

Kokybinių duomenų analizė

Praktinis įvadas

Vytautas Kazakevičius

2013 m. rugsėjo 5 d.

Turinys

1 Įvadas	1
2 Vieno kintamojo dažnių lentelės	2
2.1 Bendrosios žinios	2
2.2 Binominis modelis	3
2.3 Polinominis modelis	5
3 Dviejų kintamųjų dažnių lentelės	8
3.1 Bendrosios žinios	8
3.2 Nepriklausomumo tikrinimas	9
3.3 Priklausomybės stiprumo vertinimas	10
3.4 Sutapimo kriterijai	12
4 Dažnių lentelių rinkiniai	12
4.1 Bendrosios žinios	12
4.2 Sąlyginio nepriklausomumo tikrinimas	13
4.3 2×2 lentelių rinkiniai	15
5 Logistinė regresija	15
5.1 Bendrosios žinios	15
5.2 Parametrų vertinimas	16
5.3 Hipotezių tikrinimas	16
5.4 Klasifikavimas	16

1 Įvadas

Kursą derėtų vadinti *Kokybinių duomenų analizė*. „Kokybinių duomenų“ reiškia „kokybinių kintamųjų“.

Apskritai statistikoje mus domina koks nors kintamasis y ; pagrindinis klausimas — kaip jis priklauso nuo įvairių faktorių, x_1 , x_2 ir t.t. Kokybinių duomenų analizė nagrinėja situaciją, kai y kintamasis kokybinis.

Priklausomai nuo faktorių skaičiaus ir jų tipo visa kokybinių duomenų analizė išsiskaido į 4 dalis.

1. Jei nėra nė vieno faktoriaus, nagrinėjamos *vieno kintamojo dažnių lentelės*.

2. Jei yra vienas faktorius ir jis kokybinis, nagrinėjamos *dvių kintamųjų dažnių lentelės*.
3. Jei yra daugiau faktorių ir visi jie kokybiniai, nagrinėjami *dažnių lentelių rinkiniai*.
4. Jei bent vienas faktorius kiekybinis, taikoma *logistinė regresija*.

2 Vieno kintamojo dažnių lentelės

2.1 Bendrosios žinios

Kaip jau minėjau, analizuojamas kintamasis bus žymimas y , o $y_1, \dots, y_n$ bus stebėtos jo reikšmės (vektorius $(y_1, \dots, y_n)$ tada vadinamas *imtimi*, o n yra *imties dydis*). Laikysiu, kad galimos y reikšmės yra $1, 2, \dots, s$. Tada pradiniai duomenys gali būti sutraukti į tokią *y kintamojo dažnių lentelę*:

y	1	2	...	s	
	n_1	n_2	...	n_s	n

Čia n_k yra k reikšmės pasikartojimų skaičius imtyje.

Kaip paprastai statistikoje, laikysiu, kad turima imtis yra tam tikros *atsitiktinės imties* $(Y_1, \dots, Y_n)$ realizacija. $Y_1, \dots, Y_n$ yra nepriklausomos atsitiktinio dydžio Y kopijos, o Y skirstinys aprašomas lentele

Y	1	2	...	s	
	p_1	p_2	...	p_s	1

Žymėsiu

$$p = (p_1, \dots, p_s);$$

tai — *nežinomų modelio parametrų vektorius*.

Irodoma, kad p vektoriaus didžiausio tikėtinumo įvertis aprašomas lygybėmis

$$\hat{p} = (\hat{p}_1, \dots, \hat{p}_s), \quad \hat{p}_k = \frac{n_k}{n} \text{ su } k = 1, \dots, s.$$

Atitinkamos atsitiktinės imties statistikos (parametrų maksimalaus tikėtinumo *įvertiniai*) žymimi taip pat:

$$\hat{p}_k = \frac{N_k}{n};$$

čia N_k yra k reikšmės pasikartojimų skaičius imtyje $(Y_1, \dots, Y_n)$.

Atsiskaitant prie kompiuterio, aš duodu du uždavinius — „praktinį“ ir „teorinį“. Praktiniame uždavinyje prašau panagrinėti duomenis iš **b1994** lentelės. Šioje lentelėje yra 1994 metais atliktos Lietuvos gyventojų apklausos duomenys. Buvo apklausta virš 1000 gyventojų; kiekvieną atitinka viena duomenų lentelės eilutė. Kiekvienas respondentas turėjo atsakyti į beveik 200 klausimų; kiekvieną klausimą atitinka vienas lentelės stulpelis (kintamasis).

Kintamųjų vardai prasideda **a** raide, po kurios parašytas atitinkamo klausimo numeris; pavyzdžiui, **a6**, **a145** ir t.t. Visi atsakymai užkoduoti skaičiais. Tikrieji anketos klausimai ir galimi atsakymai į juos surašyti **b1994_info.doc** faile.

Dauguma kintamųjų kokybiniai. Kiekybiniai kintamieji yra arba vertinimai (sveikatos būklės, šalies ekonominės padėties ir pan.; vertinama balais nuo -5 iki 5), arba parodo, kokią pajamų dalį respondentas skiria maistui, butui, drabužiams ir pan. Dar vienas kiekybinis kintamasis yra **gimimo metai**. Jis ypatingas tuo, kad nurodyti tik du paskutiniai metų skaitmenys. Taigi reikšmė 54 reiškia, kad respondentas gimė 1954 metais. Šį kintamąjį galima panaudoti apskaičiuojant respondento amžių: kadangi apklausa atlikta 1994 metais, reikia iš 94 atimti kintamojo reikšmę. Pavyzdžiui, $94 - 54 = 40$.

2.2 Binominis modelis

Jei y įgyja tik dvi reikšmes (t.y. $s = 2$), sakoma, kad nagrinėjamas *binominis modelis*. Binominiame modelyje formuluojami dviejų rūšių statistiniai uždaviniai:

- įvertinti kokią nors skirstinio charakteristiką $g(p)$;
- patikrinti kokią nors hipotezę apie nežinomą parametą.

Praktinio uždavinio pavyzdys. (b1994 lentelė) Su 90% patikimumu įvertinkite, kokia dalis vilniečių tikėjosi geresnių arba žymiai geresnių 1995 metų (6, 151).

Skliaustuose prieš sąlygą aš primenu, kad reikės naudoti duomenis iš b1994 lentelės. Skliaustuose po sąlygos surašau kintamųjų, kuriuos reikės panaudoti analizėje, numerius. Atvėręs b1994_info.doc failą randu tokią informaciją:

a6 Ko tikėtės iš 1995 metų?

1. Bus žymiai geresni.
2. Bus geresni.
3. Bus tokie pat.
4. Bus blogesni.
5. Bus žymiai blogesni.

a151 Kur gyvenate?

1. Vilniuje.
2. Kaune.
3. ...

Uždaviniui atlikti rašoma, pavyzdžiui, tokia SAS programa.

```
data duom;
  set b1994;
  if a151=1;
  if a6=. then tikejosi=.;
  else if a6<=2 then tikejosi=1;
  else tikejosi=2;
  keep a6 tikejosi;
run;

proc freq data=duom;
  tables tikejosi/ alpha=0.1 binomial(level=1);
run;
```

Dar vienas praktinis uždavinys. (b1994 lentelė) Ar galima tvirtinti, kad tik ketvirtadalis vyrų savo šeimos materialinę padėtį vertina geriau nei kaimyno? (145, 20)

b1994_info.doc faile randu:

a145 Lytis

1. Vyras.
2. Moteris.

a20 Palyginkite savo šeimos materialinę padėtį su kaimyno.

1. Žymiai geresnė.
2. Šiek tiek geresnė.
3. Tokia pat.
4. Šiek tiek blogesnė.
5. Žymiai blogesnė.

Uždaviniui atlikti tiktų tokia SAS programa.

```
data duom;
  set b1994;
  if a145=1;
  if a20=. then vertina=.;
  else if a20<=2 then vertina=1;
  else vertina=2;
  keep a20 vertina;
run;

proc freq data=duom;
  tables vertina/ binomial(p=0.25 level=1);
run;
```

Teorinio uždavinio pavyzdys. Teorija sako, kad jei binominiame modelyje n didelis, tai

$$\sqrt{n}(\hat{p}_1 - p_1) \approx N(0, p_1(1 - p_1)).$$

Iš čia išplaukia, kad daug kartų skaičiuojant tikimybės įvertį

$$\frac{1}{k} \sum_{j=1}^k n(\hat{p}_1^{(j)} - p_1)^2 \approx p_1(1 - p_1);$$

čia k yra įverčių skaičiavimo skaičius, o $\hat{p}_1^{(j)}$ — suskaičiuotas įvertis j -ojoje imtyje. O kaip bus, jei n mažas? Paimkite, pavyzdžiui, $p_1 = 0.3$ ir $n = 10$.

Čia tiktų tokia SAS programa.

```
data duom;
  do imtis=1 to 500;
    do i=1 to 10;
      u=ranuni(0);
      if u<=0.3 then y=1;
      else y=2;
      w=1;
    end;
  end;
```

```

        output;
    end;
    y=1; w=0;
    output;
    y=2; w=0;
    output;
end;
run;

proc freq data=duom noprint;
    by imtis;
    tables y/ binomial;
    output out=rez binomial;
run;

data anal;
    set rez end=galas;
    suma+(10*(_bin_-0.3)*(_bin_-0.3)/500);
    if galas then put suma;
run;

```

Ką dar reikia žinoti? Jei vertinama ne kuri nors tikimybė, o išvestinis parametras, pavyzdžiui, $g(p_1)$, yra dvi galimybės.

- Jei g funkcija monotoniška ir $(a; b)$ yra p_1 tikimybės pasiklovimo intervalas (jį duoda SAS), tai $g(p_1)$ parametro pasiklovimo intervalas yra arba $(g(a); g(b))$ (kai g didėjanti), arba $(g(b); g(a))$ (kai g mažėjanti).
- Bendruoju atveju pasiklovimo intervalas konstruojamas remiantis asimptotiniu sąryšiu

$$\sqrt{n}(g(\hat{p}) - g(p)) \xrightarrow{d} N(0, p(1-p)).$$

Iš jo išplaukia, kad $1 - \alpha$ lygmens asimptotinis pasiklovimo intervalas yra $(\hat{p} - \delta; \hat{p} + \delta)$ su

$$\delta = z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = z_{\alpha} ASE(\hat{p});$$

čia z_{α} yra standartinio normalaus skirstinio α lygmens kritinė reikšmė. Dydis $ASE(\hat{p})$ vadinamas įvertinio *asimptotinė standartine paklaida*; jį duoda SAS. Kritinė reikšmė z_{α} sutampa su $1 - \alpha$ lygmens kvantiliu, kurį skaičiuoja SAS funkcija *probit*.

Teoriniuose uždaviniuose reiktų žinoti ir pagrindines statistikos sąvokas: kriterijaus lygmuo, pirmos rūšies klaidos tikimybė, galia. Turėti supratimą apie tiksluosius kriterijus.

2.3 Polinominis modelis

Jei y įgyja daugiau nei dvi reikšmes (t.y. $s > 2$), sakoma, kad nagrinėjamas *polinominis modelis*. Polinominiame modelyje formuluojami tie patys statistiniai uždaviniai:

- įvertinti kokią nors skirstinio charakteristiką $g(p)$;
- patikrinti kokią nors hipotezę apie nežinomą parametą.

Praktinio uždavinio pavyzdys. (b1994 lentelė) 78% Lietuvos žmonių dėl sunkios šalies padėties kaltina valdžią, 6% — save. Ar Kaune padėtis kitokia? (25, 151).

b1994_info.doc faile randu

```
a25 Kas kaltas dėl dabartinės sunkios šalies padėties?
      1 Ankstesnė vadovybė
      2 Dabartinė valdžia
      3 Ir tie, ir tie
      4 Mes patys
      5 Nežinau
      6 Kiti
```

Taigi tinka tokia SAS programa.

```
data duom;
  set b1994;
  if a151=2;
  if a25=. then kaltas=" ";
  else if a25<=3 then kaltas="valdzia";
  else if a25=4 then kaltas="mes";
  else kaltas="kiti";
  keep a25 kaltas;
run;

proc freq data=duom;
  tables kaltas/ testp=(78,6,16);
run;
```

Teorinio uždavinio pavyzdys. Tikslieji kriterijai hipotezei $p = p^*$ tikrinti yra konservatyvūs, t.y. jų pirmos rūšies klaidos tikimybė yra mažesnė negu kriterijaus lygmuo. Patikrinkite, kiek konservatyvus yra 5% lygmens tikslus kriterijus, kai $n = 7$ ir $p^* = (0.1, 0.6, 0.3)$.

```
data duom;
  do imtis=1 to 500;
    do i=1 to 7;
      u=ranuni(0);
      if u<0.1 then y=1;
      else if u<0.7 then y=2;
      else y=3;
      w=1;
      output;
    end;
    do i=1 to 3;
      y=1; w=0;
      output;
    end;
  end;
```

```

        end;
    end;
run;

proc freq data=duom noprint;
    by imtis;
    weight w/zeros;
    tables y/ testp=(0.1,0.6,0.3);
    exact chisq;
    output out=rez chisq;
run;

data anal;
    set rez end=galas;
    if (xp_pchi>0.05) then atmete=1;
    else atmete=0;
    suma+(atmete/500);
    if galas then put atmete=;
run;

```

Ką dar reikia žinoti. Jei reikia įvertinti išvestinį parametą $g(p)$, remiamės tuo, kad

$$\sqrt{n}(g(\hat{p}) - g(p)) \xrightarrow{d} N(0, \sigma^2);$$

čia

$$\sigma^2 = g'(p)\Sigma(p)g'(p)^\top, \quad g'(p) = \begin{pmatrix} \frac{\partial g}{\partial p_1}(p) & \cdots & \frac{\partial g}{\partial p_s}(p) \end{pmatrix}$$

ir

$$\Sigma(p) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_s \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_s \\ \vdots & \vdots & \ddots & \vdots \\ -p_sp_1 & -p_sp_2 & \cdots & p_s(1-p_s) \end{pmatrix}.$$

Taigi jei n didelis, tai su $\approx 1 - \alpha$ tikimybe

$$g(\hat{p}) - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \leq g(p) \leq g(\hat{p}) + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}};$$

čia $\hat{\sigma} = g'(\hat{p})\Sigma(\hat{p})g'(\hat{p})^\top$.

Dar vienas teorinis uždavinys. Panagrinėkime asimptotinio 95% lygmens pasiklovimo intervalą p_1/p_2 santykiui polinominiame modelyje. Koks tikrasis jo patikimumas, kai n nedidelis? Imkite, pavyzdžiui, $n = 10$ ir $p = (0.3, 0.6, 0.1)$.

```

data duom;
    do imtis=1 to 500;
        do i=1 to 10;
            u=ranuni(0);
            if u<0.3 then y=1;
            else if u<0.9 then y=2;
            else y=3;
        end;
    end;
run;

```

```

        w=1;
        output;
    end;
    do i=1 to 3;
        y=i;
        w=0;
        output;
    end;
end;
run;

proc freq data=duom noprint;
    by imtis;
    weight w/zeros;
    tables y/ out=rez;
run;

data rez2;
    set rez;
    retain p1 p2;
    if y=1 then p1=percent/100;
    else if y=2 then p2=percent/100;
    else do;
        p3=percent/100;
        output;
    end;
    keep imtis p1 p2 p3;
run;

data anal;
    set rez2 end=galas;
    g1=1/p2;
    g2=-p1/(p2*p2);
    sigma11=p1*(1-p1);
    sigma12=-p1*p2;
    sigma22=p2*(1-p2);
    sigma=sigma11*g1*g1+2*sigma12*g1*g2+sigma22*g2*g2;
    delta=probit(0.975)*sqrt(sigma/10);
    if (p1/p2-delta<=0.5 & 0.5<=p1/p2+delta) then pateko=1;
    else pateko=0;
    suma+pateko/500;
    if galas then put suma=;
run;

```

3 Dviejų kintamųjų dažnių lentelės

3.1 Bendrosios žinios

Kaip ir anksčiau, analizuojamąjį kintamąjį žymėsiu y , bet dabar dar nagrinėsiu ir kitą kokybinį kintamąjį x ; $(x_1, y_1), \dots, (x_n, y_n)$ bus stebėtos tų kintamųjų

reikšmės. Laikysiu, kad galimos x reikšmės yra $1, \dots, r$, o galimos y reikšmės — $1, 2, \dots, s$. Tada pradiniai duomenys gali būti sutraukti į tokią x ir y kintamųjų dažnių lentelę:

$x \ y$	1	2	$\dots$	s	
1	n_{11}	n_{12}	$\dots$	n_{1s}	n_{1+}
2	n_{21}	n_{22}	$\dots$	n_{2s}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\dots$	n_{rs}	n_{r+}
	n_{+1}	n_{+2}	$\dots$	n_{+s}	n

Čia n_{jk} yra (j, k) poros pasikartojimų skaičius imtyje.

Kaip ir anksčiau laikysiu, kad turima imtis yra tam tikros *atsitiktinės imties* $(X_1, Y_1), \dots, (X_n, Y_n)$ realizacija. (X_i, Y_i) yra nepriklausomos atsitiktinio vektoriaus (X, Y) kopijos, o to vektoriaus skirstinys aprašomas lentele

$X \ Y$	1	2	$\dots$	s	
1	p_{11}	p_{12}	$\dots$	p_{1s}	p_{1+}
2	p_{21}	p_{22}	$\dots$	p_{2s}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	p_{r1}	p_{r2}	$\dots$	p_{rs}	p_{r+}
	p_{+1}	p_{+2}	$\dots$	p_{+s}	1

Žymėsiu

$$p = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1s} \\ p_{21} & p_{22} & \dots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \dots & p_{rs} \end{pmatrix};$$

tai — *nežinomų modelio parametrų matrica*.

Irodoma, kad p matricos didžiausio tikėtino įvertis aprašomas lygybėmis

$$\hat{p} = (\hat{p}_{jk}), \quad \hat{p}_{jk} = \frac{n_{jk}}{n} \text{ su } j = 1, \dots, r, k = 1, \dots, s.$$

Atitinkamos atsitiktinės imties statistikos (parametrų maksimalaus tikėtino įvertiniai) žymimi taip pat:

$$\hat{p}_{jk} = \frac{N_{jk}}{n};$$

čia N_{jk} yra (j, k) poros pasikartojimų skaičius atsitiktinėje imtyje. Be to, su $j = 1, \dots, r$ ir $k = 1, \dots, s$ žymėsiu

$$N_{j+} = \sum_{k=1}^s N_{jk}, \quad N_{+k} = \sum_{j=1}^r N_{jk}.$$

3.2 Nepriklausomumo tikrinimas

Vienas pagrindinių uždavinių — hipotezės apie nepriklausomumą tikrinimas. Standartinė SAS sintaksė tokia:

```
proc freq data=duom;
  tables x*y/ chisq;
run;
```

Rezultatų lape žiūrima Q_P , Q_{MH} arba Q_{LR} .
Jei reikia tikslaus Fišerio testo, rašoma arba

```
tables x*y/ fisher;
```

arba

```
proc freq data=duom;
  tables x*y;
  exact fisher;
run;
```

Kiti tikslieji kriterijai iškviečiami sakiniu

```
exact chisq;
```

3.3 Priklausomybės stiprumo vertinimas

Pirsono koreliacija. Nors Pirsono koreliacijos koeficientas paprastai skaičiuojamas tik tarp kiekybinių kintamųjų, jis naudojamas ir kokybinių duomenų analizėje. Griežtas pagrindimas — teorinėje dalyje, o kol kas — tik bendra rekomendacija: abu kintamieji turi būti „beveik kiekybiniai“. Dichotominis kintamasis yra laikomas „beveik kiekybiniu“.

SAS procedūros sintaksė, kai reikia tik įvertinti koreliacijos koeficientą:

```
proc freq data=duom;
  tables x*y/ measures cl;
run;
```

Interpretuodami rezultatus žiūrime į koeficiento ženklą ir dydį.

Taip pat gali tekti patikrinti nekoreliuotumo arba nepriklausomumo hipotezę. Pirmoji tikrinama žiūrint į pasikliautiną intervalą arba naudojant papildomą sakinį

```
proc freq data=duom;
  tables x*y/ measures cl;
  test pcorr;
run;
```

Antroji — pagal Q_{MH} reikšmę.

Gama koeficientas. Skaičiuojamas, kai abu kintamieji bent jau ordinalūs. „Beveik kiekybiniai“ kintamieji laikomi ordinaliais. Interpretacija tokia pat, kaip Pirsono koreliacijos koeficiento.

Skaičiavimas:

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d};$$

čia

$$\Pi_c = \sum_{j,k} p_{jk} p_{jk}^{(c)}, \quad \Pi_d = \sum_{j,k} p_{jk} p_{jk}^{(d)},$$

$$p_{jk}^{(c)} = \sum_{\substack{j' > j \\ k' > k}} p_{j'k'} + \sum_{\substack{j' < j \\ k' < k}} p_{j'k'}, \quad p_{jk}^{(d)} = \sum_{\substack{j' > j \\ k' < k}} p_{j'k'} + \sum_{\substack{j' < j \\ k' > k}} p_{j'k'}.$$

SAS procedūros sintaksė, kai reikia tik įvertinti koeficientą:

```
proc freq data=duom;
  tables x*y/ measures cl;
run;
```

Interpretuodami rezultatus žiūrime į koeficiento ženklą ir dydį.

Jei reikia patikrinti nepriklausomumo hipotezę, naudojamas papildomas sakinyss

```
proc freq data=duom;
  tables x*y/ measures cl;
  test gamma;
run;
```

Neapibrėžtumo koeficientas. Skaičiuojamas, kai abu kintamieji nominalūs.

Formulės:

$$U_{y|x} = \frac{H_x + H_y - H_{xy}}{H_y};$$

čia

$$H_x = - \sum_j p_{j+} \ln p_{j+}, \quad H_y = - \sum_k p_{+k} \ln p_{+k}, \quad H_{xy} = - \sum_{j,k} p_{jk} \ln p_{jk}.$$

SAS procedūros sintaksė, kai reikia tik įvertinti koeficientą:

```
proc freq data=duom;
  tables x*y/ measures cl;
run;
```

Interpretuodami rezultatus žiūrime į koeficiento dydį.

Jei reikia patikrinti nepriklausomumo hipotezę, žiūrime į Q_{LR} .

Priklausomybės stiprumo matai 2×2 lentelėms. Tegu

$$p_1 = P(Y = 1 | X = 1) = \frac{p_{11}}{p_{1+}}, \quad p_2 = P(Y = 1 | X = 2) = \frac{p_{21}}{p_{2+}}.$$

Tada gali dominti:

$$d = p_1 - p_2, \quad r = \frac{p_1}{p_2}, \quad \text{or} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}.$$

Jei reikia įvertinti:

```
proc freq data=duom;
  tables x*y/riskdiff relrisk;
run;
```

3.4 Sutapimo kriterijai

Čia situacija šiek tiek kitokia: abu nagrinėjami kintamieji yra analizuojamieji, tik vienas stebėtas „prieš“, kitas — „po“. Todėl teorijoje vietoje X ir Y rašysiu Y ir Y' . Kiti žymenys išlieka (tik $r = s$).

Domina, ar skirstinys „po“ toks pat, kaip skirstinys „prieš“, o jei pasikeitė, tai kiek. Pasikeitimo dydis matuojamas kapa koeficientu, kuris skaičiuojamas taip:

$$\kappa = \frac{p^0 - p^e}{1 - p^e};$$

čia

$$p^0 = \sum_k p_{kk}, \quad p^e = \sum_k p_{k+} p_{+k}.$$

SAS procedūros sintaksė:

```
proc freq data=duom;
  tables x*y/agree;
run;
```

4 Dažnių lentelių rinkiniai

4.1 Bendrosios žinios

Dabar aprašysiu sprendžiamus uždavinius, kai nagrinėjama kintamojo y priklausomybė nuo kelių kokybinių kintamųjų. Vienas iš tų kintamųjų tada laikomas „pagrindiniu“, jį žymėsiu x raide. Kiti kintamieji pakeičiami vienu kokybiniu kintamuoju, kurį žymėsiu z — kiekvieną tų kintamųjų reikšmių kombinaciją atitinka tam tikra z reikšmė. Laikysiu, kad galimos x reikšmės yra $1, \dots, r$, galimos y reikšmės — $1, 2, \dots, s$ ir galimos z reikšmės — $1, \dots, L$.

$(z_1, x_1, y_1), \dots, (z_n, x_n, y_n)$ bus stebėtos kintamųjų reikšmės, arba imtis. Imties dalį, kurioje z kintamojo reikšmės lygios kažkuriam l , vadinsiu *sluoksniu* (taigi imtį sudaro L sluoksnių). Kiekvieno sluoksnio duomenis galima sutraukti į x ir y kintamųjų dažnių lentelę:

$x \ y$	1	2	...	s	
1	n_{l11}	n_{l12}	...	n_{l1s}	n_{l1+}
2	n_{l21}	n_{l22}	...	n_{l2s}	n_{l2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{lr1}	n_{lr2}	...	n_{lrs}	n_{lr+}
	n_{l+1}	n_{l+2}	...	n_{l+s}	n_{l++}

Čia n_{ljk} yra (j, k) poros pasikartojimų skaičius sluoksnyje $z = l$.

Kaip ir anksčiau laikysiu, kad turima imtis yra tam tikros *atsitiktinės imties* $(Z_1, X_1, Y_1), \dots, (Z_n, X_n, Y_n)$ realizacija. (Z_i, X_i, Y_i) yra nepriklausomos atsitiktinio vektoriaus (Z, X, Y) kopijos, o to vektoriaus skirstinys aprašomas nurodant tikimybes

$$p_{ljk} = P\{Z = l, X = j, Y = k\}.$$

Šeima

$$p = (p_{ljk} \mid 1 \leq l \leq L, 1 \leq j \leq r, 1 \leq k \leq s)$$

yra nežinomų parametrų vektorius.

Taip pat žymėsiu

$$\begin{aligned} p_{lj+} &= \sum_k p_{ljk}, & p_{l+k} &= \sum_j p_{ljk}, & p_{+jk} &= \sum_l p_{ljk}, \\ p_{l++} &= \sum_{j,k} p_{ljk}, & p_{+j+} &= \sum_{l,k} p_{ljk}, & p_{++k} &= \sum_{lj} p_{ljk}, \\ p_{jk|l} &= \frac{p_{ljk}}{p_{l++}}, & p_{k|lj} &= \frac{p_{ljk}}{p_{lj+}}. \end{aligned}$$

Aišku, kad, pavyzdžiui,

$$\begin{aligned} p_{+jk} &= P\{X = j, Y = k\}, & p_{l++} &= P\{Z = l\}, \\ p_{jk|l} &= P(X = j, Y = k \mid Z = l), & p_{k|lj} &= P(Y = k \mid Z = l, X = j). \end{aligned}$$

ir pan.

(Z, X, Y) skirstinys dažnai aprašomas, nurodant Z reikšmių tikimybes p_{l++} ir (X, Y) vektoriaus sąlygines tikimybes $p_{jk|l}$; tada

$$p_{ljk} = p_{l++} p_{jk|l}.$$

Irodoma, kad p vektoriaus didžiausio tikėtinumo įvertis aprašomas lygybėmis

$$\hat{p} = (\hat{p}_{ljk}), \quad \hat{p}_{ljk} = \frac{n_{ljk}}{n}.$$

Atitinkamos atsitiktinės imties statistikos (parametrų maksimalaus tikėtinumo *įvertiniai*) žymimos taip pat:

$$\hat{p}_{ljk} = \frac{N_{ljk}}{n};$$

čia N_{ljk} yra (l, j, k) poros pasikartojimų skaičius atsitiktinėje imtyje. Be to, žymėsiu

$$\begin{aligned} N_{lj+} &= \sum_k N_{ljk}, & N_{l+k} &= \sum_j N_{ljk}, & N_{+jk} &= \sum_l N_{ljk}, \\ N_{l++} &= \sum_{j,k} N_{ljk}, & N_{+j+} &= \sum_{l,k} N_{ljk}, & N_{++k} &= \sum_{lj} N_{ljk}. \end{aligned}$$

4.2 Sąlyginio nepriklausomumo tikrinimas

Vienas iš standartinių uždavinių nagrinėjant dažnių lentelių rinkinius — patikrinti hipotezę, kad kiekviename sluoksnyje X ir Y yra (sąlygiškai) nepriklausomi. Tiksliau ši hipotezė formuluojama taip: $p \in M$; čia

$$M = \{p \in \Theta \mid \forall l, j, k \ p_{ljk}/p_{l++} = p_{lj+}/p_{l++} \cdot p_{l+k}/p_{l++}\}$$

Dirbant su SAS, ši hipotezė patikrinama parašius tokią programą.

```
proc freq data=duom;
  tables z*x*y/chm;
run;
```

Sąlyginio nepriklausomumo hipotezės nereikia painioti su besąlyginio nepriklausomumo hipoteze, atititinkančia aibę

$$M' = \{p \in \Theta \mid \forall j, k \ p_{+jk} = p_{+j+}p_{++k}\}.$$

Apskritai neteisingas nei $M' \subset M$, nei $M \subset M'$ sąryšis.

Tegu, pavyzdžiui, Z įgyja dvi reikšmes su vienodomis tikimybėmis, jei $Z = 1$, X, Y nepriklausomi ir įgyja reikšmes 1 ir 2 su tikimybe $1/2$, o jei $Z = 2$, tai X, Y nepriklausomi ir įgyja tas pačias reikšmes su tikimybėmis $1/4$ ir $3/4$, atitinkamai. Aišku, kad tada $p \in M$. Kita vertus,

$$p_{+11} = \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{16} = \frac{5}{32}, \quad p_{+1+} = p_{++1} = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{8}.$$

Kadangi $\frac{5}{32} \neq \frac{3}{8} \cdot \frac{3}{8}$, $p \notin M'$.

O dabar tegu Z įgyja dvi reikšmes su vienodomis tikimybėmis, o sąlyginiai (X, Y) poros skirstiniai aprašomi lentelėmis

$X \ Y$	1	2
1	$1/2$	0
2	0	$1/2$

$X \ Y$	1	2
1	0	$1/2$
2	$1/2$	0

Aišku, kad X ir Y nėra sąlygiškai nepriklausomi; todėl $p \notin M$. Kita vertus, besąlyginis (X, Y) skirstinys aprašomas lentele

$X \ Y$	1	2
1	$1/4$	$1/4$
2	$1/4$	$1/4$

iš kurios matyti, kad abu dydžiai nepriklausomi ir įgyja reikšmes 1 ir 2 su vienodomis tikimybėmis. Taigi $p \in M'$.

Jeigu besąlyginį nepriklausomumą reiktų tikrinti su SAS, rašytume tokią programą:

```
proc freq data=duom;
  tables x*y/chisq;
run;
```

Sąlyginio nepriklausomumo uždavinio nereikia maišyti ir su uždaviniu, kuris sprendžiamas parašius tokią SAS programą:

```
proc freq data=duom;
  tables x*y/chisq;
  by z;
run;
```

Šiuo atveju tikrinamas sąlyginis nepriklausomumas, bet tikrinama ne viena hipotezė $p \in M$, o L hipotezių $p \in M_l$ ($l = 1, \dots, L$); čia

$$M_l = \{p \in \Theta \mid \forall j, k \ p_{ljk}/p_{l++} = p_{lj+}/p_{l++} \cdot p_{l+k}/p_{l++}\}.$$

4.3 2×2 lentelių rinkiniai

Jei nagrinėjamos 2×2 lentelių rinkiniai (t.y. $r = s = 2$), galima spręsti dar du uždavinius: patikrinti hipotezę, kad visuose sluoksniuose $Y = 1$ įvykio šansų santykiai vienodi ir, jei vienodi, įvertinti tą bendrą šansų santykio reikšmę.

Įvykio šansais l -ajame sluoksnyje vadinami santykiai $p_{1|lj}/p_{2|lj}$ ($j = 1, 2$), o tų šansų santykiu — dydis

$$or_l = \frac{p_{1|l1}/p_{2|l1}}{p_{1|l2}/p_{2|l2}} = \frac{p_{l11}/p_{l12}}{p_{l21}/p_{l22}} = \frac{p_{l11}p_{l22}}{p_{l12}p_{l21}}.$$

Taigi hipotezė užrašoma pavidalu $p \in M$ su

$$M = \{p \in \Theta \mid or_1 = \dots = or_L\}.$$

Jei hipotezė teisinga, tai

$$or_1 = \dots = or_L = \theta$$

su tam tikru $\theta > 0$ ir galima vertinti tą nežinomą parametą θ .

Sprendžiant šiuos uždavinius su SAS, rašoma ta pati programa

```
proc freq data=duom;
  tables z*x*y/cmh;
run;
```

5 Logistinė regresija

5.1 Bendrosios žinios

Dabar išsiaiškinsime, kokie uždaviniai sprendžiami, kai mus domina y kintamojo priklausomybė nuo kelių kintamųjų, iš kurių bent vienas yra kiekybinis.

Standartinė programa, rašoma analizuojant duomenis su SAS, atrodo taip:

```
proc logistic data=duom;
  class a b;
  model y= a b u v;
run;
```

Čia a ir b yra kokybiniai, o u ir v — kiekybiniai kintamieji.

Prieš pradėdamas analizę, SAS kiekvieną kokybinį kintamąjį transformuoja į keletą kiekybinių kintamųjų, vadinamų *pseudokintamaisiais* (angliškai jie vadinami *design variables*). Pavyzdžiui, jei a kintamasis įgyja tris reikšmes 1, 2 ir 3, bus sukurti du nauji kintamieji $a1$ ir $a2$, o jų reikšmės bus suskaičiuotos pagal tokią lentelę:

a	a1	a2
1	1	0
2	0	1
3	-1	-1

Kitaip tariant, jei a reikšmė yra 1, $a1$ reikšmė bus 1, o $a2$ reikšmė bus 2; jei a bus lygus 2, $a1$ bus 0, o $a2$ bus 1; jei a lygus 3, abu pseudokintamieji lygūs -1 .

Schema, pagal kurią skaičiuojamos pseudokintamųjų reikšmės, vadinama *parametrizacija*. Panorus ją galima pakeisti. Pavyzdžiui, jei **class** sakinyse atrodytų taip

```
class a (param=ref);
```

pseudokintamieji būtų kuriami pagal tokią schemą:

a	a1	a2
1	1	0
2	0	1
3	0	0

Numatytoji **param** parinktės reikšmė yra **param=effect**; ją atitinka aukščiau aprašyta schema.

Abiejose aprašytose parametrizacijose trečioji **a** kintamojo reikšmė yra ypatinga; ji vadinama *bazinė reikšmė*. Norėdami galime bazinė padaryti kitą reikšmę. Pavyzdžiui, jei **class** sakinyse atrodytų taip

```
class a (param=effect, ref="2");
```

bazinė būtų reikšmė 2 ir pseudokintamieji būtų skaičiuojami pagal tokią schemą:

a	a1	a3
1	1	0
2	-1	-1
3	0	1

Atkreipiu dėmesį, kad keistųsi ir pseudokintamųjų vardai. Apskritai jie sudaromi prie kokybinio kintamojo vardo prirašant iš dešinės atitinkamą jo reikšmę. Bazinės reikšmės neatitinka joks pseudokintamasis.

5.2 Parametrų vertinimas

5.3 Hipotezių tikrinimas

5.4 Klasifikavimas