Verjetnost in statistika - priprava na teoretični del izpita

Jernej Vivod

July 13, 2018

Ta skripta vsebuje strnjene zapiske snovi predmeta Verjetnost in statistika na prvi stopnji Bolonjskega programa študija Računalništvo in informatika na Fakuleteti 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 pokaće 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:

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

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:

- $A \cap \bar{A} = N$,
- $A \cup \bar{A} = G$,
- $\bar{N} = G, \bar{G} = N$,
- $\bar{\bar{A}} = A$,
- $A \bar{\cup} B = \bar{A} \cap \bar{B}$,
- $A \bar{\cap} B = \bar{A} \cup \bar{B}$ (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 $S = A_1, A_2, ..., A_n$ 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.

Verjetnosti 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

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

1.4 Trije aksiomi Kolmogorova

Pomni 1.5: Aksiomi Kolmogorova

1. ***Prvi aksiom***
   Verjetnost dogodka je nenegativno realno število: $P(E) \in R$,
   $P(E) \ge 0$
   $\forall E \in F$ *Kjer je* 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 nezdružjivih dogodkov) zadosti
   $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

1.5 Pogojna verjetnost

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

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

Iz te formule sledita naslednji zvezi:

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

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 $P(A|B) = P(A|\bar{B})$.

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:

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

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.

1.6 Bernoullijevo zaporedje neodvisnih dogodkov

O zaporedju neodvisnih poskusov $X_1, X_2, ..., X_n, ...$ govorimo tedaj, ko so verjetnosti izidov v enem poskusu neodvisne od tega, kaj se zgodi v drugih poskusih.

Pomni 1.8: Bernoullijevo zaporedje

Zaporedje neodvisnih poskusov se imenuje Bernoullijevo zaporedje, če se more zgoditi v vsakem poskusu iz zaporedja neodvisnih poskusov le dogodek A z verjetnostjo $P(A) = p$ ali dogodek $\bar{A}$ z verjetnostjo $P(\bar{A}) = 1 - P(A) = 1 - p = q$.

V Bernoullijevem zaporedju neodvisnih poskusov nas zanima, kolikšna je verjetnost, da se v n zaporednih poskusih zgodi dogodek A natanko k-krat.

Verjetnost, da se na začetku poskusov k-krat zgodi dogodek A, ki ima verjetnost p, je enaka $p^k \cdot (1 - p)^{n-k}$. To se lahko zgodi na mnogo načinov - toliko, na kolikor načinov si lahko izberemo k poskusov.

Pomni 1.9: Bernoullijev obrazec

Verjetnost, da se v n poskusih k-krat zgodi dogodek A z verjetnostjo p, je enaka:

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Število načinov, da izmed n poiskusov izberemo k uspešnih pomnožimo z verjetnostjo, da se je zgodilo k dogodkov A.

Pomni 1.10: Strilingov obrazec

$$n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$$

Pomni 1.11: De Moivrov točkovni obrazec

Za velike n velja:

$$P_n(k) \approx \frac{1}{\sqrt{\frac{\pi n}{2}}} e^{-\frac{(k-n/2)^2}{n/2}}$$

Pomni 1.12: Laplaceov točkovni obrazec

Laplaceov točkovni obrazec smemo uporabljati, ko je n velik in p blizu $1/2$:

$$P_n(k) \approx \frac{1}{\sqrt{\frac{2\pi npq}{2}}} e^{-\frac{(k-np)^2}{2npq}}$$

1.7 Slučajne spremenljivke

Slučajno spremenljivko predstavimo z naslednjimi njenimi lastnostnmi:

1. kakšne vrednosti mora imeti (zaloga vrednosti, oznaka Z) in
2. Kolikšna je verjetnost vsake izmed možnih vrednosti ali intervala vrednosti. Predpis, ki določa te verjetnosti, imenujemo **porazdelitveni zakon**.

Pomni 1.13: Kdaj je poznan porazdelitveni zakon slučajne spremenljivke X?

Porazdelitveni zakon slučajne spremenljivke X je poznan, če je mogoče za vsako realno število x določiti verjetnost $F(x) = P(X \leq x)$.

Pomni 1.14: Pričakovana vrednost (končne) diskretne spremenljivke X

Pričakovana vrednost (končne) diskretne spremenljivke X, oznaka $E(X)$ oz. μ_X, je posplošitev povprečne vrednosti.

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_m p_m + \cdots .$$

1.8 Enakomerna diskretna porazdelitev

Diskretna slučajna spremenljivka se porazdeljuje enakomerno, če so vse njene vrednosti enako verjetne. Slučajna spremenljivka, ki je porazdeljena enakomerno, mora imeti vedno končno zalogo vrednosti, tj $n < \infty$, pri čemer je $n := |Z|$. V primeru enakomerne diskretne porazdelitve se pričakovana vrednost slučajne spremenljivke ujema s povprečjem njene zaloge vrednosti.

INSERT IMAGES HERE

1.9 Binomska porazdelitev

Pomni 1.15: Binomska porazdelitev

Binomska porazdelitev ima zalogo vrednosti {0, 1, 2,...,n} in verjetnosti, ki jih izračunamo po Bernoullijevem obrazcu

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, ..., n$$

Binomska porazdelitev je natanko določena z dvema podatkoma - parametroma: $n \in N$ in $p \in [0, 1]$.

Pomni 1.16: Kako zapišemo, da se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p?

$X \sim B(n, p)$

Pomni 1.17: Bernoullijeva oz. indikatorska slučajna spremenljivka

Za $n = 1$ pravimo slučajni spremenljivki X, ki je porazdeljena binomsko, Bernoullijeva (oz. indikatorska) slučajna spremenljivka.
Velja $E(X) = np$.

1.10 Poissonova porazdelitev - $P(\lambda)$

Dobljena limita predstavlja verjetnost pri porazdelitvi, ki jo želimo vpeljati v tem razdelku. Poissonova porazdelitev izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem intervalu, če vemo, da se ti dogodki pojavijo s poznano povprečno frekvenco in neodvisno od časa, ko se je zgodil zadnji dogodek. Uporabimo jo lahko tudi za število dogodkov v drugih intervalih, npr. razdalja, prostornina,... Ima zalogo vrednsoti $0, 1, 2, ...$, njena verjetnostna funkcija pa je

Pomni 1.18: Verjetnostna funkcija za Poissonovo porazdelitev

$$p_k = P(številodogodkov = k) = \lambda^k \frac{e^{-\lambda}}{k!}.$$

Pomni 1.19: Poissonov obrazec

Za velike n in majhne vrednosti, tj. p blizu 0 velja

$$P_n(k) \approx \frac{(np)^k e^{-np}}{k!}.$$

Pomni 1.20: Še ena oblika zapisa Poissonovega obrazca

$$B(n, p) \approx P(np)$$

1.11 Negatina binomska oziroma Pascalova porazdelitev - P(m, p)

Negativna binomska porazdelitev, ki je znana tudi pod imenom Pascalova porazdelitev, ima zalogo vrednosti $m, m + 1, m + 2, ...$, njena verjetnostna funkcija pa je...

Pomni 1.21: Verjetnostna funkcija Pascalove porazdelitve

$$p_k = \binom{k-1}{m-1}(1-p)^{k-m}p^m$$

Parameter p je verjetnost dogodka A. Verjetnostna funkcija pa opisuje verjetnost, da se bo dogodek A zgodil m-krat.

2 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 oeniti 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. Ustreznem mestu v ranžirni vrsti pravimo rang.

2.1 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: $\mu = \frac{x_1+\cdots+x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$

Povprečje vzorca: $\bar{x} = \frac{x_1+\cdots+x_n}{n} = \frac{\sum_{i=1} nx_i}{n}$

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.

2.2 Normalna porazdelitev

Pomni 2.1: Empirična pravila

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

1. *približno 63.3% 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 2.2: Centralni momenti

Za $l \in N$ je $l-ti$ centralni moment enak
$m_l = \frac{(x_1-\bar{x})^l+\cdots+(x_n-\bar{x})^l}{n}$. $m_1 = 0, m_2 = \sigma^2, \ldots$

Pomni 2.3: Koeficient asiemtrije (s centralnimi momenti)

$g_1 = \frac{m_3}{m_2^{\frac{3}{2}}}$ *Mere asiemtrije 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 2.4: 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 (playkurtic),*
- *$K > 3(ali > 0)$ bolj špičasta kot normalna porazdelitev, z daljšimi repi (leptokurtic).*

2.3 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 2.5: Posledično za standardizirano spremenljivko velja

$\mu(Z) = 0,$
$\sigma(Z) = 1.$

2.4 Vzorčenje

Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njen izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor $(X_1, \cdots, X_n)$, vrednostnim meritev $(x_1, \cdots, x_n)$ pa rečemo vzorec. Število n je velikost vzorca.

Predpostavimo, da velja naslednje.

1. vsi členi X_i vektorja imajo isto porazdelitev, kot spremenljivka X,
2. členi X_i so med seboj neodvisni.

Takemu vzorcu rečemo enostavni slučajni vzorec. Večina statistične teorije temelji na predpostavki, da imamo opravka z enostavnim slučajnim vzorcem.

Pomni 2.6: Osnovni izrek statistike

todo

Pomni 2.7: Sredinske mere pri vzorcih

- *Vzorčni modus - najpogostejša vrednost (smiselna tudi za imenske).*
- *VZorčna mediana - srednja vrednost, glede na urejenost, (smiselna tudi za urejenostne).*
- *Vzorčno povprečje - povprečna vrednost (smiselna za vsaj razmične)*
- *Vzorčna geometrijska sredina - (smiselna za vsaj razmerostne):* $G(x) = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Pomni 2.8: Mere razpršenosti pri vzorcih

- *Vzorčni razmah* $= max_i x_i - min_i x_i$
- *Vzorčna disperzija* $s_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- *Popravljena vzorčna disperzija* $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- *Vzorčna odklona* s_o *in* s

2.5 Porazdelitev vzorčnih povprečij

Pomni 2.9: Število vseh možnih vzorcev

- *V primeru slučajnih vzorcev s ponavljanjem jih je N^n, kjer je N velikost populacije in n velikost vzorca.*
- *Število slučajnih vzorcev brez ponavljanja je $\binom{N}{n}$.*
- *Če upoštevamo vrstni red izbranih enot v vzorcu jih je $\binom{N+n-1}{n}$.*

Pomni 2.10: Ocenjevanje lastnosti, ki jih na populaciji spremlja s.s X s končnima parametroma $E(X) = \mu$ in $D(X) = \sigma^2$

z te populacije izberemo naključni slučajni vzorec $(x_1, ..., x_n$. Če so $X_1, ..., X_n$ slučajne spremenljivke, ki spremljajo elemente tega vzorca po koordinatah, potem so neodvisne in zanje velja:

$$E(X_i) \approx \mu$$

in

$$D(X_i) \approx \sigma^2$$

$(1 \leq i \leq n)$.

Nadalje lahko pričakovano vrednost in standardni odklon vzorčnega povprečja $\bar{X}$ ocenimo z $\mu_{\bar{X}} \approx \mu$ in $\sigma_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}}$.

Iz druge zveze vidimo, da pada standardni odklon $\sigma_{\bar{X}}$ proti 0 z naraščanjem velikosti vzorca, tj $\bar{X} \Rightarrow \mu$ (enako nam zagotavlja tudi krepki zakon velikih števil)