

Verjetnost in statistika - priprava na teoretični del izpita

Jernej Vivod

July 10, 2018

Ta skripta vsebuje strnjene zapiske snovi predmeta Verjetnost in statistika na prvi stopnji Bolonjskega programa študija Računalništvo in informatika na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Povzeta je po skripti profesorja dr. Aleksandra Jurišića ter po predavanjih. Služi lahko kot priprava na teoretični del izpita.

1 Verjetnost

1.1 Osnovno o verjetnosti

Verjetnostni račun obravnava zakonitosti, ki se pokažejo v velikih množicah enakih ali vsaj zelo podobnih pojavov. Osnovni pojmi verjetnosti so: poskus, dogodek in verjetnost dogodka.

Dogodek je slučajen, če so posamezni izidi negotovi, vendar pa je na dolgi rok vzorec velikega števila posameznih izidov napovedljiv. Za slučajnost velja neke vrste red, ki se pokaže šele na dolgi rok, po velikem številu ponovitev.

- gotov dogodek - oznaka G : ob vsaki ponovitvi poskusa se zgodi.
- nemogoč dogodek - oznaka N : nikoli se ne zgodi
- slučajen dogodek: včasih se zgodi, včasih ne.

Dogodek A je poddogodek ali način dogodka B , kar zapišemo $A \subseteq B$, če se vsakič, ko se zgodi dogodek A , zagotovo zgodi tudi dogodek B .

Nekaj lasnosti, ki veljajo za poljubna dogodka A in B :

- $A \cup B = B \cup A$,
- $A \cup A = A$,
- $A \cup N = A$,
- $A \cup G = G$,
- $B \subseteq A \Leftrightarrow A \cup B = A$,
- $A \cup (B \cap C) = (A \cup B) \cap C$,
- $A \cap B = B \cap A$,
- $A \cap A = A$,
- $A \cap N = N$,
- $A \cap G = A$,
- $B \subseteq A \Leftrightarrow A \cap B = B$,
- $A \cap (B \cap C) = (A \cap B) \cap C$,
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Dogodku A nasproten dogodek \bar{A} , je tisti, ki se zgodi natanko takrat, ko se dogodek A ne zgodi, in ga imenujemo tudi negacija dogodka A .

Nekaj lastnosti, ki se tičejo nasprotnih dogodkov:

- $A \cap \bar{A} = N$,
- $A \cup \bar{A} = G$,
- $\bar{\bar{N}} = G, \bar{\bar{G}} = N$,
- $\bar{\bar{A}} = A$,
- $A \cup B = \bar{A} \cap \bar{B}$,
- $A \cap B = \bar{A} \cup \bar{B}$ (De Morganovi pravili).

Če lahko dogodek A izrazimo kot vsoto nezdružljivih in mogočih dogodkov, rečemo, da je A sestavljen dogodek. Dogodek, ki ni sestavljen, imenujemo osnoven ali elementaren dogodek.

Množico dogodkov $S = A_1, A_2, \dots, A_n$ imenujemo popoln sistem dogodkov, če se v vsaki ponovitvi poskusa zgodi natanko eden od dogodkov iz množice S .

Število $f(A) = \frac{k}{n}$ imenujemo relativna frekvenca (pogostost) dogodka A v opravljenih poskusih. Ta vrednost se z naraščanjem števila poskusov ustaljuje.

Pomni 1.1: Statistična definicija verjetnosti

Verjetnost dogodka A v danem poskusu je število $P(A)$, pri katerem se navadno ustali relativna frekvenca dogodka A v velikem številu ponovitev tega poskusa.

Nekaj lastnosti, ki se tičejo tega števila:

- $P(A) \geq 0$,
- $P(G) = 1, P(N) = 0$ in $A \subseteq B \rightarrow P(A) \leq P(B)$,
- Če sta dogodka A in B nezdružljiva, potem je $P(A + B) = P(A) + P(B)$.

1.2 Vzorčni prostor in verjetnostni model

Pomni 1.2: Vzorčni prostor

Vzorčni prostor S slučajnega pojava je množica vseh možnih izidov. V tem kontekstu je dogodek katerikoli izid ali množica izidov slučajnega pojava. Dogodek je torej podmnožica verjetnostnega prostora.

Pomni 1.3: Verjetnostni model

Verjetnostni model je matematični opis slučajnega pojava, sestavljen iz dveh delov: verjetnostnega prostora S in predpisa, ki dogodkom priredi verjetnosti.

Verjetnostni model za končen vzorčni prostor podamo tako, da predpišemo verjetnost vsakemu posameznemu izidu. Te verjetnosti morajo biti števila med 0 in 1 in njihova vsota mora biti enaka

1. Verjetnost poljubnega dogodka je vsota verjetnosti izidov, ki ga sestavljajo.

1.3 Posplošitev znane zveze $P(A+B) = P(A) + P(B) - P(AB)$ na poljubno število dogodkov

Pomni 1.4: Princip inkluzije in ekskluzije posplošen na n dogodkov

$$|A_1 \cup A_2 \cup \dots \cup A_p| = \sum_{1 \leq i \leq p} |A_i| - \sum_{1 \leq i_1 \leq i_2 \leq p} |A_{i_1} \cap A_{i_2}| + \sum_{1 \leq i_1 \leq i_2 \leq i_3 \leq p} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| - \dots + (-1)^{p-1} |A_1 \cap A_2 \cap \dots \cap A_p| \quad (1)$$

1.4 Trije aksiomi Kolmogorova

Pomni 1.5: Aksiomi Kolmogorova

1. Prvi aksiom

Verjetnost dogodka je nenegativno realno število: $P(E) \in \mathbb{R}$,

$$P(E) \geq 0$$

$\forall E \in \mathcal{F}$ Kjer je \mathcal{F} prostor dogodkov.

2. Drugi aksiom

Verjetnost, da se bo vsaj eden od elementarnih dogodkov v celotnem prostoru dogodkov zgodil, je 1.

$$P(\Omega) = 1$$

3. Tretji aksiom

Vsako števno zaporedje disjunktnih množic (medsebojno nezdružljivih dogodkov) zadošči

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

1.5 Pogojna verjetnost

$P(A|B)$ je verjetnost, da se je zgodil dogodek A, če vemo, da se je zgodil dogodek B.

$$\text{velja: } P(A|B) = \frac{P(AB)}{P(B)}$$

Iz te formule sledita naslednji zvezi:

$$P(AB) = P(B)P(A|B) \quad P(AB) = P(A)P(B|A)$$

Velja tudi: $P(A)P(B|A) = P(B)P(A|B)$

Dogodka A in B sta neodvisna, če velja $P(AB) = P(A) \cdot P(B)$, $P(A|B) = P(A)$ in $P(A|\bar{B}) = P(A)$.

Za par nezdružljivih dogodkov velja $P(A|B) = 0$.

Pomni 1.6: Izrek o popolni verjetnosti

a popoln sistem dogodkov $H_i, i \in I$ in poljuben dogodek A velja

$$P(A) = \sum_{i \in I} P(AH_i) = \sum_{i \in I} P(H_i)P(A|H_i)$$

Pomni 1.7: Bayesov obrazec

a popoln sistem dogodkov $H_i, i \in I$ in poljuben dogodek A velja:

$$P\left(\frac{H_k}{A}\right) = \frac{P(H_k) \cdot P(A|H_k)}{\sum_{i \in I} P(H_i) \cdot P(A|H_i)}.$$

Intuicija za Bayesov obrazec na primeru testa za bolezen: Verjetnost, da ima oseba, ki je bila na testu za bolezen pozitivna, res to bolezen, je verjetnost, da je test pozitiven krat verjetnost, da je pozitiven zaradi prisotne bolezni ulomljeno z vsemi možnostmi, da je test pozitiven.

2 Bernoullijevo zaporedje neodvisnih dogodkov

3 Statistika - uvod

Nekaj temeljnih osnovnih pojmov:

Enota - posamezna proučevana stvar ali pojav

Populacija - množica vseh proučevanih enot; pomembna je natančna opredelitev populacije (npr. časovno in prostorsko).

Vzorec - Podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih celotne populacije.

Spremenljivka - lastnost enot; označimo jih npr. z X, Y, X_1 . Vrednost spremenljivke X na i-ti enoti označimo z x_i .

Parameter - zn ačilnost populacije. Običajno jih označujemo z malimi grškimi črkami.

Statistika - značilnost vzorca; običajno jih označujemo z malimi latinskimi črkami. Vrednost

statistike je lahko za različne vzorce različna.

Eno izmed osnovnih vprašanj statistike je, kako z uporabo ustreznih statistik oceniti vrednosti izbranih parametrov.

Frekvenčna porazdelitev spremenljivke je tabela, ki jo določajo vrednosti ali skupine vrednosti in njihove frekvence. Skupine vrednosti številskih spremenljivk imenujemo razredi.

Če zapišemo podatke v vrsto po njihovi numerični velikosti pravimo, da gre za urejeno zaporedje oziroma ranžirano vrsto. Ustrezem mestu v ranžirni vrsti pravimo rang.

3.1 Mere za lokacijo in razpršenost

V tem razdelku bomo obravnavali naslednjih šest ključnih pojmov:

1. srednje vrednosti
2. razpon
3. centili, kvartili
4. varianca
5. standardni odklon
6. Z-vrednosti

Modus (oznaka M_0) je tista vrednost, ki se v množici podatkov pojavi z največjo frekvenco. Mediana je srednja vrednost naraščujoče po velikosti urejene množice podatkov. Če je število podatkov sodo, vzamemo povprečje srednjih dveh vrednosti.

Povprečje populacije: $\mu = \frac{x_1 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$

Povprečje vzorca: $\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$

Razpon je razlika med največjo in najmanjšo meritvijo v množici podatkov.

100p-ti centil ($p \in [0, 1]$), je definiran kot število, od katerega ima 100p% meritev manjšo ali enako numerično vrednost.

Za računanje 100p-tega centila je najlažje, da podatke najprej uredimo po velikosti.

Varianca populacije je povprečje kvadratov odklonov od pričakovane vrednosti

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Varianco vzorca dobimo tako, da vsoto kvadratov odklonov delimo s stopnjo prostosti vzorca, tj. $n - 1$ namesto n .

Standardni odklon (deviacija) je pozitivno predznačen kvadratni koren variance, koeficient variacije pa standardni odklon deljen s povprečjem.

3.2 Normalna porazdelitev

Pomni 3.1: Empirična pravila

Empirična pravila, ki veljajo, če ima podatkovna množica porazdelitev približno zvonaste oblike:

1. približno 68.3% vseh meritev leži na intervalu $[\mu - \sigma, \mu + \sigma]$, kjer je σ standardni odklon in μ povprečje.
2. približno 95.4% vseh meritev leži na intervalu $[\mu - 2\sigma, \mu + 2\sigma]$.
3. približno 99.7% vseh meritev leži na intervalu $[\mu - 3\sigma, \mu + 3\sigma]$.

Pomni 3.2: Centralni momenti

Za $l \in \mathbb{N}$ je l -ti centralni moment enak

$$m_l = \frac{(x_1 - \bar{x})^l + \dots + (x_n - \bar{x})^l}{n}. \quad m_1 = 0, m_2 = \sigma^2, \dots$$

Pomni 3.3: Koeficient asimetrije (s centralnimi momenti)

$g_1 = \frac{m_3}{m_2^{3/2}}$ Mere asimetrije dobimo tako, da opazujemo razlike med srednjimi vrednostmi.

Le-te so tem večje čim bolj je porazdelitev asimetrična:

$$KA_{M_0} = \frac{(\bar{x} - M_0)}{s}, \quad KA_{M_e} = \frac{3(\bar{x} - M_e)}{s}.$$

Pomni 3.4: Koeficient sploščenosti (kurtosis) (s centralnimi momenti):

$$K = g_2 = \frac{m_4}{m_2^2} - 3$$

- $K = 3$ (ali 0) normalna porazdelitev zvonaste-oblike (mesokurtic),
- $K < 3$ (ali < 0) bolj kopasta kot normalna porazdelitev, s krajšimi repi (platikurtic),
- $K > 3$ (ali > 0) bolj špičasta kot normalna porazdelitev, z daljšimi repi (leptokurtic).

3.3 Standardizacija

Vsaki vrednosti x_i spremenljivke X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ : $z_i = \frac{x_i - \mu}{\sigma}$.

Pomni 3.5: Posledično za standardizirano spremenljivko velja

$$\begin{aligned} \mu(Z) &= 0, \\ \sigma(Z) &= 1. \end{aligned}$$