SEMINAR

RENYI DIVERGENCA

Aljaž Ostrež

Fakulteta za matematiko in fiziko, Ljubljana

Institut Jožefa Stefana, Ljubljana

14. junij 2020

Kazalo

Uvod

Verjetnost in statistika sta veji matematike, ki močno povezujeta teoretično matematiko in praktično uporabo matematike v znanosti.

Zakaj pogostokrat v istem stavku govorimo o verjetnosti in statistiki? Statistika in verjetnost sta zelo povezana pojma, saj v statistični teoriji opišemo naključnost in negotovost s pomočjo teorije verjetnosti.

Posebna zahvala gre prof. dr. Miranu Černetu, ki mi je pomagal pri dokazu, da je Renyi divergenca res divergenca.

1 Osnovni statistični pojmi

Na kratko predstavimo statistične pojme, ki jih moramo razumeti za nadaljna poglavja. Definicije niso formalne, saj je namen tega poglavja bralcu razumljivo predstaviti te pojme na enostaven in pregleden način. Bralec, ki dobro pozna osnovne pojme statistike in teorije verjetnosti, lahko to poglavje izpusti.

Začnimo z definicijami osnovnih statističnih pojmov:

Definicija 1.1

1. ***Statistična populacija*** *(ali* ***populacija****) je množica vseh proučevanih elementov. Potrebna je natančna časovna in krajevna opredelitev.*
2. *Dovolj velika podmnožica populacije, tj. podmnožica, na osnovi katere lahko sklepamo lastnosti celotne populacije, imenujemo* ***vzorec****.*
3. *Posameznemu proučevanemu elementu populacije/vzorca pravimo* ***enota****.*
4. ***Spremenljivka*** *je lastnost enot.*

Poglejmo si primer statistične populacije:

Zgled: *Vzemimo populacijo vseh študentov v 2. letniku matematike na Fakulteti za matematiko in fiziko (krajevna opredelitev) v šolskem letu 2018/2019 (časovna opredelitev). Če so študenti enakomerno razdeljeni v dve skupini za vaje, je vsaka skupina vzorec populacije. Posamezen študent je enota te populacije, kot spremenljivko pa lahko vzamemo na primer višino, težo, spol, povprečje ocen v tekočem študijskem letu itd.*

Spremenljivka je nasprotje od **konstante**; medtem ko ima konstanta le eno vrednost, spremenljivka zavzame različne vrednosti. Kot je že bilo omenjeno, je spremenljivka lastnost enote statistične populacije. Poznamo več delitev spremenljivk, zaradi naših potreb se bomo omejili le na **delitev spremenljivk glede na tip izražanja vrednosti**:

1. **Opisne (atributivne) spremenljivke** - vrednosti lahko opišemo z besedami (npr. spol, barva oči, ...).

2. **Številske (numerične) spremenljivke** - vrednosti izražamo s števili. Številske spremenljivke delimo še na dve veji:

    - **zvezne**: v teoriji lahko zavzamejo katerokoli vrednost na nekem intervalu (npr. teža, pretečen čas, ...),
    - **diskretne**: na nekem intervalu lahko zavzamejo le določene vrednosti (npr. število študentov v posameznem letniku).

Poznamo različne tipe statističnih analiz glede na število hkrati analiziranih spremenljivk:

- **univariatna** - analiza ene spremenljivke,
- **bivariatna** - analiza dveh spremenljivk,
- **multivariatna** - analiza več spremenljivk.

Omejili se bomo na univariatno analizo. Omenimo še definicijo slučajne spremenljivke:

Definicija 1.2 ***Slučajna spremenljivka*** *je količina, ki nastopi kot rezultat poskusa, kjer je možnih več izidov.*

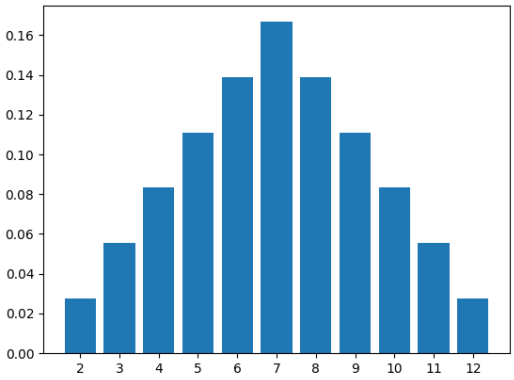
V statistiki lahko razumemo slučajno spremenljivko kot spremenljivko naključno izbrane enote statističnega vzorca.

Poglejmo si še definicije porazdelitve, verjetnostne funkcije in gostote, ker bomo te pojme masovno uporabljali v naslednjih poglavjih.

Definicija 1.3 *Če za neko slučajno spremenljivko poznamo vse možne izide in vemo, kako pogosto jih zavzame, pravimo, da poznamo njeno* ***verjetnostno porazdelitev****.*

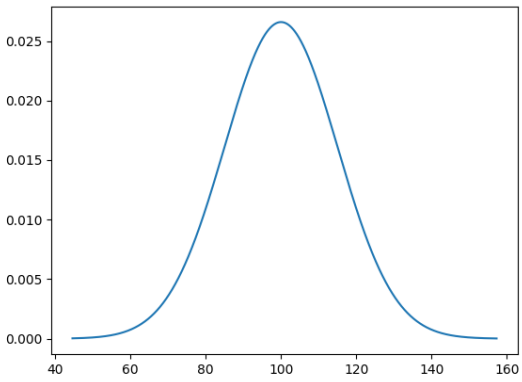
Zgled:

1. *Porazdelitev disretne spremenljivke - rezultat meta dveh kock*



Slika 1: Porazdelitev meta dveh igralnih kock.

2. *Porazdelitev zvezne spremenljivke - inteligenčni količnik (IQ)*



Slika 2: Porazdelitev IQ v populaciji ljudi.

Vse možne izide porazdelitve in pogostost izidov nam predstavi funkcija verjetnosti za diskretne spremenljivke oziroma funkcija gostote verjetnosti za zvezne spremenljivke. Navedimo natančnejši definiciji:

Definicija 1.4

- ***Funkcija gostote verjetnosti*** *(ali samo gostota verjetnosti, angl. probability density function) zvezne slučajne spremenljivke X je funkcija $f : \mathbb{R} \to \mathbb{R}^+$, tako da za vsaki realni števili $a \leq b$ velja:*

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx. \tag{1}$$

Drugače povedano, ploščina gostote verjetnosti med a in b nam pove, kakšna je verjetnost, da se zgodijo dogodki med a in b. Če nas zanima verjetnost dogodka a, nam to pove kar število $f(a)$. Veljati mora še $f(x) \geq 0$ za vse x in $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

- ***Funkcija verjetnosti*** *(angl. probability mass function) diskretne slučajne spremenljivke X je funkcija $p : \mathbb{R} \to [0,1]$, definirana kot*

$$p(a) = P(X = a) \tag{2}$$

za $a \in \mathbb{R}$, tj. funkcija, ki nam pove, kakšna je verjetnost da se zgodi dogodek a za vse a. Veljati mora še $\sum_i p(a_i) = 1$, ko i preteče vse možnosti.

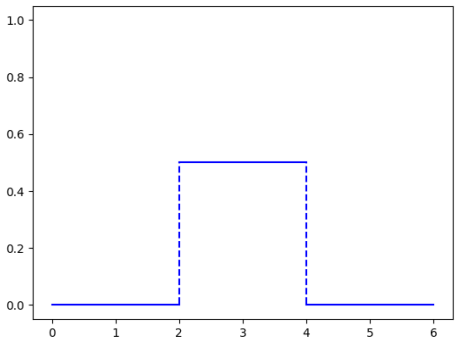
S tem, ko poznamo funkcijo verjetnosti oz. gostoto verjetnosti, poznamo tudi porazdelitev slučajne spremenljivke.

Poglejmo si še nekaj porazdelitev zveznih spremenljivk.

Zgled:

1. *Verjetno najosnovnejša porazdelitev je* ***uniformna porazdelitev****. Gostota verjetnosti te porazdelitve na intervalu $[a,b]$ je definirana kot:*

$$p(x) = \begin{cases} \frac{1}{|b-a|} & , \quad x \in [a,b] \\ 0 & , \quad sicer \end{cases} \tag{3}$$



Slika 3: Uniformna porazdelitev na intervalu $[2,4]$.

2. *Zgoraj smo se že srečali s porazdelitvijo inteligenčnega količnika. To porazdelitev imenujemo je podana z* ***normalna*** *(ali* ***Gaussova****)* ***porazdelitev****, njena gostota verjetnosti pa je funkcija:*

$$\mathcal{N}^{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{4}$$

kjer je μ *srednja vrednost (maksimum porazdelitve - pri inteligenčnem količniku je* 100*),* σ *pa standardni odklon statistične populacije:*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}, \tag{5}$$

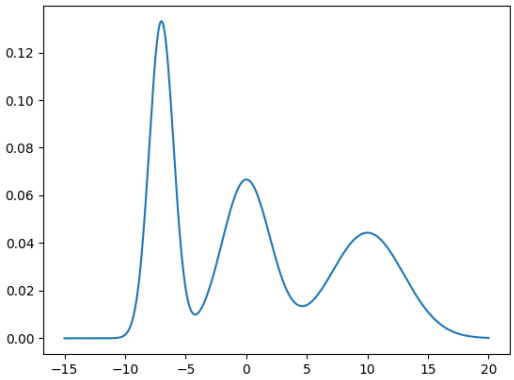
kjer je x_i *i-ta spremenljivka populacije,* $\bar{x}$ *povprečje populacije in* N *velikost populacije (pri inteligenčnem količniku je* $\sigma = 15$*). Omenimo še pojem variance.* ***Varianca*** *je kvadrat standardnega odklona:* σ^2*.*

3. *Kot zadnji primer si oglejmo porazdelitev* ***Gaussovske mešanice****. Gostota te porazdelitve je podana kot:*

$$p(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathcal{N}_i^{\mu_i,\sigma_i}(x), \tag{6}$$

kjer so $\mathcal{N}_i^{\mu_i,\sigma_i}$ normalne porazdelitve. Poglejmo si konkreten primer:

$$p(x) = \frac{1}{3} \cdot (\mathcal{N}^{-7,1}(x) + \mathcal{N}^{0,2}(x) + \mathcal{N}^{10,3}(x)). \tag{7}$$



Slika 4: Porazdelitev Gaussovske mešanice (7)

Za zaključek pregleda osnovnih pojmov omenimo še, da bomo v nadaljevanju velikokrat govorili o porazdelitvi in v mislih imeli gostoto verjetnosti te porazdelitve. Tako bomo ta dva pojma včasih kar enačili.

2 Divergenca glede na gostoto verjetnosti

Divergenca ima lahko več pomenov. V analizi se ta pojem uporablja pri obravnavanju zaporedij in vrst (takrat govorimo o konvergenci in divergenci zaporedja oz. vrste). V tem poglavju bomo predstavili še drugi pomen divergence.

2.1 Definicija divergence

Najprej navedimo motivacijski zgled za uporabo divergence.

Zgled: *V podjetju s proizvodnjo na tekočem traku nas zadolžijo za oskrbo tekočega traku. Naša naloga je, da sporočimo napako, ko jo zaznamo. Napake lahko zaznamo z odstopanjem izhodnih podatkov (npr. temperatura, vibracije, glasnost, itd.).*

Nastane problem: kako sploh primerjati trenutne podatke z optimalnimi in kdaj je odstopanje podatkov dovolj veliko? Potrebujemo metodo, s pomočjo katere bomo lahko primerjali razliko med dvema setoma podatkov. Poleg tega moramo z zagotovostjo trditi, da je prišlo do napake. Pomagamo si z divergenco.

Torej divergenca je merilo za razliko med dvema statističnima vzorcema (vzorca si lahko predstavljamo kot dve porazdelitvi). Navedimo formalno definicijo.

Definicija 2.1 *Naj bo S prostor vseh verjetnostnih porazdelitev na istem definicijskem območju (tj. porazdelitve z istimi nosilci - na istem območju niso enake 0).* ***Divergenca*** *na S je funkcija $D(\cdot\|\cdot) : S \times S \to \mathbb{R}$, tako da velja:*

1. $D(p\|q) \geq 0$ *za vsaka* $p, q \in S$,
2. $D(p\|q) = 0 \Leftrightarrow p = q$.

Dualna divergenca D^* *je definirana kot* $D^*(p\|q) = D(q\|p)$.

Divergenca ni nujno simetrična in zanjo ne velja trikotniška neenakost, zato je ne moremo enačiti z metriko.

2.2 Divergenca univariatnih porazdelitev

Definiranih je več različnih divergenc. Vsaka divegenca ima neke koristne lastnosti (npr. ena divergenca je bolj občutljiva glede na srednje vrednosti porazdelitev, tj. bo velika, ko se bodo srednje vrednosti razlikovale; spet druga divergenca je bolj občutljiva na varianco porazdelitev).

Renyi divergenci se bomo posvetili v naslednjem poglavju. Brez dokazov, da je to res divergenca, navedimo še nekaj ostalih primerov divergenc. Navedimo samo formule za zvezne spremenljivke, saj je formula za diskretne spremenljivke analogna (namesto integrala uporabimo vsoto).

Zgled:

1. ***Kullback-Leiblerjeva divergenca**:*

$$D_{KL}(p\|q) = \int_{\Omega} p(x) \cdot \log\Big(\frac{p(x)}{q(x)}\Big)\, dx, \tag{8}$$

   kjer sta p in q porazdelitvi z definicijskim območjem Ω.

2. ***f-divergenca**: To je družina divergenc, generirana s funkcijami f, za katere velja:*

   - *f konveksna na $\mathbb{R}^+$,*
   - *$f(1) = 0$.*

   Elementi te družine so oblike:

$$D_f(p\|q) = \int_{\Omega} p(x) \cdot f\Big(\frac{p(x)}{q(x)}\Big)\, dx, \tag{9}$$

   kjer sta p in q porazdelitvi definirani na definicijskem območju Ω.

3. ***Hellingerjeva distanca**:*

$$H^2(p,q) = 2\int_\Omega \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \, dx, \tag{10}$$

kjer sta p in q porazdelitvi definirani na definicijskem območju Ω.

2.3 Renyi divergenca

Glede na to, da obstaja veliko različnih divergenc, si nekoliko podrobneje oglejmo **Renyi divergenco**. Omejili se bomo le na zvezne spremenljivke, saj je definicija za diskretne analogna.

Definicija 2.2 *Naj bosta P in Q porazdelitvi z definicijskim območjem Ω, p in q po vrsti gostoti verjetnosti porazdelitev P in Q ter $\alpha > 0$, $\alpha \neq 1$. Tedaj je **Renyi divergenca** definirana kot:*

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \cdot \log \int_\Omega \Big(p(x)\Big)^\alpha \Big(q(x)\Big)^{1-\alpha} \, dx. \tag{11}$$

Dokažimo, da je Renyi divergenca res divergenca.

Dokaz. Naj bo S prostor gostot verjetnosti. Dokazati moramo:

1. $D(p\|q) \geq 0$ za vsaka $p, q \in S$,
2. $D(p\|q) = 0 \Leftrightarrow p = q$.

Najprej dokažimo, da je Renyi divergenca vedno pozitivna. Namesto $p(x)$ in $q(x)$ pišimo kar p in q. Dokazati moramo:

$$\frac{1}{\alpha - 1} \log \int_\Omega p^\alpha q^{1-\alpha} \, dx \geq 0. \tag{12}$$

Ločimo primere:

- $\alpha > 1$: ker je prvi faktor v neenakosti 12 pozitiven za $\alpha > 1$, je ekvivalentno dokazati, da

$$\int_\Omega p^\alpha q^{1-\alpha} \, dx \geq 1$$

oziroma

$$\int_\Omega \left(\frac{p}{q}\right)^\alpha q\, dx \geq 1.$$

Uporabimo Jensenovo neenakost za konveksno funkcijo ϕ:

$$\phi(\int f(x)dx) \leq \int (\phi \circ f)(x)dx,$$

kjer izberemo funkcijo $\phi(t) = t^\alpha$:

$$\int_\Omega \left(\frac{p}{q}\right)^\alpha q\, dx \geq \left(\int_\Omega \frac{p}{q} q dx\right)^\alpha = \left(\int p dx\right)^\alpha = 1,$$

kjer smo upoštevali, da je $\int_\Omega p dx = 1$ po definiciji gostote verjetnosti.

- $0 < \alpha < 1$: ker je prvi faktor v neenakosti 12 negativen za $0 < \alpha < 1$, je ekvivalentno dokazati, da

$$0 < \int_\Omega p^\alpha q^{1-\alpha}\, dx \leq 1$$

oziroma

$$0 < \int_\Omega \left(\frac{q}{p}\right)^{1-\alpha} p\, dx \leq 1.$$

Uporabimo Jensenovo neenakost za konkavno funkcijo ϕ:

$$\phi(\int f(x)dx) \geq \int (\phi \circ f)(x)dx,$$

kjer izberemo funkcijo $\phi(t) = t^{1-\alpha}$:

$$\int_\Omega \left(\frac{q}{p}\right)^{1-\alpha} p\, dx \leq \left(\int_\Omega \frac{q}{p} p\, dx\right)^{1-\alpha} = \int_\Omega q\, dx = 1,$$

kjer smo upoštevali, je $\int_\Omega q dx = 1$ po definiciji gostote verjetnosti.

Dokažimo še drugo točko, torej

$$\frac{1}{\alpha-1}\log\int_\Omega p^\alpha q^{1-\alpha}\,dx = 0 \Leftrightarrow p = q.$$

Najprej dokažimo implikacijo iz desne proti levi ($\Leftarrow$):

$$\frac{1}{\alpha-1}\log\int_\Omega \left(\frac{p}{p}\right)^\alpha p\,dx = \frac{1}{\alpha-1}\log\int_\Omega p\,dx = \frac{1}{\alpha-1}\log 1 = 0.$$

V drugo smer naredimo kratek premislek. Izraz na levi strani ekvivalence bo enak 0, ko:

1. $\alpha = 0$, kar je v protislovju s predpostavko, da je $\alpha > 0$,
2. $p = q$, saj bo takrat $\log\int_\Omega p\,dx = \log 1 = 0$.

Zadnja implikacija je dokazana površno, saj bi se lahko zgodilo tudi, da implikacija ($\Rightarrow$) iz 2. točke velja, če $p \neq q$. Zaključimo, da se to zaradi lastnosti gostot verjetnosti ne more zgoditi. □

Renyi divergenca ni definirana v $\alpha = 1$, a v tej točki poznamo njeno vrednost:

Izrek 2.1 *Naj bo $D_\alpha(P\|Q)$ Renyi divergenca porazdelitev P in Q. Tedaj velja:*

$$\lim_{\alpha\to 1} D_\alpha(P\|Q) = \int_\Omega p(x)\cdot\log\left(\frac{p(x)}{q(x)}\right)dx, \tag{13}$$

*kjer je izraz na desni ravno **Kullback-Leiblerjeva** divergenca porazdelitev P in Q, tj.*

$$\lim_{\alpha\to 1} D_\alpha(P\|Q) = D_{KL}(P\|Q). \tag{14}$$

Dokažimo izrek 2.1:

Dokaz. Izračunajmo limito $D_\alpha(P\|Q)$, ko gre α proti 1.

$$\lim_{\alpha\to 1}\frac{\log\int p(x)^\alpha q(x)^{1-\alpha}\,dx}{\alpha-1}=\text{“}\frac{0}{0}\text{”},$$

zato lahko uporabimo L'Hospitalovo pravilo:

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}.$$

Posebej izračunajmo odvoda števca in imenovalca. Odvod imenovalca je trivialen: $(\alpha-1)'=1$. Odvajajmo še števec:

$$\begin{aligned}\frac{d}{d\alpha}\Big(\log\int p(x)^\alpha q(x)^{1-\alpha}\,dx\Big)&=\frac{1}{\int p(x)^\alpha q(x)^{1-\alpha}\,dx}\quad\frac{d}{d\alpha}\int p(x)^\alpha q(x)^{1-\alpha}\,dx\overset{(*)}{=}\\&\overset{(*)}{=}\underbrace{\frac{1}{\int p(x)^\alpha q(x)^{1-\alpha}\,dx}}_{\overset{\alpha\to 1}{\longrightarrow}1}\quad\int\frac{\partial}{\partial\alpha}\Big(p(x)^\alpha q(x)^{1-\alpha}\,dx\Big)=\\&=\int\Big(p(x)^\alpha\cdot\log p(x)\cdot q(x)^{1-\alpha}-p(x)^\alpha\cdot q(x)^{1-\alpha}\cdot\log q(x)\Big)dx=\\&=\int p(x)^\alpha\cdot q(x)^{1-\alpha}\cdot\log\frac{p(x)}{q(x)}\;dx,\end{aligned}$$

kjer smo pri $(*)$ upoštevali, da je $F(\alpha)=p(x)^\alpha\cdot q(x)^{1-\alpha}$ zvezna funkcija, $\int p(x)^\alpha q(x)^{1-\alpha}\,dx$ pa gre proti 1, ko gre α proti 1 po definiciji gostote verjetnosti. Če izračunamo limito kvocienta odvodov, dobimo:

$$\lim_{\alpha\to 1}\frac{\int p(x)^\alpha\cdot q(x)^{1-\alpha}\cdot\log\frac{p(x)}{q(x)}\;dx}{1}=\int p(x)\cdot\log\frac{p(x)}{q(x)}\;dx,$$

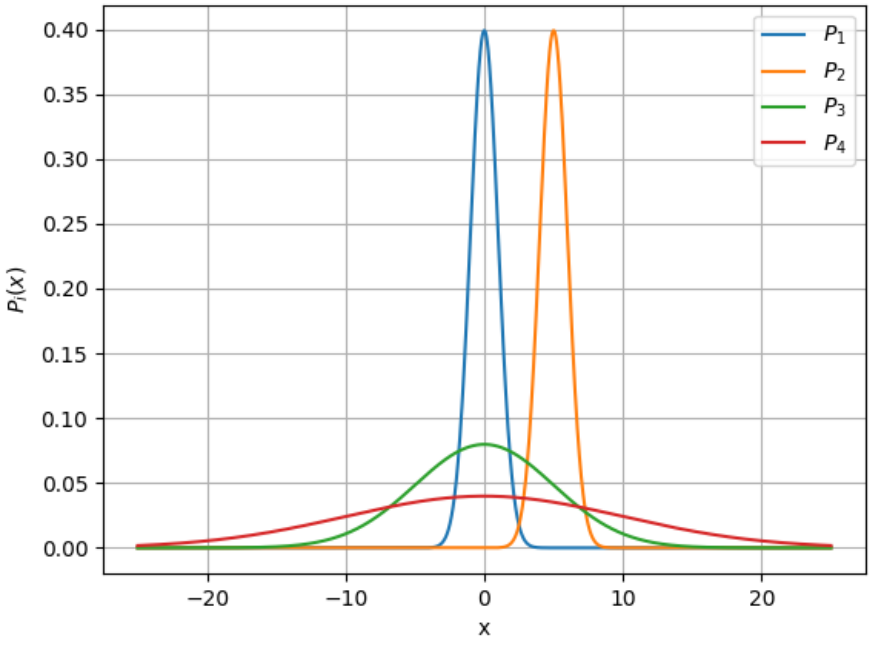
kar je po definiciji ravno Kullback-Leiblerjeva divergenca. □

2.4 Renyi divergenca normalnih porazdelitev

Kot zgled si poglejmo izračun Renyi divergence za nekaj normalnih porazdelitev. Podajmo jih z enačbami:

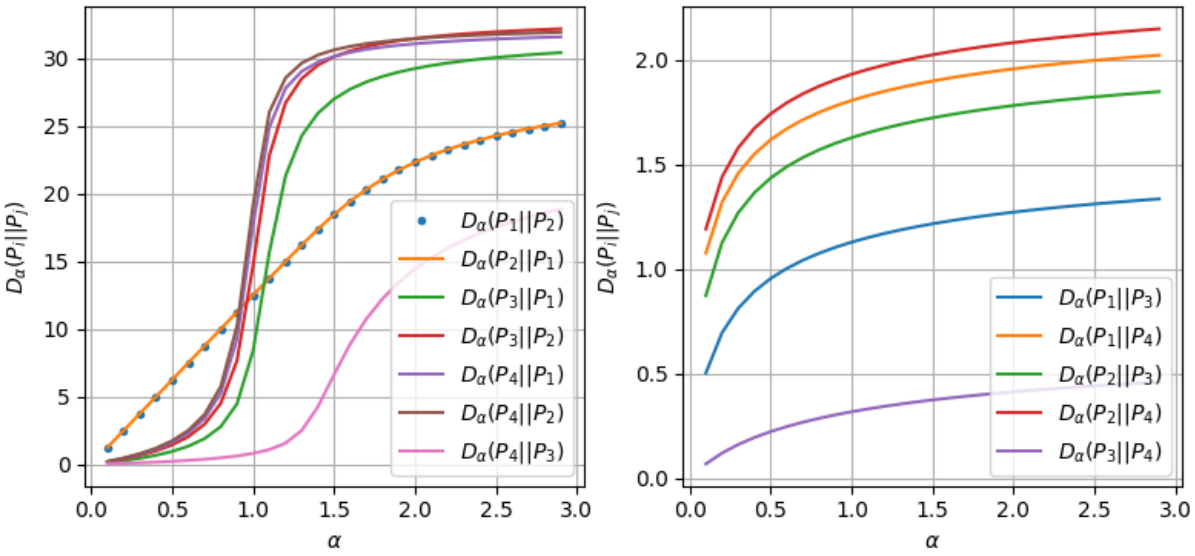
$$\begin{aligned}
\mathcal{N}_1^{0,1}(x) &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \\
\mathcal{N}_2^{5,1}(x) &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-5)^2}{2}} \\
\mathcal{N}_3^{0,5}(x) &= \frac{1}{5\sqrt{2\pi}} \cdot e^{-\frac{x^2}{50}} \\
\mathcal{N}_4^{0,10}(x) &= \frac{1}{10\sqrt{2\pi}} \cdot e^{-\frac{x^2}{200}}.
\end{aligned} \tag{15}$$

kjer je $\mathcal{N}_i^{\mu_i,\sigma_i}$ gostota verjetnosti porazdelitve P_i. Za predstavo si oglejmo še grafe teh verjetnostnih porazdelitev:



Slika 5: Normalne porazdelitve.

Brez simbolnega računanja si poglejmo divergenco med normalnimi porazdelitvami (15).



Slika 6: Divergence normalnih porazdelitev.

Na podlagi izračunov omenimo ugotovitvi:

1. Opazimo, da bo vrednost Renyi divergence zelo velika, če bo varianca prve porazdelitve večja kot varianca druge porazdelitve in majhna, če bo varianca prve porazdelitve manjša kot varianca druge porazdelitve.

2. Če je varianca porazdelitev P_1 in P_2 enaka, je Renyi divergenca simetrična, t.j.

$$D(P_1\|P_2) = D^*(P_1\|P_2) = D(P_2\|P_1).$$

Če lahko sklepamo na podlagi Renyi divergence normalnih porazdelitev smo ugotovili, da na Renyi divergenco zelo vpliva razmerje med variacama prve in druge porazdelitve. Zaradi tega je lahko razlika med vrednostjo Renyi divergence in vrednostjo dualne Renyi divergence zelo velika.

Brez dokaza omenimo še, da translacija sistema ne vpliva na izračun Renyi divergence

normalnih porazdelitev, to pomeni:

$$D_\alpha\Big(\mathcal{N}^{\mu_1,\,\sigma_1}(x)\|\mathcal{N}^{\mu_2,\sigma_2}(x)\Big) = D_\alpha\Big(\mathcal{N}^{\mu_1,\sigma_1}(x+a)\|\mathcal{N}^{\mu_2,\sigma_2}(x+a)\Big), \tag{16}$$

kjer je $a \in \mathbb{R}$.

Opomba: Izračun divergenc in izris grafov sta bila narejena s programskim jezikom `Python3`.

2.5 Numerično računanje Renyi divergence

Kljub temu, da v definiciji divergence zahtevamo, da imata porazdelitvi ista nosilca oz. sta na istih območjih neničelni, se to v praksi pogostokrat ne obnese.

Prva težava je, da bi mogoče hoteli primerjati tudi porazdelitve, ki nimajo istih nosilcev. Vsako porazdelitev lahko razširimo na $\mathbb{R}$ tako, da ji povsod, kjer ta ni definirana, priredimo vrednost 0. S tem sicer dobimo porazdelitvi z istim difinicijskim območjem, a nimata istih nosilcev. Seveda pa se poraja vprašanje, ali je sploh smiselno primerjati porazelitvi z različnima nosilcema? Na primer, ali je smiselno primerjati uniformno porazdelitev na intervalu $[0,1]$ z uniformno porazdelitvijo na intervalu $[2,3]$? Bralec naj to presodi sam.

Drugo težavo malce bolj analizirajmo, saj se pojavi tudi, ko imata porazdelitvi ista nosilca. Kot primer vzemimo normalni porazdelitvi. Gostota verjetnosti normalne porazdelitve je funkcija $\mathbb{R} \to \mathbb{R}_+$, torej bi morali pri Renyi divergenci dveh normalnih porazdelitev vedno dobiti rezultat v množici realnih števil.

Naj bosta p in q gostoti verjetnosti poljubnih realnih funkcij. Problem nastane v repih normalnih porazdelitev. Kljub temu, da je $p(x) > 0$ in $q(x) > 0$ za vsak $x \in \mathbb{R}$, pride do numeričnih težav (deljenje z ničlo). Zakaj? Pride do podkoračitve (v neki točki nam računalnik zaokroži vrednost $p(x)$ na 0). Vpeljimo pojem konstante numerične ločljivosti.

Definicija 2.3 ***Konstanta numerične ločljivosti*** *je najmanjše pozitivno število, ki ga operacijski sistem še ne zaokroži na 0 (t.j. v operacijskem sistemu najmanjše predstavljivo število).*

Zgled: *V 64-bitnem operacijskem sistemu je konstanta numerične ločljivosti enaka* $\epsilon = 2,220446049250313 \cdot 10^{-16}$. *Torej bodo vsa števila med 0 in* ϵ *zaokrožena na 0.*

Predpostavimo, da operiramo s 64-bitni sistemom. Naj bo od zdaj naprej

$$\epsilon = 2,220446049250313 \cdot 10^{-16}.$$

Torej, normalna porazdelitev p bo neničelna le na intervalu $[x_1, x_2]$, kjer sta x_1 in x_2 rešitvi enačbe $p(x) = \epsilon$, za število a na komplementu tega intervala pa bo $p(a) = 0$. Analogno sklepamo za normalno porazdelitev q.

Zaradi takšnega zaokroževanja pride pri izračunu Renyi divergence do deljenja s številom 0. Spomnimo se formule za izračun Renyi divergence:

$$D_\alpha(p\|q) = \begin{cases} \frac{1}{\alpha-1} \log \int_\Omega p(x)^\alpha q(x)^{1-\alpha} dx & , \quad \alpha \neq 1 \\ \int_\Omega p(x) \log \frac{p(x)}{q(x)} dx & , \quad \alpha = 1 \end{cases} \tag{17}$$

Najprej obravnavajmo težave pri $\alpha \neq 1$. Pride do deljenja z ničlo, ko je $q(x) = 0$ za nek x, ker je

$$q(x)^{1-\alpha} = q(x) \cdot \left(\frac{1}{q(x)}\right)^\alpha. \tag{18}$$

Temu problemu se izognemo tako, da definiramo:

- $\frac{0}{0} = 0$,
- $\frac{a}{0} = \infty$ za $a > 0$,
- $\log \infty = \infty$.

Če se torej zgodi, da so obe gostoti hkrati 0, ni težav. Ampak v trenutku, ko je $q(x) = 0$ in $p(x) \neq 0$, bo $D_\alpha(p\|q) = \infty$.

Prav tako se deljenje z 0 pojavi pri $\alpha = 1$. Poleg zgoraj definiranih pravil, s katerimi se izognemo težavam, definirajmo še:

- $\log 0 = -\infty$.

Opazimo, da lahko kar hitro pridemo do rezultata ∞ oz. $-\infty$. Poglejmo, kaj bi lahko storili, da bi vedno dobili rezultat na intervalu $(-\infty, \infty)$. Moramo pa paziti, saj s tem postopkom nekoliko posegamo v same gostote verjetnosti in na koncu za gostoto verjetnosti p ne bo več veljalo $\int_\Omega p(x)dx = 1$. Če pa bo to posegaje minimalno, kar je odvisno od primera do primera, pa lahko na pravilen način pridemo do rezultata na željenem intervalu $(-\infty, \infty)$.

Opomba: (dogovor) Število a je numerično enako 0 oziroma numerično ničelno, če je enako 0 ali če je $|a| < \epsilon$ (tedaj pride do podkoračitve).

Vzemimo poljubni gostoti verjetnosti p, q in ju razširimo na $\mathbb{R}$. Naj bo m tako število, da velja:

$$\forall x_0 < m, \forall \delta > 0 : \Big(p(x_0) < \epsilon \ \wedge \ q(x_0) < \epsilon\Big) \wedge \Big(p(m+\delta) \geq \epsilon \ \vee \ q(m+\delta) \geq \epsilon\Big),$$

t.j. m je tako število, da sta gostoti verjetnosti numerično ničelni na intervalu $(-\infty, m)$ (zaradi podkoračitve) in m je največje tako število. Vemo, da tako število obstaja, saj je $\lim_{x\to-\infty} p(x) = 0$ za poljubno gostoto verjetnosti p (ploščina med p in x-osjo je 1). Naj bo M tako število, da velja:

$$\forall x_0 > m, \forall \delta > 0 : \Big(p(x_0) < \epsilon \ \wedge \ q(x_0) < \epsilon\Big) \wedge \Big(p(m+\delta) \geq \epsilon \ \vee \ q(m+\delta) \geq \epsilon\Big),$$

t.j. M je tako število, da sta gostoti verjetnosti numerično ničelni na intervalu (M, ∞) (zaradi podkoračitve) in M je najmanjše tako število. Vemo, da tako število obstaja, saj je $\lim_{x\to\infty} p(x) = 0$ za poljubno gostoto verjetnosti p.

Na komplementu intervala $[m,M]$ bodo zaradi numerične ločljivosti obe gostoti enaki 0. Na intervalu $[m,M]$ pa definirajmo novi funkciji:

$$f(x) := \begin{cases} p(x) & , \ p(x) > \epsilon \\ \\ \epsilon & , \ p(x) \leq \epsilon \end{cases} \qquad \text{in} \qquad g(x) := \begin{cases} q(x) & , \ q(x) > \epsilon \\ \\ \epsilon & , \ q(x) \leq \epsilon \end{cases} .$$

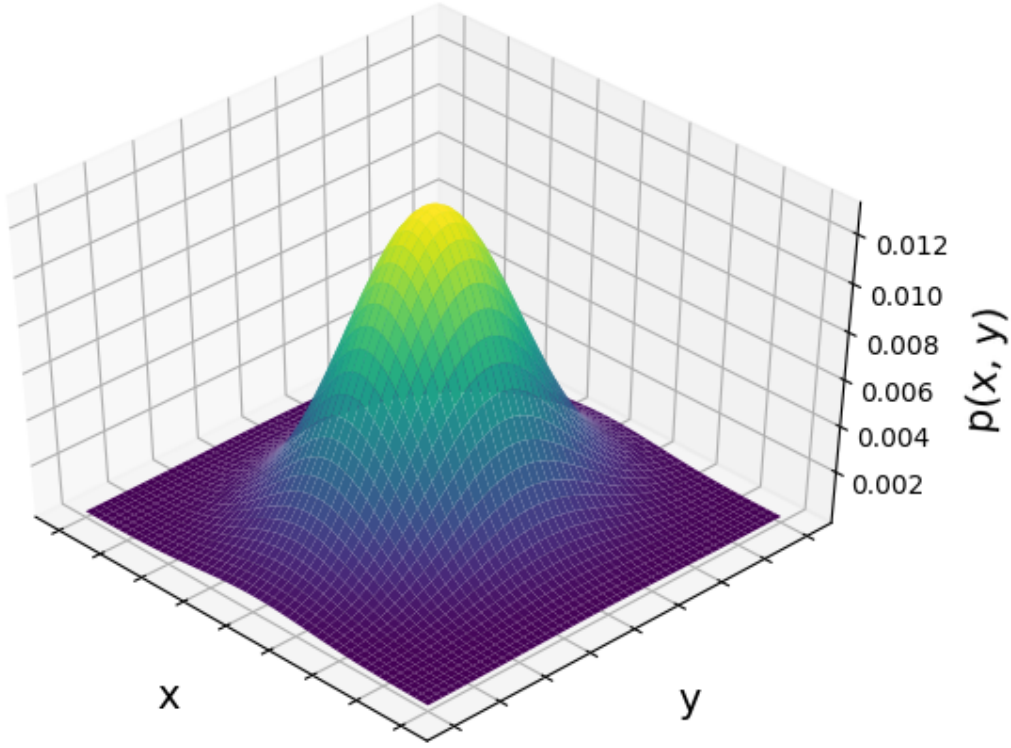
Funciji f in g se od p in q razlikujeta le tam, kjer sta p in q zaradi podkoračitve enaka 0. Preveriti moramo le še, da je $\int_{[m,M]} f(x)dx \approx 1$ in $\int_{[m,M]} g(x)dx \approx 1$. To bo res, če bo skupna dolžina podintervalov na $[m,M]$, kjer sta $p,q < \epsilon$, majhna. Sedaj lahko naredimo aroksimacijo: $D_\alpha(p\|q) \approx D_\alpha(f\|g)$, kjer zagotovo ne pride do deljenja z 0. Kljub temu pa bodo te vrednosti zaradi deljenja z ϵ lahko zelo velike.

Opomba: Kdaj lahko $\int_{[m,M]} f(x)dx$ zaokrožimo na 1, je odvisno od tega, kakšna napaka je za nas še zadovoljiva. Na primer, če je skupna dolžina podintervalov na $[m,M]$, kjer je $p(x) < \epsilon$, enaka 1000, se bo vrednost $\int_{[m,M]} f(x)dx$ absolutno razlikovala od 1 za približno $2,22 \cdot 10^{-13}$.

2.6 Divergenca bivariatnih in multivariatnih porazdelitev

Do sedaj smo obravnavali divergenco univariatnih porazdelitev, torej gostota verjetnosti je bila funkcija ene spremenljivke (npr. P porazdelitev, $p : \mathbb{R} \to \mathbb{R}$ njena gostota verjetnosti). Kaj pa če gledamo porazdelitve v več dimenzijah? Za začetek si poglejmo zgled.

Zgled: *Poglejmo si zgled bivariatne normalne porazdelitve. Kot primer v vsakdanjem življenju lahko vzamemo npr. populacijo ljudi in hkrati gledamo višino in težo.*



Slika 7: Normalna porazdelitev v dveh dimenzijah.

Izpustimo definicije in ostale primere porazdelitev, poglejmo raje, kako definiramo divergenco. Definicija divergence za bi- in multivariate porazdelitve sovpada z definicijo

divergence za univariatne porazdelitve:

1. $D(p\|q) \geq 0$ za vsaka $p, q \in S$, kjer je S množica porazdelitev,

2. $D(p\|q) = 0 \Leftrightarrow p = q$.

Recimo, da gledamo porazdelitve v n dimenzijah. Gostote verjetnosti teh porazdelitev bodo funkcije $p : \mathbb{R}^n \to \mathbb{R}$. Divergence se torej na naraven način prevedejo za n-dimenzionalne porazdelitve. Poglejmo si to na primerih.

Zgled: *Naj bo $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Naj bodo p, q gostote verjetnosti z definicijskim območjem $\Omega \subseteq \mathbb{R}^n$. Definicije divergenc razširimo na n dimenzij:*

1. ***Kullback-Leiblerjeva divergenca***

$$D_{KL}(p\|q) = \int_{\Omega} \cdots \int p(x) \cdot \log\left(\frac{p(x)}{q(x)}\right) dx_1 \ldots dx_n. \tag{19}$$

2. ***f-divergenca***

$$D_f(p\|q) = \int_{\Omega} \cdots \int p(x) \cdot f\left(\frac{p(x)}{q(x)}\right) dx_1 \ldots dx_n. \tag{20}$$

3. ***Hellingerjeva distanca***

$$H^2(p,q) = 2 \cdot \int_{\Omega} \cdots \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx_1 \ldots dx_n. \tag{21}$$

4. ***Renyi divergenca***

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1} \cdot \log\left(\int_{\Omega} \cdots \int \left(p(x)\right)^{\alpha} \left(q(x)\right)^{1-\alpha} dx_1 \ldots dx_n\right), \qquad \alpha \neq 1. \tag{22}$$

2.7 Praktična uporaba

Divergenca glede na gostoto verjetnosti je zelo uporabna v praktičnih problemih, ko glede na set podatkov naredimo oceno gostote verjetnosti z Gaussovimi jedri (angl. Gaussian kernel density estimation).

V splošnem bomo hoteli izračunati divergenco med dvema statističnima vzorcema. Skoraj nikoli divergence ne računamo analitično, saj pri obdelavi podatkov nimamo podane gostote verjetnosti teh podatkov eksplicitno in simbolno. Divergenco bomo izračunali z uporabo histogramov vzorcev ali pa z uporabo ocenjevalcev za gostoto verjetnosti.

3 Predstavitev podatkov vzorca

Naredimo preskok iz teorije v prakso. Imamo dva vzorca podatkov in želimo izračunati divergenco med njima. Pojavi se problem, saj ne vemo, kateri porazdelitvi ustrezata dana vzorca, torej nimamo teoretičnih gostot vertjetnosti. Poiskati želimo približek za gostoto verjetnosti. To naredimo z vpeljavo histogramov ali z vpeljavo različnih estimatorjev za gostoto verjetnosti.

3.1 Histogram

Naj bo S množica podakov. Razdelimo interval $[\min(S), \max(S)]$ na $n \in \mathbb{N}$ intervalov z dolžino večjo od 0:

$$[\min(S), \max(S)] = [a_0, a_1) \cup [a_1, a_2) \cup \ldots \cup [a_{n-2}, a_{n-1}) \cup [a_{n-1}, a_n].$$

Omenimo, da $\min(S) = a_0$ in $\max(S) = a_n$. Naj bo

$$N : \{0, \ldots, n-1\} \to \mathbb{N}$$

funkcija s predpisom:

$$N(i) = \sum_{x \in S} \delta_i,$$

kjer je

$$\delta_i = \begin{cases} 1, & x \in [a_i, a_{i+1}) \\ 0, & sicer \end{cases}.$$

Če je $i = n-1$, je v pogoju pri δ_{n-1} interval zaprt. Funkcija N torej prešteje, koliko elementov iz S je znotraj intervala $[a_i, a_{i+1})$.

Definicija 3.1 ***Histogram*** *je funkcija* $H : \mathbb{R} \to \mathbb{N}$ *s predpisom:*

$$H(x) = \begin{cases} N(i) & , \quad \exists i \in \{0, \ldots, n-1\} \ni : x \in [a_i, a_{i+1}) \\ 0 & , \quad sicer \end{cases}, \tag{23}$$

kjer v pogoju pri $i = n-1$ *dopuščamo zaprti interval.*

Pri obdelavi podatkov je bolj kot zgornja različica histograma uporabna naslednja:

Definicija 3.2 *Naj bo* $N = |S|$. ***Normalizirani histogram*** *je funkcija* $h : \mathbb{R} \to \mathbb{R}_+$ *s predpisom:*

$$h(x) = \begin{cases} \frac{N(i)}{N \cdot (a_{i+1} - a_i)} & , \quad \exists i \in \{0, \ldots, n-1\} \ni : x \in [a_i, a_{i+1}) \\ 0 & , \quad sicer \end{cases}, \tag{24}$$

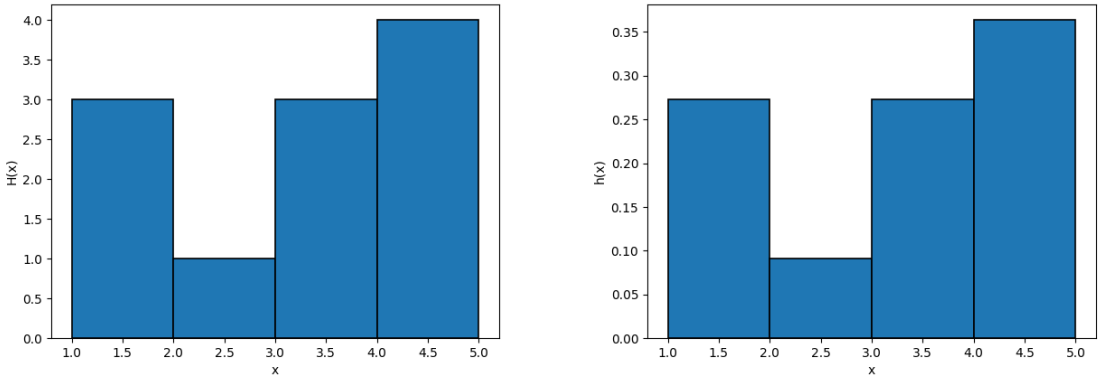
kjer v pogoju pri $i = n-1$ *dopuščamo zaprti interval.*

Opomba: Normalizirani histogram ima lastnost: $\int_{\mathbb{R}} h(x)\,dx = 1$.

Poglejmo si zgled histograma in razliko med histogramom in normaliziranim histogramom.

Zgled: $S_0 = \{1, 1.3, 1, 2.5, 3, 3.5, 3.6, 4, 4.1, 4.2, 5\}$.

$[\min(S_0), \max(S_0)]$ *razdelimo na 4 podintervale:* $[1,2), [2,3), [3,4), [4,5]$. *Poglejmo, kako izgledata histogram in normaliziran histogram (slika 8).*



Slika 8: Na levi je (navaden) histogram, na desni pa normaliziran histogram vzorca S_0. Oblika je ista, razlika je v skali na y-osi.

Od zdaj naprej bomo enačili pojma histogram in normaliziran histogram, z obema pa bomo v mislih imeli normaliziran histogram.

V zgledu opazimo, da histogram grafično ni podan kot funkcija. Opišimo grafičen prikaz histogramov.

3.2 Grafični prikaz histogramov

Histogram je odsekoma konstantna funkcija, kar je razvidno iz definicije ($h(x)$ je konstantna na vsakem intervalu $[a_i, a_{i+1})$). Zato si histogram lahko predstavljamo kot množico pravokotnikov, ki jim pravimo **stolpci** (angl. bins) histograma.

Naj bo na intervalu $[a_i, a_{i+1}), i \in \{0, \ldots, n-1\}$ i-ti stolpec oziroma stolpec i histograma, torej ima histogram n stolpcev. Stolpec i ima širino $w_i = a_{i+1} - a_i$. Če so intervali enako dolgi, imajo vsi stolpci isto širino w. To širino lahko tudi izračunamo:

$$w = \frac{\max(S) - min(S)}{n} = \frac{a_n - a_0}{n}, \tag{25}$$

kjer je n število stolpcev histograma.

Višina i-tega stolpca h_i pa je vrednost funkcije h v poljubni točki znotraj intervala $[a_i, a_{i+1})$, torej:

$$h_i = h(x) \quad \text{za} \quad \forall x \in [a_i, a_{i+1}]. \tag{26}$$

Zgornji način predstavitve histograma ni uporaben le grafično, temveč tudi računsko. Namesto integrala funkcije h lahko izračunamo vsoto ploščin stolpcev, da dokažemo, da je ploščina histograma enaka 1. Numerično se nam bo to izplačalo, saj je numerično integriranje bolj zahtevno kot računanje osnovnih operacij. Enako velja za izračun divergence, kjer uporabimo integracijo.

3.3 Optimalno število stolpcev histograma

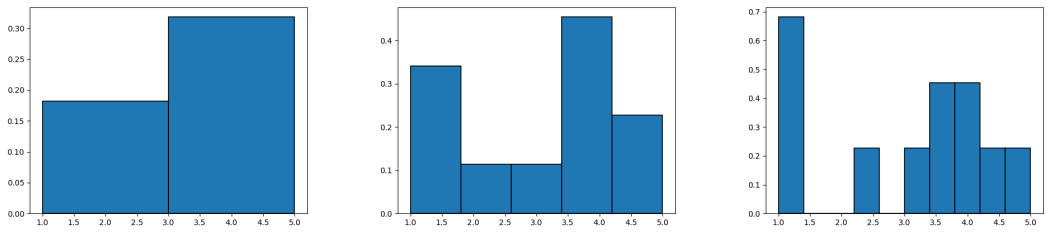
Recimo, da ima množica S porazdelitev P. Naj bo p gostota verjetnosti te porazdelitve, h pa histogram množice S z n stolpci. Velja:

$$\lim_{\substack{|S|\to\infty \\ n\to\infty}} h(x) = p(x) \quad \text{za} \quad \forall x \in S. \tag{27}$$

Histogram je torej lahko ocena gostote verjetnosti množice S.

V nobenem vzorcu nimamo na voljo ∞ podatkov, zato moramo paziti na izbiro števila stolpcev histograma. Če je $n > |S|$, bo najmanj en stolpec prazen (oz. bo imel višino 0), torej bo to lahko zelo slab približek za gostoto verjetnosti (npr. normalna porazdelitev je povsod večja od 0). Druga skrajnost je, ko izberemo premalo stolpcev in tako izgubimo podatke o porazdelitvi (moduse, če jih ima vzorec več).

Zgled: *Vzemimo množico* $S_0 = \{1, 1.3, 1, 2.5, 3, 3.5, 3.6, 4, 4.1, 4.2, 5\}$. *Poglejmo si histograme ob različni izbiri števila stolpcev.*



(a) $n = 2$ **(b)** $n = 5$ **(c)** $n = 10$

Slika 9: Histogrami z različnimi števili stolpcev.

V primeru 9a smo izbrali premalo stolpcev, s čimer lahko zgrešimo pomembne lastnosti množice S. V primeru 9c pa smo izbrali preveč stolpcev, s čimer dobimo prazne stolpce.

Obstaja veliko izbir za optimalno število stolpcev. metod za iskanje optimalnega števila stolpcev. Naštejmo jih. Za vse izbire naj bo S vzorec in $m = |S|$ moč vzorca.

3.3.1 Korenska izbira

Pri tej izbiri optimalno število stolpcev histograma izračunamo po naslednji formuli:

$$n = \lceil \sqrt{m} \rceil. \tag{28}$$

Izračunamo torej kvadratni koren od števila podatkov v vzorcu in zaokrožimo to število na najmanjše celo število, ki je večje od dobljenega korena.

3.3.2 Rice

Pri tej izbiri optimalno število stolpcev za histogram izračunamo na naslednji način:

$$n = \lceil \sqrt[3]{m} \rceil. \tag{29}$$

3.3.3 Sturges

Pri tej izbiri optimalno število stolpcev izračunamo po naslednji formuli:

$$n = \lceil \log_2 m \rceil + 1. \tag{30}$$

Ta formula je zelo varčna, saj je logaritem narašča zelo počasi. Medtem ko bo pri korenski izbiri in 10000 podatkov optimalno število stolpcev 100, bo pri Sturgesovi formuli število stolpcev enako 11.

Vse tri do zdaj naštete izbire delujejo zelo naivno, saj upoštevajo le velikost vzorca, ne pa ostalih njegovih lastnosti, npr. variance. Kljub temu pa se izkaže, da te metode zelo dobro konkurirajo z ostalimi, če gledamo lepe porazdelitve, npr. normalne porazdelitve. Težava so bolj kompleksne porazdelitve, ki imajo visoke repe in so na eni strani omejene. Pri teh hitro dobimo premalo stolpcev in moramo nujno upoštevati še kakšno drugo lastnost, ne le velikosti vzorca.

Poglejmo si sedaj še izbiri, ki upoštevata bolj specifične lastnosti vzorca. Ti dve formuli sicer vrneta optimalno širino stolpca, a zlahka iz tega podatka dobimo optimalno število stolpcev (glej formulo (25)).

3.3.4 Scoot

Formula Scoot poleg velikosti vzorca upošteva še njegov standardni odklon σ:

$$h = 3,49 \cdot \sigma \cdot m^{-1/3}. \tag{31}$$

3.3.5 Freedman-Diaconis

Ta izbira prav tako uporabi še dodatno lastnost poleg velikosti vzorca, in sicer medkvartilno razdaljo vzorca IQR (angl. *interquartile range*), to je razlika med zgornjim in spodnjim kvartilom.

$$h = 2 \cdot IQR \cdot m^{-1/3}. \tag{32}$$

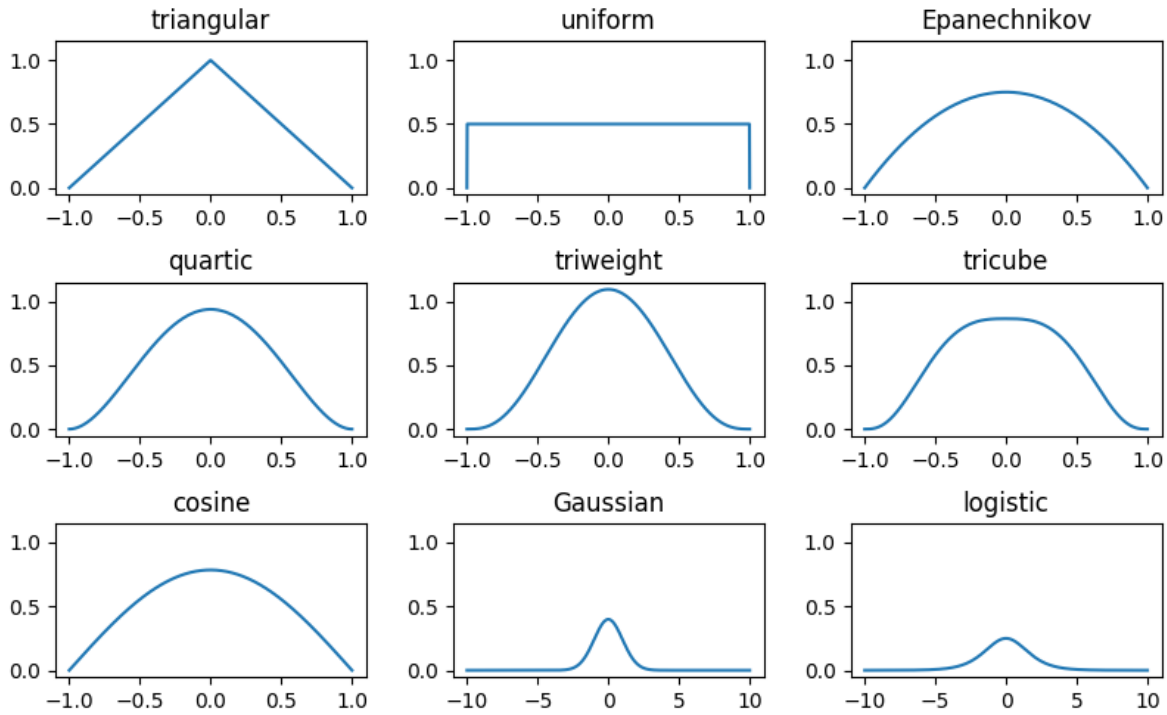
Ker smo že sproti opisali izbire, primer izpustimo. Omenimo pa, da sta zadnji dve metodi v splošnem najboljši, saj poleg velikosti vzorca upoštevata še dodatne lastnosti vzorca, poleg tega pa je izračun variance in medkvartilne razdalje hiter.

3.4 Ocena gostote verjetnosti z jedrom

Opišimo še popolnoma drugačen pristop za predstavitev podatkov. Najprej pa moramo povedati, kaj sploh mislimo z besedo "jedro".

Definicija 3.3 ***Jedro*** *je nenegativna realna integrabilna funkcija K z lastnostima:*

- $\int_{-\infty}^{\infty} K(x)\,dx = 1 \quad in$
- $K(-x) = K(x), \quad \forall x \in \mathbb{R} \quad$ *(simetrija).*



Slika 10: Nekaj primerov jeder.

Ideja ocenjevanja gostote verjetnosti z jedri (angl. *kernel density estimation*) je, da na vsako točko v vzorcu obesimo jedro, ki ga vnaprej določimo, nato pa ta jedra seštejemo in rezultat normiramo. Ker so jedra nenegativna, bo torej pridobljena funkcija ustrezala definiciji gostote verjetnosti.

Definicija 3.4 ***Ocena gostote verjetnosti z jedrom*** *je neparametričen način ocenjevanja gostote verjetnosti. Naj bo K jedro in $X = \{x_1, \ldots, x_n\}$ množica podatkov. Ocena gostote verjetnosti z jedrom K množice X je funkcija, definirana kot:*

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{33}$$

kjer je h parameter glajenja, imenovan "bandwidth".

Parameter h močno vpliva na končno oceno. Če je h premajhen, lahko dobimo močno oscilirajočo funkcijo, ob preveliki izbiri h pa lahko izgubimo pomembne podatke o porazdelitvi. Če potegnemo vzporednico s podpoglavjem o histogramih, lahko rečemo, da h igra podobno vlogo kot izbira optimalnega števila stolpcev pri histogramih. V optimizacijo paramtra h se ne bomo poglabljali, ponavadi pa ga izračunamo kot:

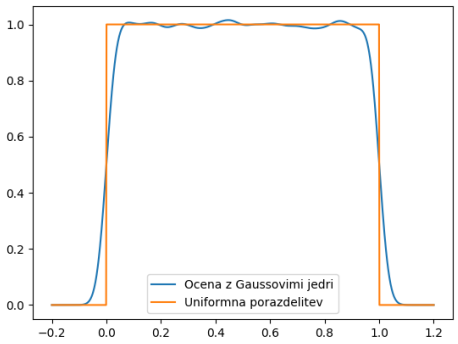
$$h = \sigma(X) \cdot \Big(\frac{4}{3|X|}\Big)^{1/5}, \tag{34}$$

kjer je $\sigma(X)$ standardni odklon množice podatkov X.

Pri ocenjevanju gostote verjetnosti z Gaussovim jedrom uporabljamo Gaussovo jedro (slika 10). To je normalna porazdelitev s srednjo vrednostjo 0 in varianco 1, torej:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{35}$$

Ocenjevanje gostote verjetnosti z Gaussovim jedrom smo izpostavili, ker je zelo pogost. Poleg tega je zelo dobra ocena pri porazdelitvah s trebuhi (npr. normalna porazdelitev, Rayleigh porazdelitev, ...). Je pa ta metoda zelo slaba pri porazdelitev s končnimi nosilci (npr. uniformna podrazdelitev), saj zaradi neomejenosti definicijskega območja Gaussovega jedra funkcija, ki jo dobimo z oceno, preseže ta nosilec.



Slika 11: Vidi se, da ocena z Gaussovimi jedri preseže nosilec uniformne porazdelitve.

Predstavili smo dva načina predstavitve podatkov. Obe načina predstavitve imata svoje prednosti in slabosti. Predstavitev s histogrami je zelo kompaktna, poleg tega pa histogram nikoli ne bo segal ven iz nosilca porazdelitve, saj je omejen s podatkovno množico. Pri ocenjevanju gostote verjetnosti z jedri se pa lahko zelo dobro približamo porazdelitvi, ki jo podatki predstavljajo, s čimer ne bomo izgubili nobene pomembne lastnosti podatkovne množice. Tvegamo pa, da lahko pademo ven iz nosilca.

Za naključen set podatkov, o katerem ne vemo nič, je težko reči, kateri način predstavitve je boljši. Vemo pa, da je za porazdelitve, ki imajo neskončne nosilce $(-\infty,\infty)$ in so gladke, veliko boljša predstavitev z oceno z Gaussovimi jedri. Pri porazdelitvah z omejenimi nosilci pa ta možnost odpade, tako da smo primorani operirati s histogrami.

Za konec tega poglavja pa predstavimo način, kako lahko optimalno število stolpcev iščemo z uporabo divergence. Kot primer vzemimo kar Kullback-Leibler divergenco D_{KL}.

3.5 Kullback-Leibler metoda za optimalno število stolpcev

Naj bo $X = \{x_1, x_2, \ldots, x_n\}$ množica podatkov. Z oceno gostote verjetnosti množice X z Gaussovim jedrom dobimo približek gostote verjetnosti p množice X. Ta približek bomo primerjali s histogramom h_n z n-stolpci s pomočjo Kullback-Leibler divergence po formuli:

$$D_{KL}(p\|h_n) = \int_\Omega p(x) \cdot \log\left(\frac{p(x)}{h_n(x)}\right) \quad dx. \tag{36}$$

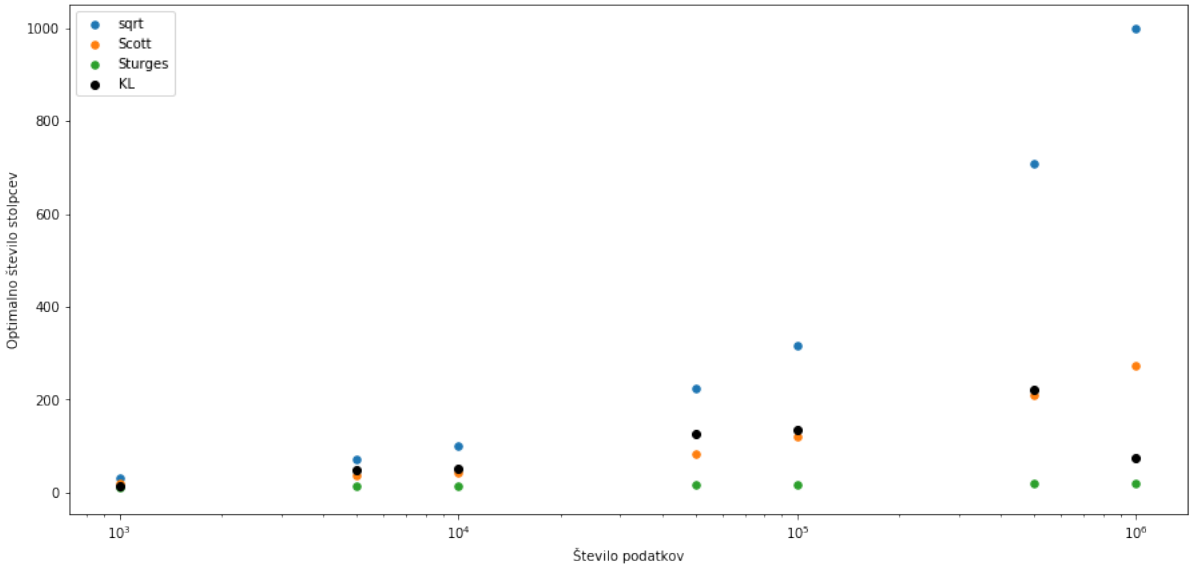
Naredimo iteracijo po številu stolpcev $i = 1, \ldots, m$ za nek $m \in \mathbb{N}$. V vsaki iteraciji izračunajmo $D_{KL}(p\|h_i)$. Tisti i, pri katerem izraz doseže minimum, bomo vzeli za optimalno število stolpcev histograma množice X.

Opomba: Zaradi zahtevnosti numeričnega računanja Kullback-Leibler divergenco raje računamo kot:

$$D_{KL}(p\|h_i) = \int p(x) \cdot ln(p(x))dx - \int p(x) \cdot ln(h_i(x))dx. \tag{37}$$

Na ta način lahko prvi člen izračunamo izven zanke iteracije, kar nam po izračunih privarčuje nekaj časa.

Poglejmo si, kakšne rezultate daje ta metoda v primerjavi z ostalimi.



Slika 12: Primerjava med optimalnim številom stolpcev glede na korensko metodo in metodami Scott, Sturges in Kullback-Leibler. Os s številom podatkov je v logaritemski skali.

Zaradi preglednosti smo metodi Freedman-Diaconis in Rice izpustili. Iz slike je razvidno, da Kullback-Leibler metoda daje optimalna števila stolpcev podobna tistim, ki jih dajo ostale metode. Je pa izračun optimalnega števila stolpcev po Kullback-Leibler metodi zelo počasen, ko je število podatkov veliko. Prav tako pa procesa ne moremo optimizirati, saj funkcija $f(n) = D_{KL}(p\|h_n)$ ni konveksna, kar bi nam omogočilo iskanje minimuma, torej moramo direktno m-krat izračunati vrednost Kullback-Leibler divergence, kar pa je zaradi integriranja počasno.

4 Renyi divergenca glede na histogram

Zdaj, ko dobro poznamo pojem histograma, lahko definiramo Renyi divergenco glede na histogram. Histogram bo torej ocena za gostoto verjetnosti vzorca.

V prejšnjem poglavju smo omenili, da lahko histogram predstavimo kot funkcijo, a bo za računanje Renyi divergence boljše, da si histogram predstavljamo grafično. Torej predstavljamo si ga kot skupek pravokotnikov (stolpcev) z določenimi mejami in višinami.

Izrek 4.1 *Naj bo H_1 histogram, ki ima meje stolpcev $(x_0, x_1, \ldots, x_{n-1}, x_n)$ in višine stolpcev $(h_{(1,0)}, \ldots, h_{(1,n-1)})$. Naj bo H_2 histogram, ki ima meje stolpcev iste kot H_1 in višine stolpcev $(h_{(2,0)}, \ldots, h_{(2,n-1)})$ (torej imata H_1 in H_2 isto število stolpcev in enake meje, le različne višine). Tedaj lahko Renyi divergenco med histogramoma H_1 in H_2 izračunamo kot:*

$$D_\alpha(H_1 \| H_2) = \frac{1}{\alpha - 1} \log\Big(\sum_{i=0}^{n-1} h_{(1,i)}^{\alpha} \cdot h_{(2,i)}^{1-\alpha} \cdot (x_{i+1} - x_i) \Big), \tag{38}$$

če $\alpha \neq 1$, oziroma

$$D_1(H_1 \| H_2) = \sum_{i=0}^{n-1} h_{(1,i)} \cdot \log \frac{h_{(1,i)}}{h_{(2,i)}} \cdot (x_{i+1} - x_i). \tag{39}$$

Dokaz. Vemo: če H_1 in H_2 zapišemo kot funkciji ($H_1 = H_1(x)$ in $H_2 = H_2(x)$), dobljeni funkciji ustrezata pogojem za gostoto verjetnosti. Torej lahko izračunamo Renyi divergenco po definiciji:

$$D_\alpha(H_1 \| \, H_2) = \frac{1}{\alpha - 1} \log \int_{x_0}^{x_n} H_1(x)^{\alpha} \cdot H_2(x)^{1-\alpha} \, dx. \tag{40}$$

Ker sta funkciji H_1 in H_2 nezvezni v istih točkah (v mejah stolpcev), zapišemo integral kot vsoto integralov na intervalih, kjer sta H_1 in H_2 konstantni:

$$\int_{x_0}^{x_1} H_1(x)^{\alpha} \cdot H_2(x)^{1-\alpha} \, dx + \ldots + \int_{x_{n-1}}^{x_n} H_1(x)^{\alpha} \cdot H_2(x)^{1-\alpha} \, dx. \tag{41}$$

Če upoštevamo še konstantne vrednosti funkcij H_1 in H_2 na intervalih integriranja, in v

vsakem členu integriramo $\int_a^b dx = b - a$, dobimo:

$$h^{\alpha}_{(1,0)} \cdot h^{1-\alpha}_{(2,0)} \cdot (x_1 - x_0) + \ldots + h^{\alpha}_{(1,n-1)} \cdot h^{1-\alpha}_{(2,n-1)} \cdot (x_n - x_{n-1}) = \sum_{i=0}^{n-1} h^{\alpha}_{(1,i)} \cdot h^{1-\alpha}_{(2,i)} \cdot (x_{i+1} - x_i), \quad (42)$$

torej je:

$$D_{\alpha}(H_1 \| H_2) = \frac{1}{\alpha - 1} \log\Big(\sum_{i=0}^{n-1} h^{\alpha}_{(1,i)} \cdot h^{1-\alpha}_{(2,i)} \cdot (x_{i+1} - x_i) \Big). \quad (43)$$

Podobno dokažemo za $\alpha = 1$ - izračunamo Kullback-Leibler divergenco. □

Zdaj znamo izračunati Renyi divergenco med dvema histogramoma, a imamo zelo strogo omejitev: meje stolpcev prvega histograma morajo biti enake mejam stolpcev drugega histograma. Tej omejitvi pa se lahko elegantno izognemo.

4.1 Sprostitev definicijskega območja histogramov

Če hočemo, da imata poljubna histograma enake meje stolpcev, moramo najprej uskladiti definicijski območji histogramov.

Naj bo X urejen seznam z mejami stolpcev prvega histograma in Y urejen seznam z mejami stolpcev drugega histograma. Za spodnjo mejo:

- Če je $\min(X) > \min(Y)$, potem na začetek seznama X vstavimo element $\min(Y)$. Tako v prvem histogramu dobimo še en stolpec z mejami $\min(Y)$ in $\min(X)$. Dodelimo mu vrednost 0.

- Če je $\min(X) < \min(Y)$, potem na začetek seznama Y vstavimo element $\min(X)$. Tako v drugemu histogramu dobimo še en stolpec z mejami $\min(Y)$ in $\min(X)$. Dodelimo mu vrednost 0.

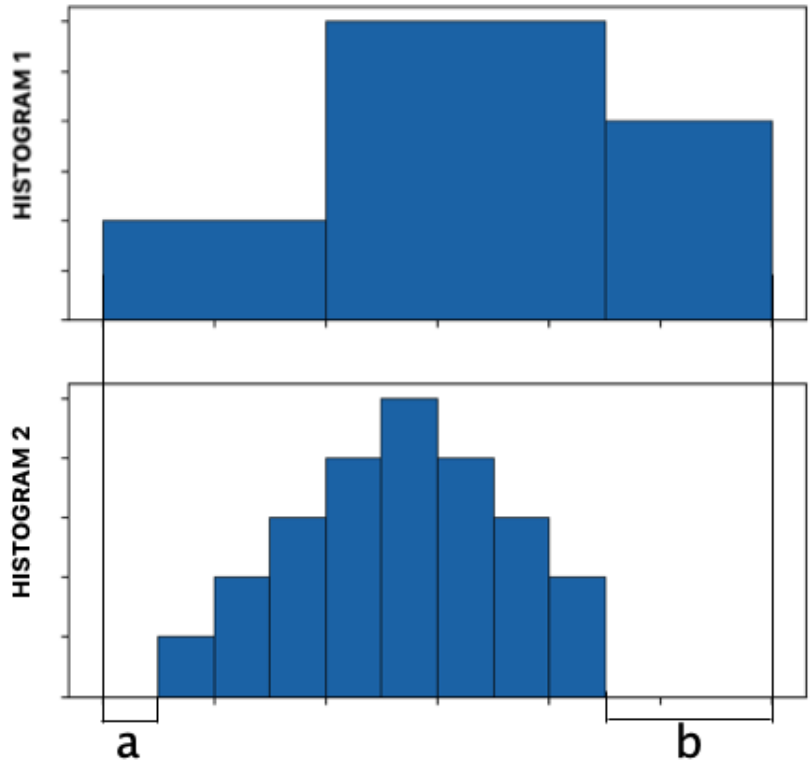
Za zgornjo mejo:

- Če je $\max(X) < \max(Y)$, potem na konec seznama X vstavimo element $\max(Y)$. Tako v

prvem histogramu dobimo še en stolpec z mejami $\max(X)$ in $\max(Y)$. Dodelimo mu vrednost 0.

- Če je $\max(X) > \max(Y)$, potem na konec seznama Y vstavimo element $\max(X)$. Tako v drugem histogramu dobimo še en stolpec z mejami $\max(Y)$ in $\max(X)$. Dodelimo mu vrednost 0.

Dobili smo torej največ dva nova stolpca. Ker smo jima dodelili vrednost 0, ploščina histograma ostaja enaka, torej je bil to dober korak k našemu cilju (da bodo ploščine histogramov čim bližje 1). Definicijsko območje obeh histogramov je na tej točki enako.



Slika 13: Sprostitev definicijskega območja histograma 2. Dobili smo dva nova stolpca a in b.

5 Zaključek

Literatura

[1] Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. in Meester, L. E., 2005. A Modern Introduction to Probability and Statistic: Understanding Why and How. 2. izdaja. London: Springer-Verlag London Limited.

[2] Eguchi, S., 1985. A differential geometric approach to statistical inference on the basis of contrast functional. Dostopno na: https://projecteuclid.org/download/pdf_1/euclid.hmj/1206130775 [9. 3. 2019]

[3] Ferligoj, Anuška, idr. 2015, Osnove statistike na prosojnicah: študijsko gradivo pri predmetu Statistika, Ljubljana: Fakulteta za družbene vede.

[4] Gil, M., 2011. On Rényi Divergence Measures for Continuous Alphabet Sources. Dostopno na: https://pdfs.semanticscholar.org/8ebc/dd48e51d9f18a0025794ae088ce754dd47ce.pdf [9. 3. 2019]

[5] Košmelj, K., 2007. Uporabna statistika. 2. izdaja. Ljubljana: Biotehniška fakulteta.

[6] Sason, I. in Verdú, S., 2018. f-Divergence Inequalities. Dostopno na: http://webee.technion.ac.il/people/sason/f-divergence%20inequalities.pdf [9. 3. 2019]

[7] van Erven, T. in Harremoës, P., 2007. Renyi Divergence and Kullback-Leibler Divergence. Dostopno na: https://arxiv.org/pdf/1206.2459.pdf [9. 3. 2019]