793 dalį šanso, kad vyras juokiasi, bet skirtumas statistiškai nereikšmingas.

Kadangi $e^{\hat{\beta}_2} = 1,131$, tai padidėjus amžiui vienais metais šansas juoktis, o ne žvengti padidėja 1,131 karto. 0,95 - pasikliovimo intervalas šiam šansų santykiui yra $(1,011; 1,266)$.

Kadangi kintamasis "Lytis" pripažintas nereikšmingu, tai galima pakartoti skaičiavimus išmetus šį nepriklausomą intamąjį.

Perskaičiavus gauname

"Model Summary" lentelėje pateikiamos Kokso-Snelo $R^2 = 0,401$ ir Nagelkerkės $R^2 = 0,536$ reikšmės panašios į buvusias.

Koeficientų β_1, β_0 įverčiai $\hat{\beta}_1 = 0,128$, $\hat{\beta}_0 = -4,460$, atitinkamų Voldo statistikų reikšmės $W_1 = 5,833$, $W_0 = 5,896$, p-reikšmės hipotezėms $H_j : \beta_j = 0$ tikrinti: $pv_1 = 0,016$, $pv_0 = 0,015$.

Hipotezė H_1 atmetama. Taigi linksmumas priklauso nuo amžiaus. Pateiktos ir šansų santykių $e^{\hat{\beta}_1} = 1,136$ reikšmės bei parametr0 β_1 pasikliovimo intervalas.

Kadangi $e^{\hat{\beta}_2} = 1,136$, tai padidėjus amžiui vienais metais šansas juoktis, o ne žvengti padidėja 1,136 karto. 0,95 - pasikliovimo intervalas šiam šansų santykiui yra $(1,024; 1,261)$.

Logistinė regresija gali būti panaudojama ir duomenų klasifikavimui. Jei $\hat{Y}_j \geq 0.5$, SPSS klasifikuoja šį įvykį kaip "teigiamą", jei $\hat{Y}_j < 0.5$„ šį įvykį klasifikuoja kaip "neigiamą". Jei i-jam objektui gaunamas "teigiamas" įvykis, tai šiam objektui priskiriama prognozė $\tilde{Y}_i = 1$, priešingu atveju $\tilde{Y}_i = 0$. Galimi variantai: $Y_i = \tilde{Y}_i = 1$; $Y_i = \tilde{Y}_i = 0$; $Y_i = 1, \tilde{Y}_i = 0$; $Y_i = 0, \tilde{Y}_i = 1$. Pažymėkime $D_{kl} = \sum_{i=1}^{n} \mathbf{1}_{\{Y_i = k, \tilde{Y}_i = l\}}$ skaičių objektų, kuriems $Y = k, \tilde{Y} = l$.

"Classification table" lentelėje pateikiamos dažnių D_{kl} reikšmės. Nagrinėjamu konkrečiu atveju $D_{00} = 10, D_{01} = 1, D_{10} = 2, D_{11} = 9$. Be to, paskutiniame stulpelyje pateikiama $D_{00}/(D_{00} + D_{01}) \times 100\% = 90,9$ – žvengiančių žiūrovų dalis (procentais), kuriai teisingai prognozuojamas žvengimas, $D_{11}/(D_{10} + D_{11}) \times 100\% = 81,8$ – besijuokiančių žiūrovų dalis (procentais), kuriai teisingai prognozuojamas juokimasis, $(D_{00} + D_{11}/n \times 100\% = 86,4$ – visų žiūrovų dalis (procentais), kuriai teisingai prognozuojamas elgesys. Visi rodikliai pakankamai aukšti, taigi pagal amžių galima gerai atskirti besijuokiančius žiūrovus nuo žvengiančių.

Užrašas po lentele "The cut value is .500" nurodo, kad tikimybės lyginamos su 0.5.

(b) Tas pats.

(c) *Duomenų įvedimas*

Duomenis suvedame į penkis stulpelius su tais pačiais pavadinimais kaip uždavinio formuluotėje.

Logistinės regresijos atveju priklausomas kintamasis yra ne pats naujagimio svoris (kurį galima būtų panaudoti tiesinėje regresijoje), o kategorinis kintamasis, kuris įgyja tik dvi reikšmes: 1, kai naujagimio svoris neviršija 2500 g, ir 0, kai jo svoris viršija 2500 g. Pavadinkime jį kategorizuotu naujagimio svoriu: : "Naujagsvkateg". Šio kintamojo reikšmėms apibrėžti atliekame tokias operacijas:

Transform-Recode into different values

Naujagsvor į langą "Numeric variable-Output variable"

Į langelš Output variable name - Naujagsvkateg

Change

Norint galima jam priskirti ir etiketę.

Paspaudžiame "Old and new values". Pažymime "Range, LOWEST through value" ir įveskite 2500.

Prie "New value" įveskite 1. Add. Pažymėkite "All other values" ir prie "New value" įveskite 0.

Continue, OK. Duomenų faile atsiras naujas kintamasis "Naujagsvkateg".

Kadangi prašoma pateikti neišnešioto kūdikio prognozę 38 metų būsimai motinai, kuri rūko, serga hipertonija ir sveria 85 kg, tai į atitinkamus stulpelius dar įvedame 38, 1, 1 ir 85. **Beje, sąlygoje praleista, kad prognozė reikalinga 38 metų moteriai.**

Skaičiavimas

Analyze - Regression - Binary Logistic

Priklausomą kintamąjį "Naujagsvkateg" į Dependent langelį, o nepriklausomus kintamuosius "Motinamz" ir "Rukymas", "Hipertonija", "Motinsvor" į Covariates langelį.

Nuspauskite Categorical mygtuką. Naujame lange Define Categorical Variables kategorinius kintamuosius "Rukymas" ir "Hipertonija" į Categorical Covariates langelį,

"Change Contrast" zonoje Last opciją pakeiskie į First opciją. Bus lyginami neišnešioti kūdikiai atžvilgiu išnešiotų.

Continue

Nuspauskite *Save* ir prie "Predicted values" pažymėkite "Probabilities".

Continue, OK

Rezultatai Duomenų faile atsiras trečias stulpelis PRE_1, kuriame pateikti tikimybių gimti neišnešiotam kūdikiui įvertiniai. 38 metų būsimai motinai, kuri rūko, serga hipertonija ir sveria 85 kg šis įvertinys yra 0,01265, t.y. tokiai moteriai neišnešiotas kūdikis gimsta su labai maža tikimybe. Tai yra todėl, kad motina yra pakankamai solidaus amžiaus.

"Model Summary" lentelėje pateikiamos Kokso-Snelo $R^2 = 0,338$ ir Nagelkerkės $R^2 = 0,458$ reikšmės, abu metodai vertina nepriklausomais kintamaisiais paaiškintą priklausomo kintamojo sklaidos dalį.

"Variables in the Equation" lentelėje pateikiami koeficientų $\beta_1, \beta_2, \beta_3, \beta_4, \beta_0$ įverčiai $\hat{\beta}_1 = -0,307$, $\hat{\beta}_2 = 0,341$, $\hat{\beta}_3 = 1,736$, $\hat{\beta}_4 = -0,008$, $\hat{\beta}_0 = 6,276$, atitinkamų Voldo statistikų reikšmės bei p-reikšmės hipotezėms $H_j : \beta_j = 0$ tikrinti: $pv_1 = 0,022$, $pv_2 = 0,764$, $pv_3 = 0,243$, $pv_4 = 0,792$.

Matome, kad rūkymas, hipertonija ir motinos svoris nėra reikšmingi faktoriai neišnešioto kūdikio gimimui. Tuo tarpu motinos amžius yra reikšmingas faktorius. Padidinus motinos svorį 1kg neišnešioto kūdikio gimimo šansas sumažėja, nes $\hat{\beta}_1 = -0,307$ yra neigiamas ir šansų santykis yra $e^{\hat{\beta}_1} = 0,736$. Metais jaunesnės šansas pagimdyti neišnešiotą kūdikį yra $1/0,736 = 1,36$ karto didesnis.

Modelyje paliekame tiktai motinos amžių kaip nepriklausomą kintąmąjį. Šiuo atveju $\hat{\beta}_1 = -0,279$, $pv_1 = 0,011$. Ir dabar gauname, kad motinos amžius yra reikšmingas faktorius

neišnešioto kūdikio gimimui. Šansų santykis yra $e^{\hat{\beta}_1} = 0,756$. Metais jaunesnės šansas pagimdyti neišnešiotą kūdikį yra $1/0,756 = 1,32$ karto didesnis.

Jei vietoje logistinės regresijos modelio naudojamas tiesinės regresijos modelis naujagimio svorio prognozavimui, tai priklausomas kintamasis yra Naujagsv, o ne Naujagsvkateg (šis apskritai nenaudojamas). Atlikus tiesinę regresinę analizę gauname, kad hipotezė $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ priimama, nes p-reikšmė yra 0,650. Taigi naujagimio svorio negalime reikšmingai prognozuoti pagal motinos amžių, sergamumą hipertonija, rūkymą ir motinos svorį. Apskritai, modelis duoda vidutinio naujagimio svorio įvertinį 38 metų būsimai motinai, kuri rūko, serga hipertonija ir sveria 85 kg: 3060 g., bet hipotezės priėmimas reiškia, kad tai reikšmingai nesiskiria nuo priklausomo kintamojo vidurkio $\hat{Y} = 2783$ g.

12. Klasterinė analizė

Klasteriu vadinsime panašių objektų grupę. Klasterinės analizės tikslas - sugrupuoti objektus į klasterius.

Naudojami įvairūs objektų panašumo matai. Tarkime, kad tiriama n objektų grupė, o i-asis objektas charakterizuojamas kelių požymių, kurių galimų reikšmių sritys yra intervalai, vektoriumi $\mathbf{X}_i = (X_{i1}, ..., X_{ik})^T$. Vienas iš galimų dviejų objektų panašumo matų yra Euklidinis atstumas

$$d(\mathbf{X}_i, \mathbf{X}_j) = ||\mathbf{X}_i - \mathbf{X}_j|| = (\sum_{l=1}^{k}(X_{il} - X_{jl})^2)^{1/2}.$$

Kadangi vektorių $\mathbf{X}_i$ koordinačių matavimo vienetai dažnai yra skirtingi, tai ši matą natūralu naudoti ne patiems vektoriams $\mathbf{X}_i$, bet jų z reikšmėms $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{ik})^T$; čia

$$Z_{il} = (X_{il} - \bar{X}_{.l})/S_l, \quad \bar{X}_{.l} = \frac{1}{n}\sum_{i=1}^{n} X_{il}, \quad S_l^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{il} - \bar{X}_{.l}))^2.$$

Taigi naudojamas atstumas

$$d(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = ||\boldsymbol{Z}_i - \boldsymbol{Z}_j|| = (\sum_{l=1}^{k}(Z_{il} - Z_{jl})^2)^{1/2}.$$

Naudojami ir kitokie atstumai: $\max_l |Z_{il} - Z_{jl}|$ (Čebyšovo atstumas), $\sum_{l=1}^{k}(Z_{il} - Z_{jl})^2$ (Euklido atstumo kvadratas) ir kiti.

Klasifikavimas naudojant hierarchinius jungimo metodus atliekama tokiu būdu.

1) Pradiniai klasteriai yra tiesiog patys objektai. Skaičiuojami atstumai $d_{ij} = d(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ tarp visų klasifikuojamų objektų porų (įskaitant atstumus $d_{ii} = 0$). Gaunama atstumų $n \times n$ matrica $[d_{ij}]$. Nustatome du objektus i_1 ir j_1, kurių atstumas mažiausias. Juos sujungiame į klasterį $U = \{i_1, j_1\}$. Taigi po šio žingsnio yra $n-1$ klasteris: vienas susideda iš dviejų objektų, kiekvienas iš kitų $n-2$ klasteriū susideda iš vieno objekto.

2) Skaičiuojami atstumai tarp klasterio U ir kiekvieno iš vienetinių klasterių. Atstumų matrica jau yra eilės $(n-1) \times (n-1)$. Vėl atrenkama klasterių pora, kuriai atstumas mažiausias. Iš šios poros sudaromas naujas klasteris. Jis gali susidėti iš dviejų arba trijų objektų (trys objektai gaunami tuo atveju, jei mažiausias atstumas yra tarp klasterio U ir kažkurio vieno objekto klasterio). Taigi po šio žingsnio yra $n-2$ klasteriai.

3) Šis procesas tęsiamas tol, kol lieka vienas klasteris, sudarytas iš visų objektų.

Šio proceso schema vaizduojama grafiku, vadinamu *dendrograma*. Vertikalėje nurodomi objektų numeriai, horizontalėje atstumai tarp klasterių. Klasteriai (pradžioje tiesiog objektai) , kurie jungiami į didesnį klasterį, jungiami laužte, o laužtės aukštis yra atstumas tarp sujungtų klasterių.

Atliekant jungimą į klasterius reikia apibrėžti, kaip nustatomas atstumas tarp klasterių U ir V, kai bent viename iš jų yra daugiau negu vienas objektas.

Atstumas tarp klasterių (nusakantis jų jungimo metodą) skaičiuojamas įvairiais metodais:

1) Vienetinės jungties (artimiausio kaimyno, *single linkage, nearest neighbor*) metodas:

$$d(U, V) = \min_{i \in U, j \in V} d(\boldsymbol{Z}_i, \boldsymbol{Z}_j);$$

2) Pilnosios jungties (tolimiausio kaimyno, *complete linkage, furthest neighbor*) metodas:

$$d(U, V) = \max_{i \in U, j \in V} d(\boldsymbol{Z}_i, \boldsymbol{Z}_j);$$

3) Vidutinėss jungties (*between groups or average linkage*) metodas:

$$d(U, V) = \sum_{\boldsymbol{Z}_i \in U} \sum_{\boldsymbol{Z}_j \in V} d(\boldsymbol{Z}_i, \boldsymbol{Z}_j)/(n_U n_V);$$

čia n_U, n_V yra klasterių objektų skaičius;

4) Centrų jungties (*centroid linkage*) metodas:

$$d(U,V) = d(\bar{\boldsymbol{Z}}_U, \bar{\boldsymbol{Z}}_V)/n_V),$$

where $\bar{\boldsymbol{Z}}_U = \sum_{i\in U} \boldsymbol{Z}_i/n_U$, $\bar{\boldsymbol{Z}}_V = \sum_{j\in V} \boldsymbol{Z}_j$;

5) Vordo (*Ward's*) metodas:

$$d(U,V) = ||\bar{\boldsymbol{Z}}_U - \bar{\boldsymbol{Z}}_V||/(1/n_U + 1/n_V).$$

Be hierarchinių klasifikavimo metodų dar naudojami nehierarchiniai metodai. Populiariausias yra

k-vidurkių metodas:

Klasterių skaičius k pasirenkamas iš anksto. Pradžioje atsitiktinai parenkama k objektų ir juos atitinkantys vektoriai imami kaip pradinių klasterių centrai. Tada skaičiuojami visų objektų atstumai iki centrų ir objektas priskiriamas klasteriui U_i, $i = 1, ..., k$, nuo kurio centro tas objektas mažiausiai nutolęs. Taip gaunama k pradinių klasterių.

Kiekviename iš pradinių klasterių perskaičiuojami centrai.

Skaičiuojami kiekvieno objekto atstumai iki klasterių centrų. Jei objektas artimiausias ne savo klasterio centrui, tai jis perkeliamas į tą artimiausią klasterį. Taigi gaunama k naujų klasterių.

Perskaičiuojami šių klasterių centrai ir procedūra kartojama tol, kol nei vienas objektas nebekeliamas į kitus klasterius.

(a) *Duomenų įvedimas*

Pirmame stulpelyje "Apskritis" įvedami apskričių pavadinimai. Kituose stulpeliuose įvedamos kovariančių a_1, a_2, a_4, a_5, a_6 reikšmės.

Skaičiavimas.

Analyze - Classify -Hierarchical Cluster

"Hierarchical cluster analysis" lange:

"Variables": a_1, a_2, a_4, a_5, a_6.

"Label cases by": Apskritis.

"Cluster" langelyje nurodyti, ką norite klasterizuoti: cases (taigi klasterizuojami objektai-apskritys).

"Display" langelyje pažymėkite "statistics" ir "plots".

Statistics:

Pažymėkite "Agglomeration schedule".

"Cluster membership": none

Continue

Plots:

Pažymėkite "Dendrogram".

"Orientation": vertical

Continue

Method:

Nurodykite pirmą kartą, kad norite naudoti Pilnosios jungties (Furthest neighbor) metodą, kitą kartą - Vordo (Ward's) metodą.

"Interval": Squared Euclidian distances

Laukelyje "Transform variables"-"Standartize": pažymėkite "z scores".

Continue, OK

SPSS standartizuoja visus kintamuosius, kad jų vidurkiai būtų 0, o dispersijos 1, taigi suskaičiuoja z reikšmes.

Rezultatai

"Agglomeration Schedule" rodo, kad

1 žingsnyje į klasterį sujungiami 6 ir 8 objektai ir gaunamas klasteris {6,8}. Tarp jų atstumas lygus 0.424 (stulpelis Coefficients). Taigi jau yra 9 klasteriai $\{6,8\}, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{7\}, \{9\}, \{10\}$.

2 žingsnyjee į klasterį sujungiami 4 ir 6 objektai, tarp jų atstumas lygus 1.609. Žiūrint į Stage cluster first appears pastebime, kad 6 (žr. Cluster 2) jau buvo jungiamas į klasterį (su 8), taigi suformuotas vienas klasteris {4,6,8}, kuriame yra trys objektai. Taigi jau yra 8 klasteriai $\{4,6,8\}, \{1\}, \{2\}, \{3\}, \{5\}, \{7\}, \{9\}, \{10\}$.

3 žingsnyje į klasterį sujungiami 1 ir 9 objektai. Tarp jų atstumas lygus 2,304. Šie objektai anksčiau dar nebuvo jungti į klasterius (Stage cluster first appears abiems 0). Taigi jau yra 7 klasteriai $\{4,6,8\}, \{1,9\}, \{2\}, \{3\}, \{5\}, \{7\}, \{10\}$.

4 žingsnyje į klasterį sujungiami 4 ir 7 objektai. Tarp jų atstumas lygus 2,706. 4 (žr. Cluster 1) jau buvo jungiamas į klasterį (su 6 ir 8), taigi suformuotas klasteris $\{4,6,7,8\}$. Jau yra 6 klasteriai $\{4,6,7,8\}, \{1,9\}, \{2\}, \{3\}, \{5\}, \{10\}$.

5 žingsnyje į klasterį sujungiami 2 ir 5 objektai. Tarp jų atstumas lygus 3,410. Šie objektai anksčiau dar nebuvo jungti į klasterius, taigi suformuotas klasteris $\{2,5\}$. Taigi jau turime penkis klasterius $\{4,6,7,8\}, \{1,9\}, \{2,5\}, \{3\}, \{10\}$. Kiti objektai dar su nieko nesujungti.

6 žingsnyje į klasterį sujungiami 1 ir 2 objektai. Tarp jų atstumas lygus 8,099. 1 (žr. Cluster 1) jau buvo jungiamas į klasterį su 9, o 2 jau buvo jungiamas į klasterį su 5, taigi suformuotas klasteris $\{1,2,5,9\}$. Taigi jau turime keturis klasterius $\{4,6,7,8\}, \{1,2,5,9\}, \{3\}, \{10\}$.

7 žingsnyje į klasterį sujungiami 3 ir 10 objektai. Tarp jų atstumas lygus 9,753. Šie objektai anksčiau dar nebuvo jungti į klasterius, taigi suformuotas klasteris $\{3,10\}$.

Taigi jau turime tris klasterius $\{4,6,7,8\}, \{1,2,5,9\}, \{3,10\}$.

8 žingsnyje į klasterį sujungiami 1 ir 4 objektai. Tarp jų atstumas lygus 11,683. 1 (žr. Cluster 1) jau buvo jungiamas į klasterį (su 6 ir 8), taigi suformuotas klasteris $\{1,2,4,5,6,7,8,9\}$. Taigi jau turime du klasterius $\{1,2,4,5,6,7,8,9\}, \{3,10\}$.

9 žingsnyje į klasterį sujungiami 1 ir 3 objektai. Tarp jų atstumas lygus 32,490. 1 Taigi suformuotas vienas klasteris $\{1,2,3,4,5,6,7,8,9,10\}$, kuris susideda i6 vis7 objekt7.

"Vertical Icicle" rodo tą pačią informaciją kaip ir Agglomeration schedule, tiktai nerodo atstumų tarp klasterių:

Žiūrint iš apačios matome, kad pradžioje sudaromi 9 klasteriai, tai yra klasteris $\{8,6\}$, kiti vienetiniai. Kai sudaromi 8 klasteriai, tai yra klasteris $\{8,6,4\}$, kiti vienetiniai. Kai sudaromi 7 klasteriai, tai yra klasteriai $\{8,6,4\}$, $\{9,1\}$, kiti vienetiniai. Kai sudaromi 6 klasteriai, tai yra klasteriai $\{7,8,6,4\}$, $\{9,1\}$, kiti vienetiniai. Kai sudaromi 5 klasteriai, tai yra klasteriai $\{7,8,6,4\}$, $\{5,2\}$, $\{9,1\}$, kiti vienetiniai, ir t.t.

"Dendogram" rodo tą pačią informaciją kaip ir Agglomeration schedule, tiktai grafine forma.

k-vidurkių metodu suskirstysime apskritis į tris klasterius.

Skaičiavimas.

Pradžioje

Analyze - Descriptive statistics - Descriptives

"Descriptives" lange:

"Variables": a_1, a_2, a_4, a_5, a_6.

Pažymėkite "Save standardized values as variables".

OK

Duomenų lentelėje Atsiras stulpeliai Za_1, Za_2, Za_4, Za_5, Za_6, Taigi SPSS standartizuoja visus kintamuosius, kad jų vidurkiai būtų 0, o dispersijos 1, t.y. suskaičiuoja z reikšmes.

Analyze - Classify - k Means Cluster

"k-Means Cluster" lange:

"Variables": Za_1, Za_2, Za_4, Za_5, Za_6.

Save:

Pažymėkite "Cluster membership" ir "Distance from cluster center".

Rezultatai

Lentelėje "Number of Cases in each Cluster" nurodyta po kiek apskričių yra 1,2 ir 3 klasteriuose .

Grįžkime prie duomenų lentelės.

Randame du naujus kintamuosius. QCL_1 nurodo, kuriems iš trijų klasterių priklauso objektai. QCL_2 nurodo objektų atstumus iki klasterių, kuriems jie priskirti, centrų.