Prilagodba vjerojatnosnog modela opaženim frekvencijama i χ^2-test

Snježana Lubura[1] i Miljenko Huzak[2]

1. Uvod

U statistici su dva problema osnovna. Jedan od njih je kako na osnovi uzorka prikupljenog opažanjem vrijednosti neke statističke varijable procijeniti njenu populacijsku distribuciju, a drugi je problem kako iz uzorka ocijeniti je li populacijska distribucija jednaka nekoj pretpostavljenoj razdiobi. Ti su problemi usko povezani budući da se u primjenama često pretpostavlja da populacijska distribucija pripada nekoj užoj klasi razdioba koje su parametrizirane nekim parametrima nepoznatih vrijednosti. Kažemo da je zadan parametarski model za populacijsku distribuciju izučavane varijable. U tom slučaju prvo se prilagođava pretpostavljeni model opaženim podacima, a zatim se sprovodi statistički test kojim se ocjenjuje je li model prihvatljiv (reprezentativan) za populacijsku distribuciju.

Prilagodba modela podacima svodi se na procjenu vrijednosti populacijskih parametara korištenjem nekog statističkog kriterija (ili metode) procjene. U ovome članku koristit ćemo dvije metode procjene, u stručnoj literaturi poznate kao *metoda maksimalne vjerodostojnosti i metoda mimimuma* χ^2 („*hi-kvadrata*"). Pokazuje se da u kontekstu statističkog modela za podatke opisanog u sljedećem poglavlju, bilo koja od tih metoda daje *procjenitelje* jednakih graničnih razdioba kada veličina uzorka teži u neizmjerno. U tom slučaju se kaže da su dobiveni procjenitelji *asimptotski ekvivalentni*.

Statistički test za ocjenu prihvatljivosti pretpostavljenog modela, koji ćemo opisati u članku, poznat je pod imenom *Pearsonov* χ^2*-test* ili, jednostavno, χ^2*-test* („*hi-kvadrat-test*"). U uskoj je vezi s metodom minimuma χ^2 jer je minimalna vrijednost kriterijske funkcije za nalaženje procjenitelja ujedno i testna statistika za taj test.

Svrha ovog članka je dvostruka. S jedne strane namjera nam je opisati primjenu χ^2-metode u procjeni i testiranju parametarskih modela s posebnim naglaskom na diskusiju određenih praktičnih rješenja koja se često koriste.

[1] Snježana Lubura, dipl. ing., PMF-Matematički odsjek, Zagreb, snjezana.lubura@math.hr
[2] doc. dr. sc. Miljenko Huzak, PMF-Matematički odsjek, Zagreb, huzak@math.hr

S druge strane, navedenu temu želimo učiniti prihvatljivom i s metodičke strane kako bi se izloženi sadržaj mogao učinkovito približiti učenicima kroz nastavu statistike već u srednjoj školi. U vezi s time pretpostavljamo da su sljedeći statistički pojmovi poznati: uzorak, statistika, statistička varijabla, populacijska i empirijska distribucija, statistički test i opći pojmovi vezani uz njega. Također, pretpostavljamo da su poznati i sljedeći pojmovi vezani uz vjerojatnosnu distribuciju: funkcija vjerojatnosti, funkcija distribucije i gustoća slučajne varijable. Novouvedeni i u članku obrađeni statistički pojmovi i rezultati naznačeni su kurzivom (italic) kroz tekst.

Problem procjene parametara usko je vezan uz matematički problem nalaženja ekstrema neke funkcije. Za važne primjere navedene u članku dovoljno je poznavati *Shannon-Kolmogorovljevu informacijsku nejednakost*. U izvodu te nejednakosti koristi se geometrijska interpretacija konveksnosti funkcije kroz Jensenovu nejednakost. Općenito, za nalaženje ekstrema funkcije trebalo bi poznavati pojam derivacije i stacionarne točke, te vezu između točke ekstrema derivabilne funkcije i stacionarnosti, ali ni to nije presudno ukoliko učenici na raspolaganju imaju primjerenu programsku podršku uz koju mogu dovoljno precizno nacrtati graf funkcije i iz njega očitati traženu točku ekstrema. Iako bi se statistički problem prilagodbe vjerojatnosnog modela podacima mogao iskoristiti kao motivacija za temu nalaženja ekstrema funkcije, u članku nismo slijedili taj put.

Struktura članka je sljedeća. U sljedećem (drugom) poglavlju kroz primjere uvodimo i opisujemo χ^2-metodu. Primjenu te metode na probleme testiranja nezavisnosti dviju diskretnih varijabli i homogenosti distribucija obrađujemo u trećem i četvrtom poglavlju. U petom poglavlju diskutiramo primjenu χ^2 -testa na testiranje modela neprekidnih razdioba.

2. Pearsonov χ^2-test

Pokažimo prvo na jednostavnom primjeru kako ćemo pomoću Pearsonovog χ^2 -testa testirati da li podaci, do kojih smo došli nezavisnim mjerenjem neke diskretne slučajne veličine, predstavljaju realizaciju slučajnog uzorka iz zadane konkretne populacijske distribucije.

Primjer 1. Igraća kocka baca se 300 puta. Pri tome se svako bacanje provodi pod istim uvjetima i na slučajan način. Zapisujemo broj strane koja se okrenula na kocki. Dolazimo do sljedećih rezultata: 53 puta okrenuo se broj 1, 43 puta broj 2, 58 puta broj 3, 41 puta broj 4, 52 puta broj 5 i 53 puta broj 6. Te *opažene frekvencije* možemo zapisati u tablici:

x_i	1	2	3	4	5	6
n_i	53	43	58	41	52	53

Tablica 1: Rezultat bacanja kocke iz Primjera 1.

Ovdje su x_i, $i = 1, 2, \ldots, 6$ mogući elementarni ishodi pri pokusu bacanja igraće kocke i ujedno vrijednosti diskretne varijable X = „broj koji se okrenuo", a n_i, $i = 1, 2, \ldots, 6$ su njihove opažene (izmjerene) frekvencije.

Zanima nas je li korištena kocka simetrična. Kada bi bila simetrična, onda bi vjerojatnost da se okrene bilo koja od strana, označena brojevima 1, 2, 3, 4, 5 ili 6, bila ista i iznosila bi 1/6. Označimo s π_i vjerojatnost da se prilikom bacanja kocke okrenula strana s brojem i, $i \in \{1,2,3,4,5,6\}$. Dakle, ako je kocka simetrična, onda je $p_i = 1/6$ za svaki $i \in \{1,2,3,4,5,6\}$. Na taj način u potpunosti je zadana distribucija slučajne varijable X.

Simetričnost kocke provjerit ćemo tako da testiramo nul-hipotezu

$$H_0 : p_i = \frac{1}{6}, i \in \{1,2,3,4,5,6\};$$

nasuprot alternativnoj hipotezi da kocka nije simetrična, tj.

$$H_1 : p_i \neq \frac{1}{6}, \text{za barem jedan } i.$$

Kako testirati nul-hipotezu H_0?

Primijetimo da će nakon n bacanja kocke *očekivana frekvencija* ishoda x_i, biti jednaka np_i, $i \in \{1,2,3,4,5,6\}$ ako je n vrlo velik broj. Primijetimo da je u ovome primjeru $n = 300$. Razumno je pretpostaviti da tada *udaljenost* između opaženih i očekivanih frekvencija nije velika. Tu udaljenost mjerimo pomoću *Pearsonove* χ^2 *-statistike*:

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i}. \tag{1}$$

Za razliku od euklidske udaljenosti vektora opaženih $(n_1, \ldots, n_6)$ i očekivanih $(np_1, \ldots, np_6)$ frekvencija, udaljenost komponenti tih vektora u χ^2 -statistici je normalizirana, s očekivanim vrijednostima odgovarajućih frekvencija, kako bi se ujednačili doprinosi vrijednosti x_i s velikim i onih s malim očekivanim frekvencijama.

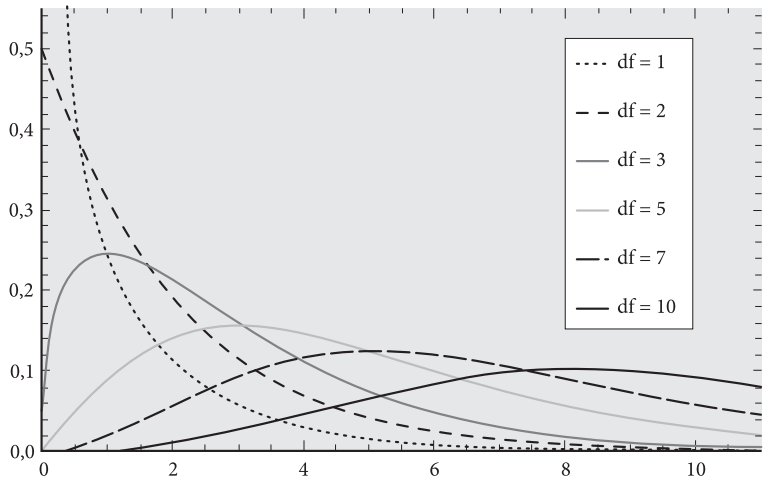
Na osnovi vrijednosti Pearsonove χ^2 -statistike, uz pretpostavku istinitosti nul-hipoteze i na osnovi opaženih podataka, donosi se odluka o odbacivanju nul-hipoteze ili o njenom neodbacivanju u korist alternativne hipoteze. Ako je nul-hipoteza H_0 točna, tada je očekivana frekvencija svakog od elementarnih ishoda jednaka $n \cdot 1/6 = 300 \cdot 1/6 = 50$. Za podatke iz tablice, vrijednost χ^2 -statistike tada je jednaka:

$$\chi^2 = \frac{(53-50)^2}{50} + \frac{(43-50)^2}{50} + \frac{(58-50)^2}{50} + \frac{(41-50)^2}{50} + \frac{(52-50)^2}{50} + \frac{(53-50)^2}{50} =$$

$$= \frac{9+49+64+81+4+9}{50} = \frac{216}{50} = 4.32.$$

Idući korak je odlučiti odbacujemo li nul-hipotezu H_0 ili ne. H_0 ćemo odbaciti ako je vrijednost testne χ^2 -statistike veća od neke unaprijed zadane tzv. kritične vrijednosti. Naime, velika vrijednost χ^2 -statistike znači da je udaljenost opaženih od pretpostavljenih očekivanih frekvencija velika, a to upućuje na zaključak da treba odbaciti hipotezu H_0 kojom su takve očekivane frekvencije pretpostavljene. Kritičnu vrijednost određujemo na osnovi vlastite odluke koliko veliku pogrešku odbacivanja H_0, a da je ona u stvarnosti istinita (to je tzv. pogreška prve vrste), možemo tolerirati. Ako dopustimo da je ta pogreška najviše 5%, onda kažemo da testiramo na razini značajnosti od 5%. To znači da je kritična vrijednost jednaka broju za koji vrijedi da je 5% svih mogućih vrijednosti χ^2 -statistike, uz istinitost hipoteze H_0 veće od tog broja. Očito, što je razina značajnosti manja, kritična vrijednost je veća, pa ćemo teže odbaciti nul-hipotezu. S druge strane, u tom slučaju povećat će se vjerojatnost pogreške druge vrste.

Da bismo za zadanu razinu značajnosti odredili kritičnu vrijednost, trebamo poznavati zakon razdiobe slučajne varijable χ^2 (tzv. uzoračku razdiobu χ^2 -statistike), uz pretpostavku istinitosti nul-hipoteze. Ne postoji općeniti rezultat o tome koja je to razdioba. Nije jednostavno odrediti tu razdiobu ni za χ^2 -statistiku u ovome našem konkretnom primjeru. Ipak, iz statističke teorije poznato je da za velike uzorke (dakle u graničnom slučaju, kada duljina uzorka teži u neizmjerno), te uz pretpostavku da vrijedi nul-hipoteza, Pearsonova χ^2 -statistika ima asimptotski χ^2 -*distribuciju s* $r - 1$ *stupnjem slobode*, gdje je r broj razreda[3]. χ^2 -distribucija je neprekidna[4]. Grafovi njenih gustoća za razne stupnjeve slobode prikazani su na Slici 1.



Slika 1: Gustoće χ^2 *-razdioba s df stupnjeva slobode, za razne df.*

[3] Taj se rezultat pripisuje Karlu Pearsonu (vidjeti [6]) koji ga je objavio u svom radu [5] iz 1900. Suvremeni dokaz tog teorema može se naći u [7], str. 499.-500.).

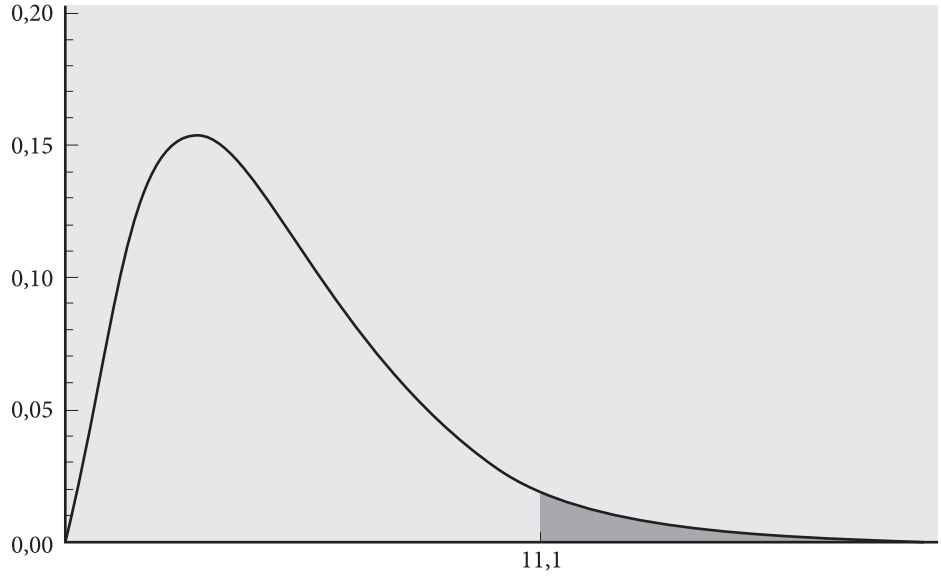
[4] Gustoća χ^2 -distribucije s m stupnjeva slobode jednaka je (Γ je oznaka za gama-funkciju):

$$f(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{2^{-m/2}}{\Gamma(m/2)} \cdot x^{(m/2)-1} e^{-x/2}, & x > 0. \end{cases}$$

Slikom 2. pokazan je grafički položaj kritične vrijednosti za zadanu razinu značajnosti. Inače, kritičnu vrijednost možemo odrediti pomoću tablica u kojima su tabelirani kvantili ili funkcija distribucije χ^2-razdiobe. Ta vrijednost može se i preciznije odrediti pomoću statističkih ili matematičkih programskih paketa (na primjer: R, Mathematica, MATLAB,...).[5]

Vratimo se na naš primjer. Budući da je broj razreda r jednak broju mogućih ishoda bacanja $r = 6$ i kritičnu vrijednost tražimo u odnosu na χ^2-distribuciju s $r - 1 = 5$ stupnjeva slobode. Za razinu značajnosti od $\alpha = 5\%$, kritična vrijednost, u oznaci $\chi^2_{0.05}(5)$, jednaka je 11.1. Budući da je $\chi^2 = 4.32 \leq 11.1$, ne odbacujemo H_0, odnosno pretpostavku o simetričnosti igraće kocke, na razini značajnosti od 5%.

Sada se možemo pitati što ako ne možemo pretpostaviti da je kocka simetrična? Naime, iako u prethodnom primjeru nismo odbacili pretpostavku o simetričnosti kocke (na zadanoj razini značajnosti od 5%), ne znači da je ta pretpostavka točna. Još uvijek postoji mogućnost da smo napravili pogrešku druge vrste: prihvaćamo H_0, a da je alternativna hipoteza H_1 istinita. U tom slučaju prirodno je postaviti pitanje: kako procijeniti pravu razdiobu od X? Da bismo pokušali odgovoriti na njega, vratimo se problemu i podacima iz Primjera 1.



Slika 2: Položaj kritične vrijednosti $\chi^2_{0.05}(5) = 11.1$ za razinu značajnosti od 5% u odnosu na graf gustoće χ^2-razdiobe s 5 stupnjeva slobode. Zasjenjena površina ispod grafa gustoće, a iznad intervala $[11.1, +\infty\rangle$, jednaka je razini značajnosti 0.05.

[5] www.r-project.org (R), www.wolfram.com (Mathematica), www.mathworks.com (MATLAB)

Primjer 2. Za igraću kocku, podatke i varijablu X iz Primjera 1., procijenimo zakon razdiobe

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{pmatrix} \tag{2}$$

od X.

Kako procijeniti populacijsku distribuciju pomoću podataka?

Jedan od načina je da se nađe takva vjerojatnosna distribucija koja je najmanje udaljena od frekvencijske distribucije podataka (tzv. empirijske distribucije). Udaljenost između vjerojatnosne distribucije $(p_1, p_2, \ldots, p_6)$ i empirijske distribucije $\left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_6}{n}\right)$ mjerimo pomoću χ^2-funkcije (1). Prikažimo je kao funkciju od vektora $(p_1, p_2, \ldots, p_6)$ za čije komponente vrijedi da su pozitivne i da im je zbroj jednak jedan. Dakle,

$$\chi^2 : P \to \mathbb{R}, \quad \chi^2(p_1, p_2, \ldots, p_6) = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{6} \frac{\left(\frac{n_i}{n} - p_i\right)^2}{p_i}, \tag{3}$$

gdje je

$$P = \{(p_1, p_2, \ldots, p_6) : (\forall i \in \{1, 2, \ldots, 6\})(p_i > 0) \,\&\, \sum_{i=1}^{6} p_i = 1\}.$$

Primijetite da je za podatke iz Tablice 1. empirijska distribucija u P. Minimalna vrijednost ovako definirane χ^2-funkcije jednaka je 0 i postiže se u jedinstvenom vektoru koji je jednak empirijskoj distribuciji. Dakle, procjena populacijske distribucije za podatke iz Tablice 1. jednaka je:

$$(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_6) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_6}{n}\right) = \left(\frac{53}{50}, \frac{43}{50}, \frac{58}{50}, \frac{41}{50}, \frac{52}{50}, \frac{53}{50}\right). \tag{4}$$

Kažemo da je procjena (4) dobivena *metodom minimuma* χ^2.

Alternativna metoda procjenjivanja populacijske distribucije (2) je *metoda maksimalne vjerodostojnosti*. Da bismo tu metodu mogli primijeniti, potrebno je odrediti zakon razdiobe opaženih frekvencija na osnovi pretpostavki o populacijskoj distribuciji mjerene varijable X i definicije slučajnog uzorka. Zbog toga se takva razdioba opažanih frekvencija zove *uzoračka razdioba*. Označimo s $(N_1, N_2, \ldots, N_6)$ slučajni vektor kojemu su opažene frekvencije $(n_1, n_2, \ldots, n_6)$ = (53, 43, 58, 41, 52, 53) jedna realizacija. Dakle, taj slučajni vektor predstavlja *slučajne* opažene frekvencije. Iz prirode slučajnog eksperimenta kojim je biran uzorak slijedi da je uzoračka razdioba od $\mathrm{N} = (N_1, N_2, \ldots, N_6)$ *polinomijalna* za čiju gustoću[6] f vrijedi da je

$$f(k_1, k_2, \ldots, k_6) = \mathbb{P}(N_1 = k_1, N_2 = k_2, \ldots, N_6 = k_6) = \frac{n!}{\prod_{i=1}^{6} k_i!} \prod_{i=1}^{6} p_i^{k_i}. \tag{5}$$

[6] Detaljnija rasprava o gustoćama diskretnih slučajnih vektora, te posebno o polinomijalnoj razdiobi, može se naći u [7], str. 130.-134.

za nenegativne cijele brojeve $k_1, k_2, \ldots, k_6$ takve da je ispunjen uvjet $k_1 + k_2 + \cdots + k_6 = n$. U suprotnom, $f(k_1, k_2, \ldots, k_6)$ jednako je 0. Kraće kažemo da slučajni vektor $\mathbf{N}$ ima polinomijalnu distribuciju s parametrima $(n; p_1, p_2, \ldots, p_6)$. Pomoću vrijednosti funkcije gustoće za tu razdiobu u opaženim vrijednostima frekvencija $(n_1, n_2, \ldots, n_6)$ za svaki vektor $(p_1, p_2, \ldots, p_6)$ iz P definiramo *vjerodostojnost* tog vektora kao broj

$$L(p_1, p_2, \ldots, p_6) := f(n_1, n_2, \ldots, n_6) = \frac{n!}{\prod_{i=1}^{6} n_i!} \prod_{i=1}^{6} p_i^{n_i}.$$

Na taj način dobro je definirana funkcija $L : P \to \mathbb{R}$ koju zovemo *vjerodostojnost* za vjerojatnosni model (2). Sada smo u mogućnosti definirati *procjenu* od $(p_1, p_2, \ldots, p_6)$ *maksimalne vjerodostojnosti* kao točku $(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_6)$ iz P, za koju vjerodostojnost L poprima maksimalnu vrijednost:

$$L(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_6) = \max_{(p_1, \ldots, p_6) \in P} L(p_1, p_2, \ldots, p_6).$$

Ekvivalentno, procjenitelj maksimalne vjerodostojnosti je točka u kojoj se postiže maksimum logaritma vjerodostojnosti. Drugim riječima, neka je s

$$\ell : P \to \mathbb{R}, \quad \ell(p_1, p_2, \ldots, p_6) := \ln L(p_1, p_2, \ldots, p_6) = \sum_{i=1}^{6} n_i \ln p_i + \ln\left(\frac{n!}{\prod_{i=1}^{6} n_i!} \right), \quad (6)$$

definirana *log-vjerodostojnost* modela (2). Tada je $(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_6)$ iz P procjena maksimalne vjerodostojnosti od $(p_1, p_2, \ldots, p_6)$ ako i samo ako je

$$\ell(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_6) = \max_{(p_1, \ldots, p_6) \in P} \ell(p_1, p_2, \ldots, p_6). \quad (7)$$

Primijetimo da je dovoljno naći točku maksimuma funkcije

$$(p_1, p_2, \ldots, p_6) \mapsto \sum_{i=1}^{6} n_i \ln p_i \quad (8)$$

na skupu P jer konstanta $\ln\left(\frac{n!}{\prod_{i=1}^{6} n_i!} \right)$, za koju se ta funkcija razlikuje od log-vjerodostojnosti, ne ovisi o vektoru $(p_1, p_2, \ldots, p_6)$. Točku maksimuma te funkcije naći ćemo pomoću Shannon-Kolmogorovljeve informacijske nejednakosti[7].

Lema. (*Shannon-Kolmogorovljeva informacijska nejednakost*) *Neka su* $(p_1, p_2, \ldots, p_6)$ *i* $(q_1, q_2, \ldots, q_r)$ *dva vektora pozitivnih realnih brojeva takvih da je* $\sum_{i=1}^{r} p_i = 1$ *i* $\sum_{i=1}^{r} q_i = 1$.

[7] U ovome članku koristimo diskretnu verziju te nejednakosti. Opća verzija može se naći u [1], str. 113.

Tada vrijedi nejednakost

$$\sum_{i=1}^{r} q_i \ln\left(\frac{q_i}{p_i}\right) \geq 0,$$

pri čemu jednakost vrijedi ako i samo ako je $(p_1, p_2, \ldots, p_6) = (q_1, q_2, \ldots, q_r)$.

Dokaz. Budući da je funkcija $x \mapsto -\ln x$ strogo konveksna, prema definiciji stroge konveksnosti[8] primijenjene na konveksne kombinacije točaka $x_i = q_i / p_i$, i = 1, 2, ..., r i konveksne kombinacije točaka ln x_i i = 1, 2, ..., r s težinama q_i, i = 1, 2, ..., r, slijedi da je

$$\sum_{i=1}^{r} q_i \ln\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^{r} q_i \cdot \left(-\ln\left(\frac{p_i}{q_i}\right)\right) \geq -\ln\left(\sum_{i=1}^{r} q_i \cdot \frac{p_i}{q_i}\right) = -\ln\left(\sum_{i=1}^{r} p_i\right) = -\ln 1 = 0,$$

pri čemu jednakost vrijedi ako i samo ako se sve točke x_i, i = 1, 2, ..., r podudaraju, odnosno, ako i samo ako je za sve i, j, $p_i / q_i = p_j / q_j$, odakle slijedi tvrdnja leme.

Primijenimo Shannon-Kolmogorovljevu informacijsku nejednakost na proizvoljan vektor $(p_1, p_2, \ldots, p_6) \in P$ i vektor $(q_1, q_2, \ldots, q_6) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_6}{n}\right)$:

$$\sum_{i=1}^{6} \frac{n_i}{n} \cdot \ln\left(\frac{n_i / n}{p_i}\right) \geq 0 \Rightarrow \sum_{i=1}^{6} n_i \ln\left(\frac{n_i}{n}\right) \geq \sum_{i=1}^{6} n_i \ln p_i.$$

Odavde i iz Leme slijedi da je točka

$$(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_6) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_6}{n}\right) = \left(\frac{53}{50}, \frac{43}{50}, \frac{58}{50}, \frac{41}{50}, \frac{52}{50}, \frac{53}{50}\right) \tag{9}$$

jedinstvena točka maksimuma funkcije (8), a time i (jedinstvena) procjena maksimalne vjerodostojnosti. Primijetite da se u ovome primjeru procjene metodama χ^2 i maksimalne vjerodostojnosti podudaraju.

U Primjeru 1. testirali smo je li populacijska razdioba od *X* jednaka pretpostavljenoj razdiobi. Zbog toga smo u nul-hipotezi specificirali 6 – 1 = 5 brojeva.

[8] Realna funkcija *f* definirana na konveksnome podskupu realnih brojeva *strogo je konveksna* ako za svaki *r*, sve točke $x_1, \ldots, x_r$ iz domene od *f*, i sve nenegativne brojeve $q_1, \ldots, q_r$, takve da je $\sum_{i=1}^{r} q_i = 1$, vrijedi

$$\sum_{i=1}^{r} q_i f(x_i) \geq f\left(\sum_{i=1}^{r} q_i x_i\right),$$

pri čemu jednakost vrijedi ako i samo ako su sve točke x_i, $i = 1, \ldots, r$ jednake.

Naime, budući da je zbroj svih vjerojatnosti p_i, $i = 1, ..., 6$ jednak 1, dovoljno je zadati samo 5 brojeva, šesti broj tada je u potpunosti određen njima. Primijetite da to znači da je prava dimenzija konveksnog skupa P jednaka 5. Na taj smo način u potpunosti zadali zakon razdiobe od X. Pri tome ništa nismo morali procjenjivati iz uzorka. S druge strane, u Primjeru 2. procjenjivali smo svih 5 vjerojatnosti p_i, $i = 1, ..., 5$ (procjena za p_6 je, dakako, iz istih razloga kao prije u potpunosti određena s ostalih pet procjena). Budući da pri tome nismo ništa pretpostavljali o populacijskoj razdiobi, dakle nismo postavljali nikakvu nul-hipotezu, statističko testiranje bilo je bespredmetno. Primijetite još da su χ^2-udaljenosti (3) dobivenih procjena (po oba navedena i sprovedena kriterija) u odnosu na opaženu frekvencijsku distribuciju jednake nuli, dakle najmanje moguće.

Poopćimo sada pristupe i metode koje smo diskutirali u prvom i drugom primjeru. Pretpostavit ćemo da želimo testirati nul-hipotezu kojom se pretpostavlja da populacijska razdioba varijable koju izučavamo ovisi o nekim parametrima nepoznatih vrijednosti, ali na poznati način, pri čemu je broj parametara manji od dimenzije odgovarajućeg prostora P, vektora vjerojatnosti. U tom slučaju prvo će trebati prilagoditi pretpostavljeni parametarski model podacima tako da nepoznate parametre procijenimo jednom od navedenih metoda procjene, a zatim sprovedemo odgovarajući test nul-hipoteze. Potrebna asimptotska razdioba testne statistike slijedit će iz *Fisherovog teorema*.

Neka je dana slučajna veličina X s vrijednostima u konačnom skupu $A = \{a_1, a_2, ..., a_r\}$. Nadalje, označimo s

$$\pi_i = \mathbb{P}(X = a_i), i = 1, 2, \ldots, r$$

pravu populacijsku vjerojatnost događaja $\{X = a_i\}$. Htjeli bismo testirati nul-hipotezu da te populacijske vjerojatnosti ovise o parametru $\theta \in \Theta \subseteq \mathbb{R}^v$, u skladu s nekim pretpostavljenim parametarskim modelom, tj.

$$\pi_i = p_i(\theta), i = 1, 2, \ldots, r.$$

Dakako, ovdje nužno pretpostavljamo da vektori $(p_1(\theta), p_2(\theta), \ldots, p_r(\theta))$ pripadaju skupu

$$P = \{(\pi_1, \pi_2, \ldots, \pi_r) : (\forall i \in \{1, 2, \ldots, r\})(\pi_i > 0) \,\&\, \sum_{i=1}^{r} \pi_i = 1\}, \qquad (10)$$

za svaku vrijednost vektora parametara θ iz skupa svih mogućih vrijednosti koji smo označili s Θ i za koji pretpostavljamo da je dimenzije v. Još pretpostavljamo da različitim vrijednostima parametara θ odgovaraju različiti vektori u P. Dakako da je tada $0 \le v \le r - 1$. Naime, prava dimenzija prostora P jednaka je $\rho = r - 1$. Slučaj kada je $v = 0$ ilustriran je u Primjeru 1., a slučaj kada bi bio $v = r - 1$ ilustriran je u Primjeru 2. i njega ovdje nećemo uzeti u obzir. Dakle, zadano je injektivno preslikavanje s Θ u P:

$$\theta = (\theta_1, \theta_2, \ldots, \theta_v) \mapsto (p_1(\theta), p_2(\theta), \ldots, p_r(\theta)), \tag{11}$$

koje zovemo *parametarski model* za populacijsku distribuciju od *X*. U skladu s time, želimo testirati nul-hipotezu

$$H_0 : \pi_i = p_i(\theta), \text{ za svaki } i = 1, 2, \ldots, r \text{ i neki } \theta \in \Theta,$$

nasuprot alternativnoj hipotezi

$$H_1 : \pi_i \neq p_i(\theta), \text{ za barem jedan } i = 1, 2, \ldots, r \text{ za sve } \theta \in \Theta.$$

Primijetite da je H_0 složena hipoteza jer njome nije jednoznačno određena populacijska razdioba od *X*.

Neka je $X_1, \ldots, X_n$ slučajni uzorak[9] za *X*, te neka su $N_i = \sum_{j=1}^{n} \mathbf{1}_{\{X_j = a_i\}}$, i = 1, 2, ..., r empirijske (odnosno opažene) frekvencije[10] vrijednosti od *X*, shvaćene kao slučajne varijable. Kao i do sada, realizaciju varijable N_i označavat ćemo s n_i, za i = 1, 2, ..., r. Osnovna pretpostavka na kojoj se bazira testiranje jest da slučajni vektor $\mathbf{N} = (N_1, N_2, \ldots, N_r)$ ima polinomijalnu distribuciju. Uz pretpostavku da je ispunjena hipoteza H_0, parametri te distribucije su $(n; p_1(\theta), p_2(\theta), \ldots, p_r(\theta))$ za neki $\theta \in \Theta$.

Definirajmo pripadnu Pearsonovu χ^2 -testnu statistiku:

$$\chi^2(\theta) = \sum_{i=1}^{r} \frac{(N_i - np_i(\theta))^2}{np_i(\theta)}. \tag{12}$$

Ovu statistiku možemo shvatiti kao funkciju parametra θ definiranu na skupu Θ. Primijetite da je njena interpretacija i dalje udaljenost između vektora opaženih frekvencija $(N_1, N_2, \ldots, N_r)$ i očekivanih frekvencija u skladu s nul-hipotezom H_0, $(np_1(\theta), \ldots, np_r(\theta))$. Nadalje, za zadanu realizaciju $(n_1, n_2, \ldots, n_r)$ od $\mathbf{N}$, vrijednost od $\chi^2(\theta)$ i dalje je funkcija od θ. Za koji bismo θ iz skupa Θ trebali izračunati $\chi^2(\theta)$? Ako to pitanje postavimo na način da se zapitamo: za koji $\theta \in \Theta$ je vektor očekivanih frekvencija (u skladu s H_0!), $(np_1(\theta), \ldots, np_r(\theta))$, najmanje udaljen od vektora opaženih frekvencija $(n_1, n_2, \ldots, n_r)$, onda dolazimo do tražene procjene $\hat{\theta}$ iz Θ koja zadovoljava uvjet:

$$\chi^2(\hat{\theta}) = \min_{\theta \in \Theta} \chi^2(\theta), \tag{13}$$

---

[9] Za niz $X_1, X_2, \ldots, X_n$ kažemo da je *slučajni uzorak za X* ako je to niz nezavisnih slučajnih varijabli koje su sve jednako distribuirane i distribucija im je jednaka populacijskoj distribuciji od *X*. Za $1 \leq i \leq n$, jednostavna interpretacija varijable X_i je da predstavlja *i*-to mjerenje (ili opažanje) varijable *X* u nizu od *n* nezavisnih, po uvjetima izvođenja jednakih slučajnih pokusa.

[10] U definiciji opažene frekvencije N_i pojavljuju se *indikatorske* ili *karakteristične funkcije* događaja $\{X_j = a_i\}$ za j = 1, ..., *n*. Indikatorska funkcija događaja ***A***, u oznaci $\mathbf{1}_A$, je funkcija $\mathbf{1}_A : \Omega \to \{0,1\}$ za koju vrijedi da je $\mathbf{1}_A(\omega) = 1$ ako i samo ako je $\omega \in A$.

dakle, *procjena minimuma* χ^2. U tom slučaju je $\chi^2(\theta)$ vrijednost testne statistike za naš test. Interpretacija od $\chi^2(\hat{\theta})$ jest da je to χ^2-udaljenost opaženih frekvencija od familije svih vjerojatnosnih distribucija od X pretpostavljenih modelom u hipotezi H_0, dakle, od familije

$$P_0 := \{(p_1(\theta), p_2(\theta), \ldots, p_r(\theta)) : \theta \in \Theta\}, \tag{14}$$

koja je i ujedno prava podfamilija od P, familije svih mogućih razdioba od X.

S druge strane, ako nam je primarni cilj prilagodba modela (11) opaženim podacima, tada je uobičajeno koristiti metodu maksimalne vjerodostojnosti. Budući da opažene frekvencije $(N_1, N_2, \ldots, N_r)$ imaju polinomijalnu distribuciju, vjerodostojnost pretpostavljenog modela je funkcija

$$L : \Theta \to \mathbb{R}, \quad L(\theta) = \frac{n!}{\prod_{i=1}^{r} n_i!} \prod_{i=1}^{r} p_i^{n_i}(\theta). \tag{15}$$

Točka $\tilde{\theta}$ iz Θ u kojoj vjerodostojnost $L(\theta)$ poprima maksimalnu vrijednost je *procjena maksimalne vjerodostojnosti*. Ona je ujedno i točka maksimuma funkcije koja je, do na konstantu koja ne ovisi o θ, logaritam vjerodostojnosti:

$$\ell : \Theta \to \mathbb{R}, \quad \ell(\theta) = \sum_{i=1}^{r} n_i \ln p_i(\theta). \tag{16}$$

Dakle, procjena maksimalne vjerodostojnosti $\tilde{\theta}$ zadovoljava uvjet:

$$\ell(\hat{\theta}) = \max_{\theta \in \Theta} \ell(\theta). \tag{17}$$

Općenito, procjene metodom χ^2 i maksimalne vjerodostojnosti ne moraju biti jednake. Podacima prilagođeni model iz familije P_0 za populacijsku distribuciju od X tada je $(p_1(\tilde{\theta}), p_2(\tilde{\theta}), \ldots, p_r(\tilde{\theta}))$. Da bismo testirali reprezentativnost tog modela, postavljamo hipoteze H_0 i H_1 ranije navedene u tekstu. Testna statistika je tada $\chi^2(\tilde{\theta})$. Budući da su, uz istinitost H_0, oba procjenitelja, minimuma χ^2-a $\hat{\theta}$ i maksimalne vjerodostojnosti $\tilde{\theta}$, *asimptotski ekvivalentni*, asimptotski su ekvivalentne i testne statistike $\chi^2(\hat{\theta})$ i $\chi^2(\tilde{\theta})$. Drugim riječima, njihove razdiobe teže k istoj vjerojatnosnoj razdiobi: χ^2-razdiobi s $r - 1 - \nu$ stupnjeva slobode. To je sadržaj Fisherovog teorema. Radi potpunosti, iskazat ćemo ga:

Teorem. (*R. A. Fisher*[11]) *Pretpostavimo da za svaki prirodan broj* n *imamo definiran slučajni vektor* $\mathbf{N}^{(n)} = (N_1^{(n)}, N_2^{(n)}, \ldots, N_r^{(n)})$ *s polinomijalnom distribucijom s parametrima* $(n; p_1(\theta), p_2(\theta), \ldots, p_r(\theta))$, *pri čemu vektor vjerojatnosnih distribucija*

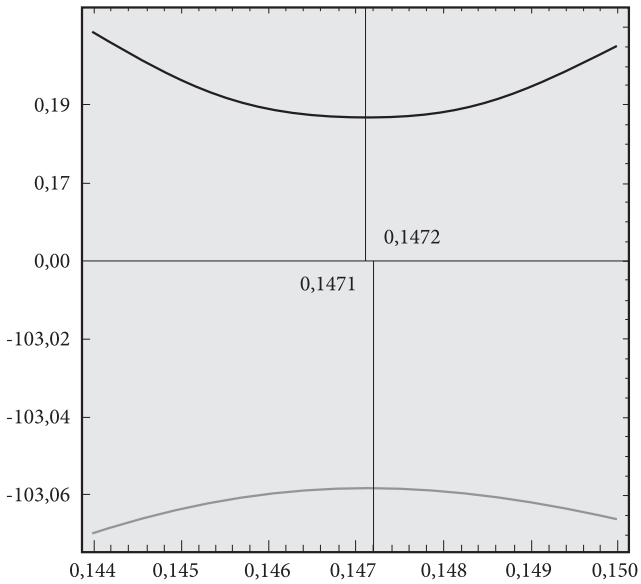
[11] Navedeni rezultat prvi je formulirao Ronald Aylmer Fisher 1924. g. u članku [2]. Suvremeni dokaz tog teorema može se naći u [8], str. 463.-467., te u diplomskom radu [4]. Jedan metodički zanimljiv pristup dokazivanju tog teorema prikazan je u članku [9]. Inače, taj teorem specijalni je slučaj općenitog rezultata koji se može naći u [1] (Theorems 23, 24, str. 152.-164.).

$(p_1(\theta), p_2(\theta), \ldots, p_r(\theta))$ pripada familiji P_0 definiranoj s (14) *i modelom* (11) *za koji još pretpostavljamo da zadovoljava neke uvjete regularnosti*[12]. *Nadalje, označimo s $\chi^2_n(\theta)$ χ^2-funkciju* (12), *te s $\hat{\theta}_n$ procjenitelj minimuma χ^2 ili procjenitelj maksimalne vjerodostojnosti za parametar θ na osnovi opaženih frekvencija, za svaki n. Neka $\rho = r - 1$ označava dimenziju prostora P svih vjerojatnosnih distribucija definiranu s* (10), *te neka je ν dimenzija parametarskog prostora Θ. Tada za sve pozitivne realne brojeve h vrijedi*:

$$\lim_{n\to\infty} \mathbb{P}(\chi^2_n(\hat{\theta}_n) > h) = \mathbb{P}(H > h),$$

gdje je H χ^2-distribuirana slučajna varijabla s $\rho - \nu$ stupnjeva slobode.

Primijetite da je, u slučaju da je P_0 jednočlana familija (tj. hipoteza H_0 je jednostavna), $\nu = 0$, pa je Pearsonov rezultat specijalan slučaj Fisherovog teorema. Primjena rezultata tog teorema na testiranje nul-hipoteze H_0 u odnosu na alternativu H_1, ilustriran je u prethodnom Primjeru 1., te u sljedećem Primjeru 3. U Primjeru 3. bit će pokazano i da se općenito procjene metodoma minimuma χ^2 i maksimalne vjerodostojnosti ne moraju podudarati.



Slika 3: Grafovi χ^2-funkcije u okolini svoje točke minimuma (crno) i funkcije log-vjerodostojnosti u okolini svoga maksimuma (sivo), iz Primjera 3.

[12] Pretpostavlja se da vrijede tzv. *Cramérovi uvjeti*: uz već spomenute uvjete da vektor $(p_1(\theta), p_2(\theta), \ldots, p_r(\theta))$ pripada familiji P, te da je preslikavanje koje svakom vektoru θ iz Θ pridružuje spomenuti vektor, injekcija, još se pretpostavlja da je to preslikavanje dva puta neprekidno diferencijabilno s Jacobijevom matricom punog ranga, te da su komponente $p_i(\theta)$, $i = 1, \ldots, r$ tog vektora uniformno ograničene odozdo pozitivnom konstantom (vidjeti [3], str. 76.). U svim našim primjerima ti su uvjeti zadovoljeni ili se može postići da budu zadovoljeni.

Primjer 3.[13] (Procjenitelj maksimalne vjerodostojnosti i minimuma χ^2 nisu jednaki)

Uzeli smo slučajni uzorak od $n = 100$ biljaka iz jednog botaničkog vrta i bilježili koliko od njih ima crvene, žute ili plave cvjetove. Crvene cvjetove imalo je 20 biljaka, žute 50, a plave 30. Želimo testirati da boja cvjetova biljaka zapravo dolazi iz diskretne distribucije oblika

$$\begin{pmatrix} \text{crvena} & \text{žuta} & \text{plava} \\ \frac{1}{3}-\theta & \frac{2}{3}-\theta & 2\theta \end{pmatrix}.$$

Dakle, moramo procijeniti parametar θ za koji vrijedi da je $0 < \theta < 1/3$. Pogledajmo χ^2-funkciju:

$$\chi^2(\theta) = \frac{(20-100(\frac{1}{3}-\theta))^2}{100(\frac{1}{3}-\theta)} + \frac{(50-100(\frac{2}{3}-\theta))^2}{100(\frac{2}{3}-\theta)} + \frac{(30-100\cdot 2\theta))^2}{100\cdot 2\theta} =$$

$$= \frac{4}{\frac{1}{3}-\theta} + \frac{25}{\frac{2}{3}-\theta} + \frac{9}{2\theta} - 100.$$

Točku njenog minimuma možemo tražiti grafički (vidjeti Sliku 3.). Dakle, procjena minimuma χ^2 za θ je:

$$\hat{\theta} = 0.1471.$$

Na isti način možemo odrediti točku maksimuma funkcije log-vjerodostojnosti koja je u ovom primjeru (do na konstantu) jednaka:

$$\ell(\theta) = 20\log(\frac{1}{3}-\theta) + 50\log(\frac{2}{3}-\theta) + 30\log(2\theta).$$

Iz grafa te funkcije (također Slika 3.) slijedi da je točka njenog maksimuma jednaka:

$$\tilde{\theta} = 0.1472.$$

Nakon što smo procijenili parametar θ (uočimo da je $\nu = 1$), trebamo testirati nul-hipotezu H_0 za naš konkretan model u ovome primjeru. Za vrijednost χ^2-testne statistike uzet ćemo vrijednost χ^2-funkcije u procjeni maksimalne vjerodostojnosti:

$$\chi^2(\tilde{\theta}) = \chi^2(0.1472) = 0.1869.$$

[13] Podaci su iz [1], Exercise 5, str. 161., a model iz iste knjige, Example 2, str. 154.

Prema Fisherovom teoremu, uz pretpostavku istinitosti nul-hipoteze H_0, razdioba testne statistike je za veliki n približno jednaka χ^2 -razdiobi s $r - 1 - v = 3 - 1 - 1 = 1$ stupnjem slobode. Zbog toga je za zadanu razinu značajnosti $\alpha = 5\% = 0.05$ kritična vrijednost jednaka $\chi^2_{0.05}(1) = 3.8415$.

Budući da je 0.1869 < 3.8, na razini značajnosti od 5% ne odbacujemo nul-hipotezu.

3. Test nezavisnosti

U sljedeća tri poglavlja primijenit ćemo u prethodnom poglavlju opisanu χ^2 -metodu na tri važna problema u statistici. U ovome poglavlju opisat ćemo kako testirati hipotezu o nezavisnosti dvaju diskretnih statističkih obilježja.

Pretpostavimo da vršimo niz od n nezavisnih i po uvjetima sprovođenja jednakih slučajnih pokusa i da pri svakom pokusu mjerimo (ili opažamo) vrijednosti obilježja X i Y. Nadalje, pretpostavimo da obilježje X poprima vrijednosti iz skupa $\{a_1,...,a_r\}$, a obilježje Y iz skupa $\{b_1,...,b_c\}$. Neka je n_{ij} broj pokusa čiji ishod ima vrijednost a_i od X i b_j od Y, i = 1, 2, ..., r, j = 1, ..., c. Ako ishod nekog pokusa, koji ima vrijednost a_i obilježja X i vrijednost b_j obilježja Y, poistovjetimo s uređenim parom (a_i, b_j), onda je n_{ij} broj pokusa koji su dali ishod (a_i, b_j) dvodimenzionalnog obilježja (X, Y). Podatke onda možemo zapisati pomoću tablice:

$X \setminus Y$	b_1	b_2	$\cdots$	b_c	ukupno
a_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_2
a_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_r
ukupno	n_1	n_2	$\cdots$	n_c	n

Primijetite da su *marginalne* frekvencije vrijednosti a_i od X i b_j od Y jednake redom:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \tag{18}$$

za sve i = 1, 2, ..., r i sve j = 1, ..., c, te da je duljina uzorka jednaka:

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j}.$$

Nama je cilj testirati nul-hipotezu i dokazati da su obilježja X i Y nezavisna. X i Y su nezavisna obilježja ako je vjerojatnost da se u jednom pokusu dogodi ishod $(X,Y)=(a_i,b_j)$ jednaka produktu vjerojatnosti ishoda $X = a_i$ (bez obzira koja je vrijednost od Y) i vjerojatnosti ishoda $Y = b_j$ (bez obzira koja je vrijednost od X), za sve parove vrijednosti (a_i, b_j). Preciznije, ako uvedemo oznake:

$$\pi_{ij} := \mathbb{P}(X = a_i, Y = b_j),\ \ p_i := \mathbb{P}(X = a_i),\ \ q_j := \mathbb{P}(Y = b_j),$$

za $i = 1, 2, ..., r$ i $j = 1, ..., c$, tada želimo testirati nul-hipotezu:

$$H_0 : \pi_{ij} = p_i \cdot q_j,\ \ \text{za sve}\ \ i = 1,2,\ldots,r,\ \ \text{i sve}\ \ j = 1,\ldots,c.$$

Nju želimo testirati nasuprot alternativnoj hipotezi:

$$H_1 : \pi_{ij} \neq p_i \cdot q_j,\ \ \text{za barem jedan par}\ \ (i,j), i = 1,2,\ldots,r,\ j = 1,\ldots,c.$$

Budući da se ovdje radi o diskretnom dvodimenzionalnom obilježju (X, Y) s $r \cdot c$ različitih vrijednosti, skup svih vjerojatnosnih distribucija tog obilježja je:

$$P = \{(\pi_{ij}; i = 1,\ldots r,\ j = 1,\ldots,c) \in \mathbb{R}^{rc} : (\forall i,j)(\pi_{ij} > 0)\ \&\ \sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij} = 1\}.$$

Prava dimenzija od P je $\rho = rc - 1$. S druge strane, nul-hipotezom H_0 zadan je parametarski model koji želimo testirati, s vektorom parametara:

$$\theta = (p_1, p_2, \ldots, p_{r-1}, q_1, q_2, \ldots, q_{c-1}).$$

Za parametre vrijedi da su strogo pozitivni, te da su zadovoljene relacije:

$$p_r = 1 - \sum_{i=1}^{r-1} p_i > 0\ \ \text{i}\ \ q_c = 1 - \sum_{j=1}^{c-1} q_j > 0. \tag{19}$$

Parametarski prostor Θ je podskup od $\mathbb{R}^{(r-1)+(c-1)}$ definiran upravo navedenim uvjetima. Dakle, ukupan broj parametara je $\nu = (r-1)+(c-1)$. Primijetimo još da iz zadanog parametarskog modela u H_0 slijedi da je:

$$(\forall i \in \{1,\ldots,r\})(p_i = \sum_{j=1}^{c}\pi_{ij})\ \ \text{i}\ \ (\forall j \in \{1,\ldots,c\})(q_j = \sum_{i=1}^{r}\pi_{ij}), \tag{20}$$

odnosno da su $(p_1,\ldots,p_r)$ i $(q_1,\ldots,q_c)$ *marginalne* distribucije od X i Y redom, što odgovara polaznom zahtjevu na te parametre.

U skladu s H_0, χ^2-funkcija definirana na Θ je:

$$\chi^2(\theta) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - np_iq_j)^2}{np_iq_j}.$$

Naći ćemo procjenu za θ maksimalne vjerodostojnosti, a čitateljima ostavljamo da nađu procjenu minimuma χ^2, te da pokažu da će za ovaj model te dvije procjene biti jednake.

Uz pretpostavku da opažene frekvencije imaju polinomijalnu distribuciju, do na konstantu koja ne ovisi o θ, log-vjerodostojnost od θ jednaka je:

$$\ell(\theta) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln(p_i q_j) = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\ln p_i + \ln q_j) = \sum_{i=1}^{r} n_{i\cdot} \ln p_i + \sum_{j=1}^{c} n_{\cdot j} \ln q_j,$$

pri čemu treća jednakost slijedi iz relacija (18). Primjenom Shannon-Kolmogorovljeve informacijske nejednakosti, prvo na vjerojatnosne vektore $(p_1,\ldots,p_r)$ i $(n_{1\cdot}/n,\ldots,n_{r\cdot}/n)$, a zatim na vektore $(q_1,\ldots,q_c)$ i $(n_{\cdot 1}/n,\ldots,n_{\cdot c}/n)$, kao u Primjeru 2. dobivamo nejednakost:

$$\ell(\theta) = \sum_{i=1}^{r} n_{i\cdot} \ln p_i + \sum_{j=1}^{c} n_{\cdot j} \ln q_j \le \sum_{i=1}^{r} n_{i\cdot} \ln\left(\frac{n_{i\cdot}}{n}\right) + \sum_{j=1}^{c} n_{\cdot j} \ln\left(\frac{n_{\cdot j}}{n}\right) = \ell(\hat{\theta}),$$

pri čemu jednakost vrijedi ako i samo ako među članovima obaju navedenih parova vektora vrijedi jednakost. Odavde slijedi da je jedinstvena procjena maksimalne vjerodostojnosti za θ jednaka:

$$\hat{\theta} = \left(\frac{n_{1\cdot}}{n},\ldots,\frac{n_{(r-1)\cdot}}{n},\frac{n_{\cdot 1}}{n},\ldots,\frac{n_{\cdot(c-1)}}{n}\right).$$

Tada je testna statistika za testiranje H_0 u odnosu na H_1 jednaka:

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \left(\frac{n_{i\cdot}n_{\cdot j}}{n}\right)\right)^2}{\left(\frac{n_{i\cdot}n_{\cdot j}}{n}\right)}.$$

Pod pretpostavkom da vrijedi nul-hipoteza H_0, a prema Fisherovom teoremu, ta statistika ima asimptotski χ^2-distribuciju s $\rho - \nu = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$ stupnjeva slobode, kada n teži u neizmjerno.

Pogledajmo na jednom primjeru kako se testira nezavisnost.

Primjer 4.[14] Pretpostavimo da n = 300 studenata piše inicijalni ispit iz matematike. 97 od njih je ispit položilo, a ostalih 203 palo je ispit. Nakon tri mjeseca isti su studenti pristupili završnom ispitu iz matematike, no ovaj put ga je položilo njih 48. Od tih 48 studenata, 18 je položilo i inicijalni ispit. Želimo testirati hipotezu nezavisnosti polaganja ovih dvaju ispita.

[14] Vidjeti [3], str. 90.

Označimo s X indikator prolaza na završnom, a s Y indikator prolaza na inicijalnom ispitu. Ovdje su $a_1 = b_1 = 1$ ("prolaz") i $a_2 = b_2 = 0$ ("pad"), te $r = 2$ i $c = 2$. Tablica s opaženim frekvencijama:

$X \setminus Y$	1	0	ukupno
1	18	30	48
0	79	173	252
ukupno	97	203	300

Izračunajmo vrijednost Pearsonove χ^2-statistike uz pretpostavku nezavisnosti X i Y:

$$\chi^2 = \frac{(18 - \frac{48 \cdot 97}{300})^2}{\frac{48 \cdot 97}{300}} + \frac{(30 - \frac{48 \cdot 203}{300})^2}{\frac{48 \cdot 203}{300}} + \frac{(79 - \frac{252 \cdot 97}{300})^2}{\frac{252 \cdot 97}{300}} + \frac{(173 - \frac{252 \cdot 203}{300})^2}{\frac{252 \cdot 203}{300}} = 0.697.$$

Za razinu značajnosti od $\alpha = 1\% = 0.01$ i $(r-1)(c-1) = 1 \cdot 1 = 1$ stupnja slobode, kritična vrijednost jednaka je $\chi^2_{0.01}(1) = 2.707$. Kako je $0.697 < 2.707$, ne odbacujemo nul-hipotezu o nezavisnosti obilježja X i Y na razini značajnosti od 1%.

4. Test homogenosti

Pretpostavimo da imamo $m \geq 2$ populacija, te da promatramo isto diskretno statističko obilježje X u svakoj populaciji. Na primjer, zanima nas razdioba obilježja „boja kose stanovništva" u dvije populacije: u stanovništvu Sjeverne Amerike i u stanovništvu Azija.

Općenito, neka je $A = \{a_1, ..., a_r\}$ skup svih različitih vrijednosti obilježja X, te neka je s π_{ij} označena vjerojatnost vrijednosti a_j tog obilježja u populaciji i, za $j = 1, 2, ..., r$ i $i = 1, 2, ..., m$. Dakle, u populaciji i ($i = 1, 2, ..., m$), zakon razdiobe od X jednak je:

$$\begin{pmatrix} a_1 & a_2 & \cdots & a_r \\ \pi_{i1} & \pi_{i2} & \cdots & \pi_{ir} \end{pmatrix}.$$

Želimo testirati nul-hipotezu da su razdiobe od X u svih promatranih m populacija jednake, odnosno homogene. Dakle, testiramo nul-hipotezu:

$$H_0 : \pi_{ij} = p_j, \text{ za sve } i = 1, ..., m \text{ i sve } j = 1, ..., r$$

nasuprot alternativnoj hipotezi:

$$H_1 : \pi_{ij} \neq p_j, \text{za barem jedan par } (i, j), i = 1, ..., m, j = 1, ..., r.$$

Familija svih razdioba obilježja X u svih m populacija jednaka je:

$$P = \{(\pi_{ij}; j = 1,\ldots,r, i = 1,\ldots,m) : (\forall i, j)(\pi_{ij} > 0)\ \&\ (\forall i)(\sum_{j=1}^{r} \pi_{ij} = 1\}.$$

Prava dimenzija od P jednaka je $\rho = m(r-1)$. Nadalje, nul-hipotezom H_0 zadan je parametarski model koji želimo testirati. Vektor parametara tog modela jednak je

$$\theta = (p_1, p_2, \ldots, p_{r-1}),$$

s parametarskim prostorom Θ definiranog uvjetima

$$p_i > 0, i = 1,2,\ldots,r-1,\ p_r = 1 - \sum_{i=1}^{r-1} p_i > 0.$$

Odavde slijedi da je ukupan broj parametara danih modelom u H_0 jednak $v = r - 1$.

Pretpostavimo sada da, neovisno od drugih populacija, iz i-te populacije uzimamo slučajni uzorak zadane duljine n_i za svaki i = 1, 2, ..., m. Označimo s n_{ij} broj ponavljanja vrijednosti a_j od X u opaženom uzorku iz populacije i, j = 1, 2, ..., r, i = 1, 2, ..., m. Neka je $n = n_{1\cdot} + \ldots + n_{m\cdot}$ ukupna duljina svih m uzoraka spojenih zajedno u jedan uzorak. Nadalje, neka je $n_{\cdot j} = \sum_{i=1}^{m} n_{ij}$ ukupna frekvencija vrijednosti a_j od X u tom zajedničkom uzorku, j = 1, 2, ..., r. Opažene frekvencije možemo prikazati u tablici:

Populacija $\setminus X$	a_1	a_2	$\cdots$	a_r	ukupno
1	n_{11}	n_{12}	$\cdots$	n_{1r}	n_1
2	n_{21}	n_{22}	$\cdots$	n_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mr}	n_m
ukupno	n_1	n_2	$\cdots$	n_c	n

Primijetite da je za svaki i = 1, 2, ..., m, $n_{i\cdot} = \sum_{j=1}^{r} n_{ij}$ jer je zbroj frekvencija svih vrijednosti u uzorku jednaka veličini uzorka. Iz istog je razloga i $\sum_{j=1}^{c} n_{\cdot j} = n$.

Pripadna χ^2-funkcija jednaka je:

$$\chi^2(\theta) = \sum_{i=1}^{m}\sum_{j=1}^{r} \frac{(n_{ij} - n_{i\cdot} p_j)^2}{n_{i\cdot} p_j}.$$

Nepoznatu vrijednost vektora parametara procijenit ćemo metodom maksimalne vjerodostojnosti. Označimo s N_{ij} slučajnu frekvenciju vrijednosti a_j od X,

za $j = 1, 2, ..., r$, u slučajnom uzorku uzetom iz populacije i (n_{ij} je opažena vrijednost od N_{ij}), za $i = 1, 2, ..., m$. Tada za svaki $i = 1, 2, ..., m$, uz pretpostavku istinitosti hipoteze H_0, slučajni vektor $\mathbf{N}^{(i)} = (N_{i1}, N_{i2}, \ldots, N_{ir})$ ima polinomijalnu distribuciju s parametrima $(n_{i\cdot}; p_1, p_2, \ldots, p_r)$. Budući da su uzorci uzimani tako da su vektori $\mathbf{N}^{(i)}$, $i = 1, 2, ..., m$, i nezavisni, vjerodostojnost od θ jednaka je:

$$L(\theta) = \prod_{i=1}^{m} \frac{n_{i\cdot}!}{\prod_{j=1}^{r} n_{ij}!} \prod_{j=1}^{r} p_j^{n_{ij}}.$$

Odavde slijedi da je log-vjerodostojnost od θ (do na konstantu koja ne ovisi o parametrima) jednaka:

$$\ell(\theta) = \sum_{i=1}^{m} \sum_{j=1}^{r} n_{ij} \ln p_j = \sum_{j=1}^{r} n_{\cdot j} \ln p_j.$$

Sada na potpuno isti način kao u Primjeru 2. zaključujemo da je procjena maksimalne vjerodostojnosti za θ jednaka:

$$\hat{\theta} = \left(\frac{n_{\cdot 1}}{n}, \frac{n_{\cdot 2}}{n}, \ldots, \frac{n_{\cdot (r-1)}}{n} \right).$$

Može se pokazati da se ta procjena podudara s procjenom minimuma χ^2. Dakle, χ^2-testna statistika iznosi:

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \left(\frac{n_{i\cdot} n_{\cdot j}}{n} \right) \right)^2}{\left(\frac{n_{i\cdot} n_{\cdot j}}{n} \right)}.$$

Pod pretpostavkom da vrijedi nul-hipoteza H_0, a prema Fisherovom teoremu,[15] ta statistika ima asimptotski χ^2-distribuciju s $\rho - \nu = m(r-1) - (r-1) = (m-1)(r-1)$ stupnjeva slobode, kada n teži u neizmjerno.

Primjer 5.[16] Pretpostavimo da dvije grupe od po 300 studenata pristupaju istovremeno ispitu i rješavaju iste ispitne zadatke. U prvoj je grupi 144 studenta dobilo ocjenu *A*, 80 ocjenu *B*, 43 ocjenu *C*, a 33 je palo. U drugoj je grupi distribucija ocjena bila: 154 s ocjenom *A*, 72 s ocjenom *B*, 35 s ocjenom *C*, a 39 ih je palo. Želimo testirati hipotezu da je distribucija ocjena na ispitu neovisna o grupi. Tablica opaženih frekvencija za ovaj primjer je:

[15] Ovdje se radi o nešto općenitijoj verziji tog teorema. Naime, slučajni vektori opažanih frekvencija su nezavisni (i, dakako, polinomijalno distribuirani), odakle slijedi da zbroj χ^2-statistika koje se na njima temelje (a što je, u stvari, naša testna statistika), također ima asimptotski χ^2-razdiobu. Za precizno objašnjenje pogledajte u [1], Theorem 24.
[16] Vidjeti [3], str. 84.-86.

grupa \ ocjene	*A*	*B*	*C*	pad	ukupno
1	144	80	43	33	300
2	154	72	35	39	300
ukupno	298	152	78	72	600

Vrijednost testne χ^2 -statistike je

$$\chi^2 = \frac{\left(144 - 300\cdot\frac{298}{600}\right)^2}{149} + \frac{\left(80 - 300\cdot\frac{152}{600}\right)^2}{76} + \frac{\left(43 - 300\cdot\frac{78}{600}\right)^2}{39} + \frac{\left(33 - 300\cdot\frac{72}{600}\right)^2}{36} +$$

$$+ \frac{\left(154 - 300\cdot\frac{298}{600}\right)^2}{149} + \frac{\left(72 - 300\cdot\frac{152}{600}\right)^2}{76} + \frac{\left(35 - 300\cdot\frac{78}{600}\right)^2}{39} + \frac{\left(39 - 300\cdot\frac{72}{600}\right)^2}{36} =$$

$$= 2.077.$$

Za razinu značajnosti od $\alpha = 5\% = 0.05$ i $(m-1)(r-1) = (2-1)(4-1) = 3$ stupnja slobode, kritična vrijednost iznosi $\chi^2_{0.95}(3) = 7.815$. Kako je 2.077 < 7.815, ne odbacujemo nul-hipotezu o homogenosti distribucije ocjena po grupama, na razini značajnosti od 5%.

5. Neprekidne distribucije

Na kraju ćemo se osvrnuti na slučaj kada imamo slučajni uzorak duljine *n*, ali svaki pokus može dati kao ishod bilo koji realan broj iz nekog intervala realnih brojeva. Tada su podaci realizacije slučajne varijable s *neprekidnom* razdiobom. Kažemo da slučajna varijabla *X* ima neprekidnu razdiobu ukoliko postoji nenegativna funkcija $f : \mathbb{R} \to \mathbb{R}$ takva da za svaka dva realna broja *a* < *b* vrijedi da je vjerojatnost događaja da *X* poprimi vrijednosti u intervalu [*a*, *b*] jednaka površini omeđenoj grafom od *f*, osi *x* i pravcima *x* = *a* i *x* = *b*.[17] Takvu funkciju *f* zovemo *gustoćom* razdiobe od *X*. Ukoliko s $F(x) := \mathbb{P}(X \le x)$, $x \in \mathbb{R}$, označimo *funkciju distribucije* od X, tada za sve *a* < *b* vrijedi:

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a < X \le b) = \mathbb{P}(a \le X < b) = \mathbb{P}(a < X < b) = F(b) - F(a).$$

[17] Preciznije, X ima neprekidnu razdiobu ako postoji nenegativna realna funkcija *f* takva da je za sve realne brojeve $a < b$, $\mathbb{P}(a \le X \le b) = \int_a^b f(x)dx$, što je ekvivalentno tome da je pripadna funkcija distribucije od *X* jednaka $F(x) = \int_{-\infty}^{x} f(x)dx$, $X \in \mathbb{R}$. Uvođenjem funkcije distribucije *F* može se izbjeći korištenje integrala za računanje vjerojatnosti događaja opisanih (generiranih) s *X* u kontekstu ove teme (bilo da je *F* zadana formulom ili tablicom, ili da se računa pomoću kalkulatora ili računala).

Vjerojatnosni modeli za neprekidne razdiobe često su parametrizirani. Da bismo to naglasili (uz argument bilo gustoće, bilo funkcije distribucije od X), navodimo i parametar: $f(x;\theta)$, $F(x;\theta)$, gdje je θ parametar ili, ako ih ima više, vektor parametara, s vrijednostima u skupu (parametarskom prostoru) Θ. Na primjer, u ovome poglavlju (Primjeri 6. i 7.) posebno ćemo gledati normalnu razdiobu s parametrima μ i σ^2, u oznaci $N(\mu,\sigma^2)$. Ovdje su: vektor parametara $\theta=(\mu,\sigma^2)$, parametarski prostor $\Theta=\mathbb{R}\times\langle 0,+\infty\rangle$, gustoća

$$f(x;\theta)=f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in\mathbb{R},$$

i funkcija distribucije:

$$F(x;\theta)=F(x;\mu,\sigma^2)=\Phi\left(\frac{x-\mu}{\sigma}\right),\ x\in\mathbb{R}.$$

Φ označava funkciju distribucije *jedinične* normalne razdiobe, tj. normalne razdiobe $N(0, 1)$ koja je tabelirana.

U Primjeru 6. pokazat ćemo kako se pomoću χ^2-testa testiraju jednostavne, a u Primjeru 7. složene hipoteze o neprekidnim razdiobama. Za oba testiranja zajedničko je da prvo treba odrediti intervale vrijednosti od X na osnovi čijih ćemo frekvencija, izvedenih iz opaženog uzorka, sprovesti test. Dakle, moramo podijeliti skup realnih brojeva $\mathbb{R}$ (ili barem sliku od X) na intervale koji su međusobno disjunktni i koji prekrivaju cijeli skup $\mathbb{R}$. Te intervale zovemo još i *razredi*. Pretpostavimo da ih ima r. Tada su razredi:

$$A_1=\langle-\infty,a_1],A_2=\langle a_1,a_2],\ldots,A_{r-1}=\langle a_{r-2},a_{r-1}],A_r=\langle a_{r-1},+\infty\rangle,$$

pri čemu su $a_1<a_2<\ldots<a_{r-1}$ granice razreda. Za svaki i = 1, 2, ..., r označimo s π_i vjerojatnost da varijabla X poprimi vrijednost u intervalu A_i. Dakle,

$$\pi_i=\mathbb{P}(X\in A_i),\ i=1,2,\ldots,r.$$

U načelu broj razreda i razrede zadajemo po volji. Ipak, i tu postoje neka ograničenja. Na primjer, broj razreda r ne smije biti prevelik u odnosu na broj n podataka u uzorku jer će u suprotnom biti previše intervala s premalom frekvencijom. Nadalje, intervali moraju biti odabrani tako da očekivana frekvencija svakog od njih ne bude premala (obično se uzima da očekivane frekvencije bude barem 5 za „unutarnje" ili „središnje" razrede, a barem 1 za „vanjske" ili „rubne" razrede)[18].

Ukoliko imamo jednostavnu nul-hipotezu, tj. da X ima neprekidnu razdiobu sa zadanom gustoćom f (često je $f\equiv f(\cdot;\theta_0)$, gdje je θ_0 zadana vrijednost parametra θ), tada se preporuča da se razredi biraju tako da svi imaju jednake očekivane

[18] Vidjeti [3] str. 19. i 20.

frekvencije. U tom slučaju će nam χ^2-testna statistika imati najmanju varijancu[19]. Dakle, granice razreda a_i, $i = 1, 2, ..., r - 1$, rješenja su jednadžbi (F je pripadna funkcija distribucije):

$$\begin{cases} F(a_1) & = & \frac{1}{r}, \\ F(a_i) - F(a_{i-1}) & = & \frac{1}{r}, \qquad i = 2, \ldots, r-1, \\ 1 - F(a_{r-1}) & = & \frac{1}{r}, \end{cases}$$

što je ekvivalentno s $F(a_i) = i/r$, $i = 1, 2, ..., r - 1$, odnosno, a_i je (i/r)-ti *kvantil* funkcije distribucije F za sve $i = 1, 2, ..., r - 1$. χ^2-statistika tada je oblika

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(n_i - \frac{n}{r}\right)^2}{\frac{n}{r}} = \frac{r}{n} \sum_{i=1}^{r} n_i^2 - n. \quad (21)$$

U slučaju složene nul-hipoteze, tj. da X ima neprekidnu razdiobu s gustoćom $f(\cdot;\theta)$ koja pripada nekoj parametarskoj familiji, za neku vrijednost $\theta \in \Theta$, razrede najčešće određujemo unaprijed, tako da su unutarnji razredi jednake duljine. Kada ih odredimo, računamo funkcije

$$p_1(\theta) := F(a_1;\theta), \;\; p_i(\theta) := F(a_i;\theta) - F(a_{i-1};\theta), i = 2, \ldots, r-1, \;\; p_r(\theta) := 1 - F(a_r;\theta).$$

Nul-hipoteza tada se može zapisati i na poznati način:

$$H_0 : \pi_i = p_i(\theta), \;\; \text{za svaki } \; i = 1, 2, \ldots, r \;\; \text{i neki } \; \theta \in \Theta.$$

Primijetimo da za svaki θ, vektor $(p_1(\theta), \ldots, p_r(\theta))$ pripada familiji P definiranoj s (10). χ^2-test se dalje sprovodi na uobičajeni način (opisan u drugom poglavlju).

Primjer 6. Generirajmo slučajni uzorak duljine $n = 84$ iz normalne razdiobe $N(144, 36)$ pomoću generatora slučajnih brojeva u nekom matematičkom programskom paketu (npr. MATLAB-u). Želimo testirati da su podaci dobro izgenerirani, tj. da je varijabla X, za koju su ti podaci nezavisne realizacije njenih vrijednosti, normalna s razdiobom $N(144, 36)$. Dakle, testiramo

$$H_0 : \;\; X \;\; \text{ima} \;\; N(144, 36)\text{-razdiobu},$$

a alternativna hipoteza je

$$H_1 : \;\; X \;\; \text{nema} \;\; N(144, 36)\text{-razdiobu}.$$

[19] Varijanca Pearsonove χ^2-statistike ovisi o vjerojatnostima π_i, $i = 1, \ldots, r$ i najmanju vrijednost $2(r-1)\left(1 - \frac{1}{n}\right)$ postiže u slučaju da je $\pi_1 = \cdots = \pi_r = 1/r$ (vidjeti [3], str. 5.).

Tablica opaženih i teorijskih frekvencija je:

razredi	frekvencije	očekivane frekvencije	
A_i	n_i	n/r	n_i^2
$\langle -\infty, 136.3]$	8	8.4	64
$\langle 136.3, 139.0]$	9	8.4	81
$\langle 139.0, 140.9]$	7	8.4	49
$\langle 140.9, 142.5]$	2	8.4	4
$\langle 142.5, 144.0]$	12	8.4	144
$\langle 144.0, 145.5]$	4	8.4	16
$\langle 145.5, 147.1]$	10	8.4	100
$\langle 147.1, 149.0]$	9	8.4	81
$\langle 149.0, 151.7]$	8	8.4	64
$\langle 151.7, +\infty\rangle$	15	8.4	225
ukupno	84	84	828

Broj razreda $r = 10$ uzeli smo proizvoljno, ali smo granice razreda birali tako da sve teorijske frekvencije budu jednake i iznose $n / r = 8.4$. Do granica za razrede došli smo rješavajući sustav jednadžbi:

$$\Phi\left(\frac{a_i - 144}{6}\right) = \frac{i}{10}, \quad i = 1, 2, \ldots, 9.$$

Izračunajmo vrijednost Pearsonove χ^2-testne statistike (21):

$$\chi^2 = \frac{r}{n}\sum_{i=1}^{10} n_i^2 - n = \frac{828}{8.4} - 84 = 14.5714.$$

Za razinu značajnosti od $\alpha = 5\%$ i 10 – 1 = 9 stupnjeva slobode, kritična vrijednost jednaka je $\chi^2_{0.05}(9) = 16.9190$. Kako je 14.5714 < 16.9190, ne odbacujemo nulhipotezu da su podaci normalno distribuirani s parametrima $\mu = 144, \sigma^2 = 36$.

Primjer 7.[20] U svrhu istraživanja porijekla etrurske kulture i Etruščana, izvršeno je mjerenje maksimalne širine lubanja 84 Etruščana. Podaci su izraženi u milimetrima. Aritmetička sredina tih $n = 84$ brojeva je $\overline{x} = 143.8$, a uzoračka standardna devijacija $s = 6.0$. Neka je X maksimalna širina lubanje slučajno odabranog Etruščana iz populacije. Želi se testirati

$$H_0: \quad X \text{ je } N(\mu,\sigma^2)\text{-distribuirana za neke } \mu \text{ i } \sigma^2,$$

nasuprot alternativnoj hipotezi

$$H_1: \quad X \text{ nije normalno distribuirana varijabla.}$$

Uočimo odmah da je broj nepoznatih parametara $\nu = 2$. Parametar $\theta = (\mu,\sigma^2)$ treba procijeniti metodom minimuma χ^2 ili maksimalne vjerodostojnosti u skupu $\Theta = \mathbb{R} \times \langle 0,+\infty\rangle$, kako je to opisano u drugom poglavlju[21]. Odabiremo metodu maksimalne vjerodostojnosti. Koristit ćemo logaritam vjerodostojnosti $\ell(\theta) = \sum_{i=1}^{r} n_i \ln p_i(\theta)$ za računanje procjene maksimalne vjerodostojnosti $\tilde{\theta}$. Ako pogledamo funkciju ℓ, onda vidimo da vjerojatnosti $p_i(\theta)$ ne ovise samo o parametru θ, nego i o razredima A_i ($i = 1,\ldots,r$), pa je računanje maksimuma takve funkcije kompleksan problem koji se može riješiti numeričkim metodama uz pomoć računala. Budući da je vjerojatnosni model koji testiramo normalan, do zadovoljavajuće procjene za $\theta = (\mu,\sigma^2)$ možemo doći i jednostavnije, na sljedeći način.

Neka je $r = 9$. Granice razreda odredimo tako da su unutarnji razredi jednake duljine i simetrično postavljeni u odnosu na aritmetičku sredinu podataka $\overline{x} = 143.8$. Opažene frekvencije za tako odabrane razrede dane su u tablici:

Razredi A_i	Frekvencije n_i
$\langle -\infty, 129.1]$	1
$\langle 129.1, 133.3]$	3
$\langle 133.3, 137.5]$	8
$\langle 137.5, 141.7]$	16
$\langle 141.7, 145.9]$	23
$\langle 145.9, 150.1]$	25
$\langle 150.1, 154.3]$	5
$\langle 154.3, 158.5]$	3
$\langle 158.5, +\infty\rangle$	0
ukupno n	84

[20] Podaci i problem su preuzeti iz knjige: F. DALY i ostali, *Elements of Statistics*, Addison-Wesley, 1995., str. 96., *Example* 3.1.

[21] Ukoliko u ovome primjeru procijenimo nepoznate parametre normalne razdiobe metodom maksimalne vjerodostojnosti iz *negrupiranih* podataka u razrede, dakle s $\hat{\mu} = \overline{x}$ i $\hat{\sigma}^2 = ((n-1)/n)s^2$ ($\overline{x}$ je aritmetička sredina, a s^2 uzoračka varijanca), asimptotska razdioba χ^2 -funkcije izračunate na osnovi tako dobivenih procjenitelja bit će jednaka razdiobi od $\chi^2 + \lambda_1\xi_1^2 + \lambda_2\xi_2^2$, gdje su χ^2, ξ_1, ξ_2 nezavisne slučajne varijable, χ^2 s $\chi^2(r-1)$ -razdiobom, ξ_1 i ξ_2 s $N(0, 1)$ -razdiobom, te $\lambda_1 > 0$ i $\lambda_2 > 0$ su neke konstante. To je posljedica *Černov-Lehmannovog teorema* primijenjenog na normalnu razdiobu (vidjeti [3], str. 102.-106.).

Neka su a_{i-1} i a_i granice razreda A_i, za $i = 1, \ldots, r$. Stavimo da je $a_0 = 124.9$ i $a_9 = 162.7$ tako da se čini da su i vanjski razredi jednake duljine s unutarnjim. Neka su

$$x_i = \frac{a_{i-1} + a_i}{2}, \ i = 1, 2, \ldots, r,$$

sredine tih razreda. Sada imamo frekvencijsku tablicu:

x_i	127.0	131.2	135.4	139.6	143.8	148.0	152.2	156.4	160.6
n_i	1	3	8	16	23	25	5	3	0

Procjena parametra $\theta = (\mu, \sigma^2)$ računa se kao aritmetička sredina i varijanca podataka iz te tablice. Dakle, procjena $\tilde{\theta} = (\tilde{\mu}, \tilde{\sigma}^2)$ dobije se pomoću formula:

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{r} x_i n_i = 143.7, \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{r} n_i (x_i - \tilde{\mu})^2 = 34.23.$$

Nakon što smo procijenili parametre i pomoću njih izračunali očekivane frekvencije, imamo novu tablicu:

A_i	n_i		$e_i := np_i(\tilde{\theta})$		$(n_i - e_i)^2 / e_i$
$\langle -\infty, 129.1]$	1		0.5156		
$\langle 129.1, 133.3]$	3	4	2.5952	3.1108	0.2541
$\langle 133.3, 137.5]$	8		8.8754		0.0863
$\langle 137.5, 141.7]$	16		18.5073		0.3397
$\langle 141.7, 145.9]$	23		23.5499		0.0128
$\langle 145.9, 150.1]$	25		18.2911		2.4608
$\langle 150.1, 154.3]$	5		8.6690		1.5529
$\langle 154.3, 158.5]$	3		2.5052		
$\langle 158.5, +\infty \rangle$	0	3	0.4912	2.9964	0.0000
ukupno	84		84		4.7066

Budući da su očekivane frekvencije prvog i zadnjeg razreda premale, združit ćemo prvi A_1 i drugi A_2 razred, kao i predzadnji A_8 i zadnji A_9 razred, u nove razrede $A_{1'} = A_1 \cup A_2$ i $A_{7'} = A_8 \cup A_9$. Novodobiveni broj razreda je $r' = 7$.

Sada nam samo preostaje izračunati vrijednost χ^2 -testne statistike obzirom na nove razrede, $\chi^2(\tilde{\theta})'$, koja iznosi (vidjeti tablicu):

$$\chi^2(\tilde{\theta})' = \chi^2(143.7, 34.23)' = 4.7066.$$

Za razinu značajnosti od $\alpha = 10\%$ i $r' - v - 1 = 7 - 2 - 1 = 4$ stupnjeva slobode, kritična vrijednost jednaka je $\chi^2_{0.1}(4) = 7.7794$. Budući da je 4.7066 < 7.7794, ne odbacujemo nul-hipotezu da su maksimalne širine lubanja Etruščana normalno distribuirane na razini značajnosti od 10%.

Literatura

[1] T. S. Ferguson, *A Course in Large Sample Theory*, Chapman & Hall, London, 1996.

[2] R. A. Fisher, The conditions under which χ^2 measures the discrepancy between observation and hypothesis, *Journal of the Royal statistical Society*, **87** (1924.), 442. - 450.

[3] P. E. Greenwood, M. S. Nikulin, *A Guide to Chi-Squared Testing*, Wiley, New York, 1996.

[4] S. Lubura, *Hi-kvadrat test*, diplomski rad, PMF-Matematički odjel u Zagrebu, Zagreb, 2009.

[5] K. Pearson, On the criterion that the given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling, *Philosophical Magazine (5th Series)*, **50** (1900.), 157. - 175.

[6] R. L. Plackett, Karl Pearson and the Chi-squared test, *International Statistical Review*, Vol. 51, No. 1 (1983.), 59. - 72.

[7] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 1987.

[8] M. J. Schervish, *Theory of statistics*, Springer-Verlag, New York, 1995.

[9] M. E. Solari, The distribution of the Chi square test of fit statistics, *The Statistician*, Vol. 13, No. 4 (1963.), 263. - 267.