Verjetnost je študij možnosti ali verjetja, da se bo nek dogodek zgodil.

Statistika preučuje podatke, jih zbira, klasificira, povzema, organizira, analizira in interpretira.

Glavni veji statistike:

Opisna statistika se ukvarja z organiziranjem, povzemanjem in opisovanjem zbirk podatkov (reduciranje podatkov na povzetke)

Analitična statistika jemlje vzorce podatkov in na osnovi njih naredi zaključke (inferenčnost) o populaciji (ekstrapolacija).

Populacija = vsi objekti ki jih opazujemo (npr. vsi registrirani glasovalci)

Vzorec je podmnozica populacije (100 registriranih glasovalcev)

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, diagram

Opis je samodejno ustvarjen

Tipi podatkov:

* Kvantitativni (numerični) predstavljajo kvantiteto ali količino nečesa (npr. teža), interval (enaki intervali predstavljajo enake količine, poljubna ničla), razmerje (operacije seštevanja, odštevanja, množenja in deljenja, smiselna točka nič)
* Kvalitativni (kategorije) ni kvantitativnih interpretacij (npr. spol), nominalni (kategorije brez odgovarjajočega vrstnega reda/urejenosti), ordinalni/številski (kategorije z urejenostjo)

Frekvenca, npr. število zaposlenih (točna številka)

Relativna frekvenca, npr. delež zaposlenih (%)

Grafična predstavitev kvantitativnih podatkov: stolpčni graf, strukturni krog, histogrami, škatla z brki (box plot), steblo-list predstavitev (stem-and-leaf), zaporedje (dot plot) in runs plot(X,Y plot)

Urejeno zaporedje je zapis podatkov v vrsto po njihovi numerični velikosti (ustreznemu mestu pravimo rang)

Histogram -> izračunaj razpon podatkov -> razdeli razpon na 5 do 20 razredov enake širine -> za vsak razred preštej število vzorcev, ki spadajo v ta razred (frekvenca razreda) -> izračunaj vse relativne frekvence razredov

Modus (oznaka M0) množice podatkov je tista vrednost, ki se pojavi z največjo frekvenco

Mediana (oznaka Me) -> podatke uredimo v naraščajočem vrstnem redu -> če št. Podatkov liho, je mediana na sredini, če pa sodo, je mediana povprečje sredinskih dveh vrednosti

Povprečje populacije:

Povprečje vzorca:

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

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

Določanje 100p-tega centila: izračunaj vrednost p(n + 1) in zaokroži na najbližje število. Naj bo to število enako i. Izmerjena vrednost z i-tim rangom je 100p-ti centil.

25. centil je tudi 1. kvartil, 50. centil je 2. kvartil ALI mediana, 75. centil je tudi 3. kvartil.

Varianca je kvadrat pričakovanega standardnega odklona populacije ali vsota kvadratov standardnih odklonov deljena s stopnjo prostosti vzorca.

Slika, ki vsebuje besede besedilo, pisava, diagram, posnetek zaslona

Opis je samodejno ustvarjen

Standardni odklon (deviacija) je pozitivni kvadratni koren variance.

Koeficient variacije je standardni odklon deljen s povprečjem.

Sredine:

* Aritmetična:
* Geometrična:
* Harmonična:
* Kvadratna:
* Potenčna:

Normalna porazdelitev:

* Veliko podatkovnih množic ima porazdelitev približno zvonaste oblike (unimodalna oblika – en sam vrh)
* Če ima podatkovna množica porazdelitev približno zvonaste oblike, potem veljajo naslednja pravila:
  + 68,3% vseh meritev leži na razdalji 1 standardnega odklona od povprečja
  + 95,4% meritev leži do 2 standardnega odklona od njihovega povprečja
  + Skoraj vse meritve, 99,7%, ležijo na razdalji 3 standardnih odklonov od povprečja
* Če je sprem. Približno normalno porazdeljena, potem jo povprečje in standardnih odklon zelo dobro opisujeta
* V primeru unimodalne porazdelitve sprem., ki pa je bolj asimetrična in bolj ali manj sploščena (koničasta), pa je potrebno izračunati še stopnjo asimetrije in sploščenosti.

l-ti centralni moment je: , kjer je m1 = 0 in m2 = σ2

Koeficient asimetrije (s centralnimi momenti):

Razlike med srednjimi vrednostnimi so tem večje, čiim bolj je porazdelitev asimetrična:

**in pa**

Koeficient sploščenosti: , kjer:

* K = 3 (ali 0) pomeni normalna porazdelitev zvonaste oblike (mesokurtic)
* K < 3 (ali negativna) bolj kopasta kot normalna porazdelitev, s krajšimi repi
* K > 3 (ali pozitivna) bolj špičasta kot normalna porazdelitev, z daljšimi repi

Standardizacija

* Vsaki vrednosti xi sprem. X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ.
* Za novo sprem. Z bomo rekli, da je standardizirana, Zi pa je standardizirana vrednost
* Potem je μ(Z) = 0 in σ(Z) = 1

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava

Opis je samodejno ustvarjen

Injektivni in surjektivni funkciji pravimo bijekcija. Množicama med katerima obstaja bijekcija pravimo bijektivni množici. Bijektivni množici imata enako število elementov (npr. končno, števno neskončno)

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, številka

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, posnetek zaslona, risanka, ilustracija

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, številka

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, pisava, vrstica, rokopis

Opis je samodejno ustvarjen

Poskus je realizacija neke množice skupaj nastopajočih dejstev (kompleksa pogojev). Poskus je torej vsako dejanje, ki ga opravimo v natanko določenih pogojih.

Dogodek je pojav, ki v množico skupaj nastopajočih dejstev ne spada in se lahko v posameznem poskusu zgodi ali pa ne.

Dogodke označujemo z velikimi črkami iz začetka abecede (A,B,C,..), poskuse pa iz konca (X, Y,.)

Dogodek je lahko: gotov (ob vsaki ponovitvi poskusa se zgodi, G), nemogoč (nikoli se ne zgodi, N) ali slučajen (včasih se zgodi, včasih pa ne)

Dogodek A je poddogodek ali način dogodka B, kar zapišemo A podmnožica B, če se vsakič, ko se zgodi dogodek A, zagotovo zgodi dogodek B (pri metu kocke je dogodek A, da pade šest pik, način dogodka B, da pade sodo število pik.) Če je dogodek A način dogodka B in sočasno dogodek B način dogodka A, sta dogodka enaka. Vsota dogodkov A in B, označimo z A + B, se zgodi, če se zgodi vsaj eden od dogodkov A in B (Vsota dogodka A, da vržemo sodo število pik, in dogodka B, da vržemo liho število pik, je gotov dogodek.). Produkt dogodkov A in B, označimo z AB, se zgodi, če se zgodita dogodka A in B hkrati (Produkt dogodka A, da vržemo sodo število pik, in dogodka B, da vržemo liho število pik, je nemogoč dogodek.). Dogodku A nasproten dogodek Ā imenujemo negacijo dogodka A. Dogodka A in B sta nezdružljiva, če se ne moreta zgoditi hkrati, njun produkt je torej nemogoč dogodek (N). 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 (Pri metu kocke je šest osnovnih dogodkov: E1, da pade 1 pika, E2, da padeta 2 piki, . . . , E6, da pade 6 pik. Dogodek, da pade sodo število pik je sestavljen dogodek iz treh osnovnih dogodkov (E2, E4 in E6)). Množico dogodkov imenujemo popolni sistem dogodkov, če se v vsaki ponovitvi poskusa zgodi natanko eden izmed dogodkov iz množice S (pomeni da vsak dogodek ni nemogoč, so paroma nezdružljivi (produkt dveh dogodkov je nemogoč dogodek) in njihova vsota (skupna) je gotov dogodek).

Ponovitve poskusa, v katerih se dogodek A zgodi, imenujemo ugodne za dogodek A, število f(A) = k/n pa je relativna frekvenca (pogostost) dogodka A v opravljenih poskusih. *Če poskus X dolgo ponavljamo, se relativna frekvenca slučajnega dogodka ustali in sicer skoraj zmeraj toliko bolj, kolikor več ponovitev poskusa napravimo.* ***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. (STATISTIČNA DEFINICIJA VERJETNOSTI)***

Osnovne lastnosti:

* Ker je relativna frekvenca vedno nenegativna -> P(A) >= 0
* P(G) = 1, P(N) = 0, če A podmnožica B -> P(A) <= P(B)
* Če sta A in B nezdružljiva, velja P(A + B) = P(A) + P(B)

**Klasična definicija verjetnosti:**

Verjetnosti prostor S slučajnega pojava je množica vseh možnih izidov.

Dogodek je katerikoli izid ali množica izidov slučajnega pojava, je torej podmnožica vzorčnega prostora.

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

Če je nek dogodek A sestavljen iz r dogodkov iz tega popolnega sistema dogodkov, potem je njegova verjetnost P(A) = r/s.

Geometrijska verjetnost: verjetnost sestavljenega dogodka kot razmerje dolžin dela, ki ustreza ugodnim izidom, in dela, ki ustreza možnim izidom.

Za dogodka A in B velja tudi: P(A + B) = P(A) + P(B) – P(AB), pri dogodkih A, B, C pa: P(A + B + C) = P(A) + P(B) + P(C) – P(AB) – P(AC) – P(BC) + P(ABC) (pravilo o vključitvi in izključitvi za množice)

Negacija:

Če so dogodki A*i* , i je element I paroma nezdružljivi:

Kompleksu pogojev K pridružimo mogoč dogodek B, tj. P(B) > 0. Realizacija tega kompleksa pogojev K' = KB je poskus X' in verjetnost dogodka A v tem poskusu je PB(A), ki se z verjetnostjo P(A) ujema ali pa ne. **Pravimo, da je poskus X' poskus X s pogojem B in verjetnost PB(A) pogojna verjetnost dogodka A glede na dogodek B, kar zapišemo: PB(A) = P(A/B).**

Relativna frekvenca A v opravljenih ponovitvah poskusa X':

Grafično (oz. tekstovno 😊 ): P(B) -> P(A/B) -> P(AB)

Par formul: P(AB) = P(B) \* P(A/B) in P(AB) = P(A) \* P(B/A), velja: P(A) \* P(B/A) = P(B) \* P(A/B)

Slika, ki vsebuje besede diagram, besedilo, vrstica, posnetek zaslona

Opis je samodejno ustvarjen

Verjetnost vsakega izmed dogodkov je enaka produktu verjetnosti na puščicah od začetka (koren od G) pa do samega dogodka)

Dogodka A in B sta neodvisna, če velja: P(AB) = P(A) \* P(B), zato za neodvisna dogodka A in B velja P(A/B) = P(A), za nezdružljiva dogodka A in B pa velja P(A/B) = 0.

Dogodka A in B sta neodvisna, če je , nadalje velja:

Naj bo Hi razbitje gotovega dogodka G, hkrati pa paroma nezdružljivi. Zanima nas verjetnost dogodka A, če poznamo verjetnost P(Hi) in pogojno verjetnost P(A/Hi). Ker so dogodki AHi paroma nezdružljivi, velja:

V prvem koraku se zgodi natanko eden od dogodkov H, ki ga imenujemo domneva (hipoteza – sestavljajo popoln sistem dogodkov), dogodek A je eden izmed mogočih dogodkov na drugi stopnji.

Včasih nas zanima po uspešnem izhodu tudi druge stopnje, verjetnost tega, da se je na prvi stopnji zgodil dogodek Hi, odgovor dobimo preko Bayes-ovega obrazca:

Bernoullijevo zaporedje neodvisnih poskusov: 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 Ā z verjetnostjo P(Ā) = 1 – P(A) = 1 – p = q.

Zanima nas kolikokrat se v n zaporednih poskusih zgodi dogodek A natanko k-krat. To pomeni, da se nasprotni dogodek izvede (n-k)-krat. Zvezi pravimo Bernoullijev obrazec:

, uporaba rekurzije: , za k = 1,…

Stirlingov obrazec:

Točkovni obrazci:

* De moivrov točkovni obrazec: , ki je poseben primer Laplaceovega točkovnega obrazca, ki ga smemo uporabljati, ko je p blizu ½:

**Slučajne spremenljivke in porazdelitve:**

Imamo poskus, katerega izidi so števila (npr. pri metu kocke so izidi števila pik). Se pravi, da je poskusom prirejena neka količina, ki more imeti različne vrednosti. Torej je spremenljivka. Katero od mogočih vrednosti zavzame v določeni ponovitvi poskusa, je odvisno od slučaja. Zato ji rečemo slučajna spremenljivka. Potrebno je vedeti: kakšne vrednosti more imeti (zaloga vrednosti) in kolikšna je verjetnost vsake izmed možnih vrednosti ali intervala vrednosti. Predpis, ki določa te verjetnosti, imenujemo porazdelitveni zakon.

Slučajne spremenljivke označujemo z velikimi tiskanimi črkami iz konca abecede, vrednosti spremenljivke pa z enakimi malimi črkami. Tako je npr. (X = x) dogodek, da sluč. Sprem. X zavzame vrednost x. Porazdelitveni zakon slučajne sprem. X je poznan, če je mogoče za vsako realno število x določiti verjetnost: . F(x) imenujemo porazdelitvena funkcija.

Najpogosteje uporabljamo naslednji vrsti sluč. Sprem:

* Diskretna sluč. Sprem., pri kateri je zaloga vrednosti neka števna (diskretna) množica
* Zvezna sluč. Sprem., ki lahko zavzame vsako realno število znotraj določenega intervala

Lastnosti porazdelitvene funkcije:

1. Funkcija F je definirana na vsem R in 0 ≤ F(x) ≤ 1, ∀x ∈ R.

2. Funkcija F je ne padajoča: x1 < x2 ⇒ F(x1) ≤ F(x2).

3. F(−∞) := lim x→−∞ F(x) = 0 in F(∞) := lim x→∞ F(x) = 1.

4. Funkcija je v vsaki točki z desne zvezna F(x+) := lim 0≤h→0 F(x + h) = F(x).

5. Funkcija ima lahko v nekaterih točkah skok. Vseh skokov je največ števno mnogo.

6. P(x1 < X ≤ x2) = F(x2) − F(x1).

7. P(x1 ≤ X ≤ x2) = F(x2) − F(x1+).

8. P(X > x) = 1 − F(x).

9. P(X = x) = F(x) − F(x−).

Diskretne slučajne spremenljivke: Zaloga vrednosti je števna množica (x1, x2, …, xn), torej je tudi števno neskončna, kot pri množici naravnih ali celih števil. Dogodki X = xk, k = 1,2,… sestavljajo popoln sistem dogodkov. Posamezna verjetnost dogodka je enaka P(X = xi) = pi. Vsota verjetnosti vseh dogodkov je enaka 1 (p1 + p2 + … + pn = 1).

Verjetnostna tabela prikazuje diskretno slučajno spremenljivko s tabelo tako, da so v prvi vrstici zapisane vse vrednosti xi, pod njimi pa pripisane pripadajoče verjetnosti pi: . Porazdelitvena funkcija je v tem primeru:

Slika, ki vsebuje besede besedilo, diagram, pisava, rokopis

Opis je samodejno ustvarjen

Enakomerna diskretna porazdelitev: končna diskretna slučajna sprem. Se porazdeljuje enakomerno, oznaka U(n), kjer je n velikost zaloge vrednosti, če so vse njene vrednosti enako verjetne.

Slika, ki vsebuje besede diagram, vrstica, pravokotnik, vzporedno

Opis je samodejno ustvarjen

Binomska porazdelitev: ima zalogo vrednosti {0,1,…,n} in verjetnosti, ki jih računamo po Bernoullijevemu obrazcu: . Binomska porazdelitev je natanko določena z dvema podatkoma/parametroma: n in p. Če se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p, zapišemo: X ~ B(n,p), E(X) = np

Pričakovana vrednost E(X) je posplošitev povprečne vrednosti diskretne spremenljivke X, tj:

, od koder izhaja:

Poissonova porazdelitev: definiramo sluč. Sprem. X kot X = število pojavitev nečesa na nek časovni interval, pri tem velja: vsak časovni interval je enak kot vsak drug časovni interval in pa da so si časovni intervali neodvisni (če je v enem časovnem intervalu veliko pojavitev nečesa, še ne pomeni da bo tudi v naslednjem). E(X) = λ. Najprej uporabimo binomsko porazdelitev B(n,p), kjer je n število ponovitev poskusa (v našem primeru je enako številu manjših časovnih enot), verjetnost posameznega dogodka p pa je enaka verjetnosti, da je v dani manjši časovni enoti prišel mimo vsaj en avto. Izpeljali smo Poissonov obrazec, kjer za velike n in majhne verjetnosti tj. p blizu 0 velja: . Poissonova porazdelitev P(λ) 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. Poissonovo porazdelitev lahko uporabimo tudi za število dogodkov v drugih intervalih, npr. razdalja, prostornina,... Ima zalogo vrednosti {0, 1, 2, . . .}, njena verjetnostna funkcija pa je , kjer je λ > 0 dani parameter in predstavlja pričakovano pogostost nekega dogodka. . Vidimo da zaloga vrednosti ni omejena, kar je bistvena razlika v primerjavi z binomsko, kjer število uspehov seveda ne more presegati števila Bernoullijevih poskusov n.

Pascalova porazdelitev: (oz. negativna binomska porazdelitev, negBin(m, p)). Ima zalogo vrednosti {m, m+1, m +2, ….}, verjetnostna funkcija pa je: , za k >= m, kjer je 0 < p < 1 dani parameter/verjetnost dogodka A v posameznem poskusu. Opisuje porazdelitev potrebnega števila poskusov, da se dogodek A zgodi m-krat. Če številu poskusov sledimo s slučajno sprem. X, potem verjetnost P(X = k), da se bo pri k ponovitvah poskusa dogodek A zgodil v zadnjem poskusu ravno m-tič, izračunamo po zgornji formuli za pk.

Za m = 1 dobimo geometrijsko porazdelitev G(p), le-ta opisuje porazdelitev števila poskusov, da se dogodek A v zadnji ponovitvi poskusa zgodi prvič.

Hipergeometrijska porazdelitev H(n, M, N): bolj splošno, naj bo v posodi M = R rdečih in N – M = B belih kroglic, kjer R, B je element naravnih števil0 in zato N >= M. Zanima nas verjetnost, da je med n je element naravnih števil0 izbranimi kroglicami natanko k rdečih, če izbiramo n-krat brez vračanja. . Za zalogo vrednosti Hipergeometrijske porazdeliteve bomo vzeli podmnožico množice naravnih števil, kjer je definirana zgornja verjetnostna funkcija. Veljajo še naslednje omejitve: max(0, n – B) <= k <= min(M, n) in n <= N = R + B.

Slika, ki vsebuje besede besedilo, rokopis, pisava, posnetek zaslona

Opis je samodejno ustvarjen

Določeni integral predstavlja ploščino pod krivuljo. Naj bo funkcija y = f(x) zvezna na [a, b] in nenegativna. Ploščina lika med krivuljo f(x) >= 0, in abscisno osjo na intervalu [a, b] je enaka določenemu integralu: .

Zvezne slučajne spremenljivke so zvezno porazdeljene, če obstaja taka integrabilna funkcija p, imenovana gostota verjetnosti, da za vsak x є R velja: , kjer p(x) >= 0. To verjetnost si lahko predstavimo tudi grafično v koordinatnem sistemu, kjer na abscisno os nanašamo vrednosti slučajne spremenljivke, na ordinatno pa gostoto verjetnosti p(x). Verjetnost je tedaj predstavljena kot ploščina pod krivuljo, ki jo določa p(x). Velja: in ter p(x) = F'(x).

Enakomerna porazdelitev zvezne slučajne spremenljivke: verjetnostna gostota enakomerno porazdeljene zvezne slučajne spremenljivke U[a,b] je: , ter 0 drugod. Grafično si jo predstavljamo kot pravokotnik nad intervalom (a, b) višine .

Normalna porazdelitev: zaloga vrednosti normalno porazdeljene slučajne spremenljivke so vsa realna števila, gostota verjetnosti pa je: . Porazdelitev je natanko določena z parametroma μ in σ, zapišemo: X ~ N(μ, σ)

Funkcija napake , je liha, zvezno odvedljiva, strogo naraščajoča funcija. , in ter Pn(k1, k2) ~ . V sklopu normalne porazdelitve imamo: . Porazdelitev N(0, 1) je standardizirana normalna porazdelitev. Spremenljivko X : N(μ, σ) pretvorimo z: v standardizirano spremenljivko Z: N(0, 1).

Laplace: iz Laplace-ovega točkovnega obrazca izhaja, da za p blizu ½ in velike n velja: .

Laplace-ov intervalski obrazec:

Bernoulijev zakon velikih števil: naj bo k frekvenca dogodka A v n neodvisnih ponovitvah danega poskusa, v katerem ima dogodek A verjetnost p. Tedaj velja: (primer: kolikšna je verjetnost, da se pri metu kovanca relativna frekvenca grba v 3600 metih ne razlikuje od 0,5 za več kot 0,01?) (n = 3600, p = ½, e = 0,01).

Porazdelitev Poissonovega toka, eksponentna: čas med dvema zaporednima Poissonovega dogodkoma. Gostota eksponentne porazdelitve je enaka: , x >= 0; porazdelitvena funkcija pa:

Porazdelitev Gama: naj bosta b, c > 0. Tedaj ima porazdelitev Gama gostoto: in p(x) = 0 za x <= 0. Funkcijo Gama lahko definiramo z določenim integralom za R[z] > 0: . Torej je .

Porazdelitev hi-kvadrat: je poseben primer porazdelitve gama: , (n je element naravnih števil je število prostostnih stopenj) in ima gostoto: , kjer je x > 0 in 0 sicer. Hi-kvadrat test za ugotavljanje kategoričnih spremenljivk uporabimo v dveh podobnih (a različnih primerih): kako dobro se opazovana/izmerjena porazdelitev prilega pričakovani porazdelitvi (kvaliteta prilagoditve) ter ocenjevanje ali sta obe naključni spremenljivki neodvisni. Test hi-kvadrat napravi dvoje: dejanskim in teoretičnim frekvencam priredi število, s katerim merimo odstopanje frekvenc. Čim večje je dobljeno število, tem večje je odstopanje. Za odstopanje dopuščamo dve razlagi: lahko da gre za slučajno odstopanje ali pa gre še za sistematično odstopanje (torej teoretične frekvence ne ustrezajo dejanski porazdelitvi).

, kjer so E-ji teoretične absolutne frekvence in O-ji dejanske absolutne frekvence. To pomeni, da se je pri N ponovitvah poskusa izid i dogodil Oi-krat, medtem ko smo pričakovali, da se zgodi Ei-krat. (Denimo, da ima poskus 6 izidov in je vrednost hi-kvadrat enaka 12.7. Število prostostnih stopenj je 5. Pogledamo v vrstico s petimi prostostnimi stopnjami in vidimo, da leži 12.7 med 11.1 in 15.1. Verjetnost, da so odstopanja med dejanskimi in teoretičnimi frekvencami zgolj slučajna, je manj kot 5% in več kot 1%.)

Cauchyeva porazdelitev z gostoto , ima porazdelitveno funkcijo:

**Slučajni vektorji in neodvisnost slučajnih spremenljivk:**

Verjetnostna funkcija P(X = x, Y = y) = p(x, y) je definirana z 2D tabelo. Pri tem je: , enako velja za Y. Spremenljivki X in Y sta neodvisni, če za vsako celico v 2D tabeli velja: P(X = x, Y = y) = P(X = x) \* P(Y = y). Vsaka celica v tabeli ima verjetnost: 0 <= p <= 1.

P(A + B) = P(A) + P(B), za AB = N; P(AB) = P(A) \* P(B) NEODVISNOST; E(X + Y) = E(X) + E(Y) VEDNO;

E(XY) = E(X) \* E(Y) NEKORELIRANOST.

Slučajni vektor je n-terica sluč. Sprem. X = (X1, …, Xn). Opišemo ga s porazdelitveno funkcijo:

F(x1, …, xn) = P(X1 <= x1, …, Xn <= xn), pri čemer slednja oznaka pomeni: , za katero velja: 0 <= F(x1, …, xn) <= 1. Funkcija F je za vsako spremenljivko napadajoča in z desne zvezna. F(-infinity, …, -infinity) = 0 in F(infinity,…infinity) = 1.

Funkciji Fi(xi) = F(infinity, …, infinity, xi, infinity, …, infinity) pravimo robna porazdelitvena funkcija sprem. Xi

Diskretne večrazsežne porazdelitve – polinomska porazdelitev: je določena s predpisom: P(X1 = k1, …, Xr = kr) = . Koeficient šteje permutacije s ponavljanjem. Za r = 2 dobimo binomsko porazdelitev, tj. B(n, p) = P(n; p, q). Koeficient r je število razredov. (primer: kup 52 igralnih kart, vlečemo eno in vrnemo nazaj, ponovimo 5x, verjetnost da dobimo 2x srce in 1x po pik, križ in karo? r = 4, n = 5, p1 = p2 = p3 = p4 = ¼, zanima nas P(X1 = 1, X2 = 2, X3 = 1, X4 = 1).

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, rokopis, kaligrafija

Opis je samodejno ustvarjenDvojni integral predstavlja prostornino pod ploskvijo.

Slika, ki vsebuje besede besedilo, rokopis, pisava, kaligrafija

Opis je samodejno ustvarjenZvezne večrazsežne porazdelitve: slučajni vektor X = (X1, X2, …, Xn) je zvezno porazdeljen, če obstaja integrabilna funkcija (gostota verjetnosti) p(x1, x2,…, xn) >= 0 z lastnostjo:

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, vrstica

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, rokopis, posnetek zaslona

Opis je samodejno ustvarjen

Neodvisnost slučajnih spremenljivk: Slučajne spremenljivke X1, X2, . . . , Xn so med seboj neodvisne, če za poljubne vrednosti x1, x2, . . . , xn ∈ R velja F(x1, x2, . . . , xn) = F1(x1) · F2(x2) · · · Fn(xn)

Trditev: Če sta X in Y diskretni slučajni spremenljivki in pij verjetnostna funkcija slučajnega vektorja (X, Y ), potem sta X in Y neodvisni natanko takrat, ko je pij = piqj za vsak par i, j.

Trditev: Če sta X in Y zvezno porazdeljeni slučajni spremenljivki z gostotama pX(x) in pY(y) ter je p(x, y) gostota zvezno porazdeljenega slučajnega vektorja (X, Y ), potem:

Slika, ki vsebuje besede besedilo, pisava, potrdilo, bela

Opis je samodejno ustvarjenX in Y sta neodvisni ⇐⇒ p(x, y) = pX(x) · pY (y) ∀x, y ∈ R.

Če je ρ = 0 sta X in Y neodvisni. Zvezno porazdeljeni slučajni spremenljivki X in Y sta neodvisni natanko takrat, ko lahko gostoto verjetnosti slučajnega vektorja (X, Y ) zapišemo v obliki p(x, y) = f(x) · g(y).

**Funkcije slučajnih spremenljivk/vektorjev in pogojne porazdelitve:**

Naj bo X: G -> R slučajna spremenljivka in f: R -> R neka realna funkcija. Tedaj je njen kompozitum Y = f o X določen s predpisom Y(e) = f(X(e)), za vsak e je element G, določa novo preslikavo Y: G -> R. V ta namen mora biti za vsak y je element R množica:

(Y ≤ y) = {e ∈ G : Y (e) ≤ y} = {e ∈ G : X(e) ∈ f−1(−∞, y]} dogodek, torej v D. Če je to res, imenujemo Y funkcija slučajne spremenljivke X in jo zapišemo kar Y = f(X). Njena porazdelitvena funkcija je: FY(y) = P(Y ≤ y). Kakšna mora biti množica f −1(−∞, y], da je množica v D? Mora biti Borelova množica (ali so intervali, ali števne unije intervalov, ali števni preseki števnih unij intervalov), vsekakor pa ko je f zvezna funkcija. FY = FX o f-1 -> FY(y) = P(Y ≤ y) = P(f(X) ≤ y) = P(X ≤ f-1(y)) = FX(f-1(y))

Trditev: Če sta X in Y neodvisni slučajni spremenljivki ter f in g zvezni funkciji na R, sta tudi U = f(X) in V = g(Y ) neodvisni slučajni spremenljivki.

Funkcije slučajnih vektorjev: Imejmo slučajni vektor X = (X1, X2, . . . , Xn) : G → Rn in zvezno vektorsko preslikavo f = (f1, f2, . . . , fm) : Rn → Rm. Tedaj so Yj = fj(X1, X2, . . . , Xn), j = 1, . . . , m slučajne spremenljivke – komponente slučajnega vektorja Y = (Y1, Y2, . . . , Ym). Pravimo tudi, da je Y funkcija slučajnega vektorja X, tj. Y = f(X). Porazdelitve komponent dobimo na običajen način:

FYj(y) = P(Yj ≤ y) = P(fj(X) ≤ y) = P(X ∈ f j −1(−∞, y]) in če je X zvezno porazdeljen z gostoto p(x1,…,xn), potem je:

Slika, ki vsebuje besede pisava, besedilo, rokopis, bela

Opis je samodejno ustvarjen

Če sta spemenljivki X in Y neodvisni dobimo naprej zvezo:

Gostota pZ = pX \* pY je konvolucija funkcij pX in pY

Če so X1, X2, …, Xn neodvisne standardizirano normalne slučajne spremenljivke, je slučajna spremenljivka Y = X12 + X22 + … + Xn2 porazdeljena po X2(n) (hi-kvadrat).

Če sta X : χ2(n) in Y : χ2(m) neodvisni slučajni spremenljivki, je tudi njuna vsota Z = X + Y porazdeljena po tej porazdelitvi Z : χ2(n + m).

Transformacije: Jacobijeva determinanta: za gostoto q(u, v) vektorja (U, V) dobimo od tu: q(u, v) = p(x(u, v), y(u, v)) \* |J(u, v)|.

Slika, ki vsebuje besede besedilo, potrdilo, pisava, bela

Opis je samodejno ustvarjen

Pogojne porazdelitve: naj bo B nek mogoč dogodek, tj. P(B) > 0. Potem lahko vpeljemo pogojno porazdelitveno funkcijo: F(x | B) = P(X ≤ x | B) = P(X ≤ x, B) / P(B). V diskretnem primeru je: pik = P(X = xi, Y = yk), B = (Y = yk) in P(B) = P(Y = yk) = qk. Tedaj je pogojna porazdelitvena funkcija:

Primer: zapiši pogojno verjetnostno porazdelitev slučajne sprem. X glede na pogoj Y = 2: verjetnosti v vrstici pri Y = 2 moramo deliti s P(Y = 2).

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, algebra

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, bela, potrdilo

Opis je samodejno ustvarjenZvezne pogojne porazdelitve: postavimo B = (y < Y ≤ y +h) za h > 0 in zahtevajmo P(B) > 0. Če obstaja limita (za h 🡪 0) jo imenujemo pogojna porazdelitvena funkcija slučajne spremenljivke X glede na dogodek (Y = y)

**Momenti in kovarianca:**

Pričakovana vrednost E(X) (matematično upanje) je poslošitev povprečne vrednosti diskretne spremenljivke X, tj. , od koder izhaja .

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, potrdilo

Opis je samodejno ustvarjenDiskretna slučajna spremenljivka X z verjetnostno funkcijo pk ima pričakovano vrednost , če je . Zvezna slučajna spremenljivka X z gostoto p(x) ima pričakovano vrednost: , če je: .

Lastnosti pričakovane vrednosti: naj bo a realna konstanta. Če je P(X = a) = 1, je E(X) = a. Velja, da je: |E(X)| ≤ E(|X|). Splošno: E(f(X)) obstaja in je v diskretnem primeru enaka: , v zveznem pa . E(aX) = a\*E(X). Če imata slučajni spremenljivki X in Y pričakovano vrednost, ga ima tudi njuna vsota X + Y in velja E(X + Y) = E(X) + E(Y). Torej je pričakovana vrednost E linearen funkcional, tj. E(aX + bY) = a\*E(X) + b\*E(Y). Če sta slučajni spremenljivki neodvisni: E(XY) = E(X) \* E(Y), vendar opomba: obstajajo tudi odvisne spremenljivke, za katere velja gornja zveza. Spremenljivki, za kateri velja E(XY ) != E(X) · E(Y ) imenujemo korelirani.

Slika, ki vsebuje besede pisava, besedilo, vrstica, bela

Opis je samodejno ustvarjenDisperzija ali varianca D(X) je določena z izrazom: D(X) = E(X2) – (E(X))2. Naj bo a realna konstanta. Če je P(X = a) = 1, je D(X) = 0. D(aX) = a2\*D(X). Količino imenujemo standardna deviacija ali standardni odklon.

Slučajno spremenljivko X standardiziramo s transformacijo: , kjer sta μ = E(X) in σ = sqrt(D(X)). Za XS velja E(XS) = 0 in D(XS) = 1

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, številka

Opis je samodejno ustvarjen

**Slika, ki vsebuje besede besedilo, diagram, posnetek zaslona, načrt

Opis je samodejno ustvarjenUrejene porazdelitve** (kako eno porazdelitev zapisat v drugo (transformacija):

Slika, ki vsebuje besede posnetek zaslona, besedilo, diagram, vrstica

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, pisava, potrdilo, bela

Opis je samodejno ustvarjenKovarianca Cov(X, Y) slučajnih spremenljivk X in Y je določena z izrazom:

Cov(X, Y ) = E((X − E(X))(Y − E(Y )))= E(XY) − E(X)E(Y) in velja: Cov(X, Y) = Cov(Y, X) (simetričnost) in Cov(a\*X + b\*Y, Z) = a\*Cov(X, Z) + b\*Cov(Y, Z) (bilinearnost). Spremenljivki X in Y sta nekorelirani natanko takrat, ko je Cov(X, Y) = 0

Če imata spremenljivki X in Y končni disperziji, jo ima tudi njuna vsota X + Y in velja: D(X + Y) = D(X) + D(Y) + 2\*Cov(X, Y). Če pa sta spremenljivki nekorelirani, je enostavno D(X + Y) = D(X) + D(Y).

Korelacijski koeficient slučajnih spremenljivk X in Y je določen z izrazom: . Za (X, Y) ~ N(μX, μY, σX, σY, p) je r(X, Y) = p. Torej sta normalno porazdeljeni slučajni spremenljivki X in Y neodvisni natanko takrat, ko sta nekorelirani. Velja še: -1 ≤ r(X, Y) ≤ 1. r(X, Y) = 0 ⬄ X in Y sta nekorelirani. r(X, Y) = 1 ⬄ z verjetnostjo 1. r(X, Y) = -1 ⬄ z verjetnostjo 1. Torej, če |r(X, Y)| = 1, obstaja med X in Y linearna zveza z verjetnostjo 1.

Slika, ki vsebuje besede pisava, besedilo, bela, oblikovanje

Opis je samodejno ustvarjenPogojna pričakovana vrednost je pričakovana vrednost pogojne porazdelitve: diskretna slučajna spremenljivka X ima pri pogoju Y = yk pogojno verjetnostno funkcijo pi|k = pik/qk, i = 1, 2, . . . in potemtakem pogojno pričakovano vrednost:

Slika, ki vsebuje besede besedilo, pisava, bela, diagram

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, posnetek zaslona, vrstica

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, bela

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, posnetek zaslona, rokopis

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, posnetek zaslona, pisava

Opis je samodejno ustvarjen

Kovariančna matrika:

Pričakovana vrednost slučajnega vektorja X = (X1, X2, . . . , Xn) je vektor E(X) = (E(X1), E(X2), . . . , E(Xn)). Primer: Za (X, Y) ~ N(μX, μY, σX, σY, p) je E(X, Y) = (μX, μY).

K = [Cov(Xi, Xj)] je kovariančna matrika vektorja X.

Kovariančna matrika K = [Kij] je simetrična: Kij = Kji

Diagonalne vrednosti so disperzije spremenljivk: Kii = D(Xi)

Slika, ki vsebuje besede besedilo, pisava, rokopis, bela

Opis je samodejno ustvarjenČe je kaka lastna vrednost enaka 0, je vsa verjetnost skoncentrirana na neki hiperravnini – porazdelitev je izrojena. To se zgodi natanko takrat, ko je kovariančna matrika K ni obrnljiva, oz. ko je det K = 0.

Primer: Za (X, Y) ~ N(μX, μY, σX, σY, p) je K = . Ker je |p| < 1, je det K = ox2\*oy2 \* (1-p2) > 0.

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, algebra

Opis je samodejno ustvarjen

Višji momenti so posplošitev pojmov pričakovane vrednosti in disperzije. Moment reda k є N glede na točko

a є R imenujemo količino: mk(a) = E((X – a)k). Moment obstaja, če obstaja pričakovana vrednost E(|X – a|k) < inf.

Za a = 0 dobimo začetni moment zk = mk(0), za a = E(X) pa

centralni moment mk = mk \* (E(X)). Primera: E(X) = z1 in D(X)

= m2. Če obstaja moment mn(a), obstajajo vsi mk(a), za k < n

Slika, ki vsebuje besede besedilo, pisava, potrdilo, bela

Opis je samodejno ustvarjenSlika, ki vsebuje besede pisava, besedilo, bela, kaligrafija

Opis je samodejno ustvarjenČe obstaja moment zn, obstaja tudi moment mn(a) za vse a je element R.

Asimetrija spremenljivka X imenujemo količino A(X) = .

Sploščenost spremenljivke X imenujemo količino K(X) = , kjer je

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, diagram

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, posnetek zaslona, algebra

Opis je samodejno ustvarjenZa simetrično glede na z1 = E(X) porazdeljene spremenljivke so vsi lihi centralni momenti enaki 0.

**Karakteristične funkcije in limitni izreki:**

Karakteristična funkcija realne slučajne spremenljivke X je kompleksna funkcija realne spremenljivke t določena z zvezo: . Posebej pomembni lastnosti sta: Če obstaja začetni moment zn, je karakteristična funkcija n-krat odvedljiva v vsaki točki in velja: . Za neodvisni X in Y je

Reprodukcijska lastnost normalne porazdelitve: vsaka linearna kombinacija neodvisnih in normalno porazdeljenih slučajnih spremenljivk je tudi sama normalno porazdeljena. Če so slučajne spremenljivke X1,…,Xn neodvisne in normalno porazdeljene N(μi, σi), potem je njihova vsota tudi normalno porazdeljena: . Da ne bi vsota povprečij rastla z n, nadomestimo vsoto spremenljivk Xi z njihovim povprečjem in dobimo: . Če privzamemo μi = μ in σi = σ, dobimo .

Limitni izreki: zaporedje slučajnih spremenljivk Xn verjetnostno konvergira k slučajni spremenljivki X, če za vsak ξ > 0 velja . Zaporedje slučajnih spremenljivk Xn skoraj gotovo konvergira k slučajni spremenljivki X, če velja: . Če zaporedje slučajnih spremenljivk Xn skoraj gotovo konvergira k slučajni spremenljivki X, potem za vsak ξ > 0 velja: . Od tu izhaja: če konvergira skoraj gotovo Xn -> X, potem konvergira tudi verjetnostno Xn -> X.

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, dokument

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, posnetek zaslona, pisava

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, posnetek zaslona

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, rokopis

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, posnetek zaslona, rokopis

Opis je samodejno ustvarjen

Slika, ki vsebuje besede besedilo, pisava, vrstica, rokopis

Opis je samodejno ustvarjenCentralni limitni izrek: »vsaka vsota ali povprečje, če je število členov dovolj veliko, je približno normalno porazdeljena«. Osnovni centralni limitni izrek (CLI): Če so slučajne spremenljivke Xi neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo, potem zanje velja centralni limitni zakon (v praksi uporabimo n > 30).

Iz konvergence karakterističnih funkcij ϕYn proti karakteristični funkciji standardizirano normalne porazdelitve lahko sklepamo po obratnem konvergenčnem izreku, da tudi porazdelitvene funkcije za Yn konvergirajo proti porazdelitveni funkciji standardizirano normalne porazdelitve. Torej velja centralni limitni zakon.

Uporaba verjetnosti: Ko vstopi v sobo k-ta oseba, je verjetnost, da je vseh k rojstnih dnevov različnih enaka: .

Enostavnejše kode za odpravljanje napak: V splošnem lahko odpravimo n napak z (2n + 1)-kratnim ponavljanjem in uporabo večinskega pravila. Toda ta metoda je preveč potratna.

Ramsey-eva teorija: Vsaka dovolj velika struktura vsebuje urejeno podstrukturo (popoln nered je nemogoč).

Ramsey-evo število: r(k, l) je najmanjše število za katerega vsak graf na r(k, l) vozliščih vsebuje bodisi k-kliko bodisi l-antikliko.

Ramsey-ev izrek: ∀k, ℓ ∈ N: r(k, ℓ) ≤ r(k, ℓ − 1) + r(k − 1, ℓ). Če sta obe števili na desni strani neenakosti sodi, potem velja stroga neenakost. Zgled uporabe: r(3, 3) ≤ r(3, 2) + r(2, 3) = 3 + 3 = 6.

Erdosev izrek: ∀k ∈ N: r(k, k) ≥ 2k/2. Zgled uporabe: r(3, 3) ≥ 3 in r(4, 4) ≥ 4.

**Statistika:**

Statistika je veda, ki proučuje množične pojave.

Dve veji: opisna statistika (zbiranje in urejanje podatkov o nekem pojavu) in inferenčna statistika (poskuša spoznanja iz zbranih podatkov posplošiti (razširiti, napovedovati,…) in oceneiti kakovost teh posplošitev). Lahko jo razdelimo tudi na uporabno in teoretično (računalniško in matematično) statistiko.

(Statistična) enota: posamezna proučevana stvar ali pojav. (npr. redni študent v UL leta 2009)

Populacija je množica vseh proučevanih enot, pomembna je natančna opredelitev populacije (npr. časovno in prostorsko), primer: vsi redni študentje v UL leta 2009.

Vzorec: podmnožica populacije, na osnovi katere po navadi sklepamo o lastnostih celotne populacije.

Spremenljivka: lastnost enot, označujemo jih npr. z X, Y, X1. Vrednost spremenljivke X na i-ti enoti označimo z xi.

Posamezne spremenljivke in odnose med njimi opisujejo ustrezne porazdelitve. Parameter je značilnost populacije in običajno jih označujemo z malimi grškimi črkami. Statistika je značilnost vzorca in 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**.

Vrste spremenljivk glede na vrsto vrednosti:

* Opisne (ali atributivne) spremenljivke: vrednosti opišemo z imeni razredov (npr. poklic)
* Številske (ali numerične) spremenljivke: vrednosti lahko izrazimo s števili (npr. starost)

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, številka

Opis je samodejno ustvarjenVrste spremenljivk glede na vrsto merske lestvice:

* Imenske (ali nominalne) spremenljivke: vrednosti lahko le razlikujemo med seboj: dve vrednosti sta enaki ali različni (npr. spol)
* Urejenostne (ali ordinalne) spremenljivke: vrednosti lahko uredimo od najmanjše do največje (npr. uspeh)
* Razmične (ali intervalne) spremenljivke: lahko primerjamo razlike med vrednostnima dvojic enot (npr. temperatura)
* Razmernostne spremenljivke: lahko primerjamo razmerja med vrednostnima dvojic enot (npr. starost)
* Absolutne spremenljivke: štetja (npr. število prebivalcev)

Vrste spremenljivk so urejene od tistih z najslabšimi merskimi lastnostmi do tistih z najboljšimi. Urejenostne spremenljivke zadoščajo lastnostim, ki jih imajo imenske spremenljivke; in podobno razmernostne spremenljivke zadoščajo lastnostim, ki jih imajo razmične, urejenostne in imenske spremenljivke: absolutna ⊂ razmernostna ⊂ razmična ⊂ urejenostna ⊂ imenska.

Posamezne statistične metode predpostavljajo določeno vrsto spremenljivk. Največ učinkovitih statističnih metod je razvitih za številske spremenljivke.

V teoriji merjenja pravimo, da je nek stavek smiseln, če ohranja resničnost/lažnost pri zamenjavi meritev z enakovrednimi (glede na dovoljene transformacije) meritvami.

Frekvenčna porazdelitev:

Število vseh možnih vrednosti proučevane spremenljivke je lahko preveliko za pregledno prikazovanje podatkov. Zato sorodne vrednosti razvrstimo v skupine. Posamezni skupini priredimo ustrezno reprezentativno vrednost, ki je nova vrednost spremenljivke. Skupine vrednosti morajo biti določene enolično: vsaka enota s svojo vrednostjo je lahko uvrščena v natanko eno skupino vrednosti. Frekvenčna porazdelitev spremenljivke je tabela, ki jo določajo vrednosti ali skupine vrednosti in njihove frekvence. Če je spremenljivka vsaj urejenostna, vrednosti (ali skupine vrednosti) uredimo od najmanjše do največje. Skupine vrednosti številskih spremenljivk imenujemo razredi.

xmin in xmax – najmanjša in največja vrednost spremenljivke X. xi,min in xi,max sta spodnji in zgornji meji i-tega razreda. Meje razredov so določene tako, da velja xi,max = xi+1,min. Širina i-tega razreda je di = xi,max – xi,min. Če je le mogoče, vrednosti razvrstimo v razrede enake širine. Sredina i-tega razreda je xi = (xi,min + xi,max) / 2 in je značilna vrednost/predstavnik tega razreda. Kumulativna (ali nakopičena frekvenca) je frekvenca do spodnje meje določenega razreda. Velja Fi+1 = Fi + fi, kjer je Fi kumulativa in fi frekvenca v i-tem razredu.

Slikovni prikazi:

* Stolpčni prikaz: na eni osi prikažemo (urejene) razrede. Nad vsakim naredimo stolpec/črto višine sorazmerne frekvenci razreda
* Krožni prikaz: vsakemu razredu priredimo krožni izsek s kotom stopinj.
* Histogram: drug poleg drugega rišemo stolpce – pravokotnike, katerih ploščina je sorazmerna frekvenci v razredu. Če so razredi enako široki, je višina sorazmerna tudi frekvenci.
* Poligon: v koordinatnem sistemu zaznamujemo točke (xi, fi), kjer je xi sredina i-tega razreda in fi njegova frekvenca. K tem točkam dodamo še točki (x0, 0) in (xk+1, 0), če je v frekvenčni porazdelitvi k razredov. Točke zvežemo z daljicami.
* Ogiva: grafična predstavitev kumulative frekvenčne porazdelitve s poligonom, kjer v koordinatnem sistem nanašamo točke (xi,min, Fi).

Škatle (box-and-whiskers plot, grafikon kvantilov) boxplot: škatla prikazuje notranja kvartila razdeljena z mediansko črto. Daljici – brka vodita do robnih podatkov, ki sta največ za 1.5 dolžine škatle oddaljena do nje. Ostali podatki so prikazani posamično.

Q-Q prikaz (qqnorm) je namenjen prikazu normalnosti porazdelitve danih n podatkov. Podatke uredimo in prikažemo pare točk sestavljene iz vrednosti k-tega podatka in pričakovane vrednosti k-tega podatke izmed n normalno porazdeljenih podatkov. Če sta obe porazdelitvi normalni, ležijo točke na premici. Premica qqline nariše premico skozi prvi in tretji kvartil. Obstaja tudi splošnejši ukaz qqplot, ki omogoča prikaz povezanosti poljubnega para porazdelitev. S parametroma datax=T zamenjamo vlogo koordinatnih osi.

**Vzorčenje:**

Analitična statistika je veja statistike, ki se ukvarja z uporabo vzorčnih podatkov, da bi z njimi naredili zaključek (inferenco) o populaciji. Zakaj vzorčenje? Cena, čas in destruktivno testiranje.

Glavno vprašanje statistike je: kakšen mora biti vzorec, da lahko iz podatkov zbranih na njem veljavno sklepamo o lastnostih celotne populacije.

Vzorec dobro predstavlja populacijo, če je izbran nepristransko in je dovolj velik. Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor (X1,…, Xn), ki mu rečemo vzorec. Število n je velikost vzorca.

Ker v vzorcu merimo isto spremenljivko in posamezna meritev ne sme vplivati na ostale, lahko predpostavimo: 1. vsi členi Xi vektorja imajo isto porazdelitev, kot spremenljivka X in 2. členi Xi so med seboj neodvisni. Takemu vzorcu rečemo enostavni slučajni vzorec. Večina statistične teorije temelji na predpostavki, da imamo opravka enostavnim slučajnim vzorcem. Če je populacija končna, lahko dobimo enostavni slučajni vzorec, tako da slučajno izbiramo (z vračanjem) enote z enako verjetnostjo.

Z vprašanjem, kako sestaviti dobre vzorce v praksi, se ukvarja posebno področje statistike – teorija vzorčenja. Načini vzorčenja:

* Ocena: priročnost
* Naključnost:
  + Enostavno: pri enostavnem naključnem vzorčenju je vsak član populacije izbran/vključen z enako verjetnostjo
  + Deljeno : razdeljen naključni vzorec dobimo tako, da razdelimo populacijo na disjunktne množice oziroma dele (razrede) in nato izberemo enostavne naključne vzorce za vsak del posebej
  + Grozdno: takšno vzorčenje je enostavno naključno vzorčenje skupin ali klastrov/grozdov elementov

Slika, ki vsebuje besede besedilo, posnetek zaslona, pisava, vrstica

Opis je samodejno ustvarjenOsnovni izrek statistike: spremenljivka X ima na populaciji G porazdelitev F(x) = P(X ≤ x). Toda tudi vsakemu vzorcu ustreza neka porazdelitev. Za realizacijo vzorca (x1, x2, . . . , xn) in x ∈ R postavimo K(x) = | {xi : xi < x, i = 1, . . . , n} | in Vn(x) = K(x)/n. Slučajni spremenljivki Vn(x) pravimo vzorčna porazdelitvena funkcija. Ker ima, tako kot tudi *K(x), n+1* možnih vrednosti k/n, k = 0,…,n, je njena verjetnostna funkcija B(n, F(x)): .

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, rokopis

Opis je samodejno ustvarjen

Vzorčne ocene: najpogostejša parametra, ki bi ju radi ocenili, sta:

* Sredina populacije μ glede na izbrano lastnost – pričakovano vrednost spremenljivke X na populaciji
* Povprečni odklon od sredine σ – standardni odklon spremenljivke X na populaciji

Statistike/ocene za te parametre so izračunane iz podatkov vzorca. Zato jim tudi rečemo vzorčne ocene.

Kot sredinske mere se pogosto uporabljajo:

* 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 vsak razmerostne):

Mere razpršenosti uporabimo za oceno populacijskega odklona:

* Vzorčni razmah:
* Vzorčna disperzija:
* Popravljena vzorčna disperzija: ter ustrezna vzorčna odklona s0 in s

Porazdelitve vzorčnih statistik: Denimo, da je v populaciji N enot in da iz te populacije slučajno izbiramo n enot v enostavni slučajni vzorec ali na kratko slučajni vzorec (vsaka enota ima enako verjetnost, da bo izbrana v vzorec, tj. 1/N). Če hočemo dobiti slučajni vzorec, moramo izbrane enote pred ponovnim izbiranjem vrniti v populacijo (vzorec s ponavljanjem). Če je velikost vzorca v primerjavi s populacijo majhna, se ne pregrešimo preveč, če imamo za slučajni vzorec tudi vzorec, ki nastane s slučajnim izbiranjem brez vračanja. Predstavljajmo si, da smo iz populacije izbrali vse možne vzorce. Dobili smo populacijo vseh možnih vzorcev. Teh je v primeru enostavnih slučajnih vzorcev s ponavljanjem Nn; kjer je N število enot v populaciji in n število enot v vzorcu. Število slučajnih vzorcev brez ponavljanja pa je: in , če upoštevamo vrstni red.

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, dokument

Opis je samodejno ustvarjen

Vzorčna porazdelitev povprečja: naj bo x1, x2, …, xn naključni vzorec, ki je sestavljen iz n meritev populacije s končno pričakovano vrednostjo μ in končnim standardnim odklonom σ. Potem sta povprečje in standardni odklon vzorčne porazdelitve enaka: in . Centralni limitni izrek (pri vzorčenju): če je naključni vzorec velikosti n izbran iz populacije s končno pričakovano vrednostjo μ in končno varianco σ2 ter če je n dovolj velik (n > 30), potem je porazdelitev standardiziranega vzorčnega povprečja , tj. , aproksimirana z N(0,1)

Hitrost centralne tendence pri CLI: dokaz CLI je precej tehničen, kljub temu pa nam ne da občutka kako velik mora biti n, da se porazdelitev slučajne spremenljivke X1 + … + Xn približa normalni porazdelitvi. Hitrost približevanja k normalni porazdelitvi je odvisna od tega kako simetrična je porazdelitev. To lahko potrdimo z eksperimentom: mečemo (ne)pošteno kocko, Xk naj bo vrednost, ki jo kocka pokaže pri k-tem metu.

Slika, ki vsebuje besede skica, besedilo, risanje, diagram

Opis je samodejno ustvarjenSlika, ki vsebuje besede skica, risanje, besedilo

Opis je samodejno ustvarjen

**Cenilke:**

Vzorčna statistika je poljubna simetrična funkcije (njena vrednost je neodvisna od permutacije argumentov) vzorca: Y = g(X1, X2,…, Xn). Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitev vzorca. Najzanimivejši sta značilni vrednosti njene pričakovani vrednosti E(Y) ter standardni odklon σY, ki mu pravimo tudi standardna napaka statistike Y (standard error – zato oznaka SE(Y)).

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, dokument

Opis je samodejno ustvarjenVzorčno povprečje: porazdelitev vzorčnih povprečij je normalna, kjer je pričakovana vrednost vzorčnih povprečij enako pričakovani vrednosti slučajne spremenljivke na populaciji: ter standardni odklon vzorčnih povprečij je: . Če tvorimo vzorce iz končne populacije brez vračanja, je standardni odklon vzorčnih povprečij: . Za dovolj velike vzorce (n > 30) je porazdelitev vzorčnih povprečij približno normalna, tudi če spremenljivka X ni normalno porazdeljena. Če se statistika X porazdeljuje vsaj približno normalno s standardno napako SE(X), potem se z porazdeljuje standardizirano normalno.

Slika, ki vsebuje besede besedilo, pisava, vrstica, bela

Opis je samodejno ustvarjenSlika, ki vsebuje besede besedilo, pisava, rokopis, vrstica

Opis je samodejno ustvarjenVzorčna disperzija:

Če je n zelo velik, je po centralnem limitnem izreku statistika χ2 porazdeljena približno normalno in sicer po zakonu: , cenilka za vzorčno disperzijo S02 približno po: in cenilka za popravljeno vzorčno disperzijo S2 približno po: .

Slika, ki vsebuje besede pisava, diagram, vrstica, bela

Opis je samodejno ustvarjenStudentova porazdelitev: pri normalno porazdeljeni slučajni spremenljivki X je tudi porazdelitev normalna, in sicer . Statistika je potem porazdeljena standardizirano normalno. Pri ocenjevanju parametra μ z vzorčnim povprečjem to lahko uporabimo le, če poznamo σ, sicer ne moremo oceniti standardne napake – ne vemo kako dobra je ocena za μ. Parameter σ lahko ocenimo s vzorčnima S0 in S. Toda S je slučajna spremenljivka in porazdelitev statistike ni več normalna N(0, 1) (razen, če je n zelo velik in s skoraj enak σ). Porazdelitev nove vzorčne statistike je t(n-1) z gostoto :

in n-1 prostostnimi stopnjami. Tej porazdelitvi pravimo Studentova. Za t(1) dobimo Cauchyevo porazdelitev, za n -> infinity pa ima porazdelitev gostoto standardizirane normalne porazdelitve.

Beta funkcija: vpeljemo jo lahko z Gama funkcijo na naslednji način: . Posebne vrednosti: B( ½ , ½ ) = ϖ in za m, n є N še: .

Slika, ki vsebuje besede besedilo, pisava, vrstica, bela

Opis je samodejno ustvarjenFisherjeva ali Snedercorjeva porazdelitev: poskusimo najti še porazdelitev kvocienta , kjer sta U ~ χ2(m) in V ~ χ2(n) in U in V neodvisni. Za x > 0 je gostota ustrezne porazdelitve F(m, n) enaka: in 0 drugje. Porazdelitvi F(m, n) pravimo Fisherjeva ali tudi Snedecorjeva porazdelitev F z (m, n) prostostnimi stopnjami.

Točkovna cenilka je pravilo ali formula, ki nam pove, kako izračunati numerično oceno parametra populacije na osnovi merjenj vzorca. Število, ki je rezultat izračuna, se imenuje točkovna ocena (in mu ne moremo zaupati – v smislu verjetnosti).

Cenilka parametra ζ je vzorčna statistika C = C(X1, …, Xn), katere porazdelitveni zakon je odvisen od parametra ζ, njene vrednosti pa ležijo v prostoru parametrov. Cenilka je simetrična funkcija: njena vrednost je enaka za vse permutacije argumentov. Seveda je odvisna tudi od velikosti vzorca n. Primer: vzorčna mediana in vzorčno povprečje sta cenilki za populacijsko povprečje μ.

Cenilka C parametra ζ je dosledna, če z rastočim n zaporedje Gn verjetnostno konvergira k ζ, to je, za vsak ξ > 0 velja: . Primer: vzorčno povprečje je dosledna cenilka za populacijsko povprečje µ. Tudi vsi vzorčni začetni momenti so dosledne cenilke ustreznih začetnih populacijskih momentov zk = E(Xk), če le-ti obstajajo. Vzorčna mediana je dosledna cenilka za populacijsko mediano. Izrek: Če za n -> infinity velja E(Cn) -> ζ in D(Cn) = 0, je Cn dosledna cenilka parametra ζ.

Nepristranska cenilka z najmanjšo varianco: Cenilka Cn parametra ζ je nepristranska, če je E(Cn) = ζ (za vsak n), in je asimptotično nepristranska, če je . Količino B(Cn) = E(Cn) – ζ imenujemo pristranost (angl. bias) cenilke Cn. Primer: vzorčno povprečje je nepristranska cenilka za populacijsko povprečje μ, vzorčna disperzija S02 je samo asimptotično nepristranska cenilka za σ2, popravljena vzorčna disperzija S2 pa je nepristranska cenilka za σ2.

Izmed nepristranskih cenilk istega parametra ζ je boljša tista, ki ima manjšo disperzijo – v povprečju daje boljše ocene. Če je razred cenilk parametra ζ konveksen (vsebuje tudi njihove konveksne kombinacije), obstaja v bistvu ena sama cenilka z najmanjšo disperzijo.

Srednja kvadratična napaka: Včasih je celo bolje vzeti pristransko cenilko z manjšo disperzijo, kot jo ima druga, sicer nepristranska, cenilka z veliko disperzijo. Mera učinkovitosti cenilk parametra ζ je srednja kvadratična napaka q(C ) = E(C – ζ)2, oz. zapišemo v obliki: q(C) ) = D(C) + B(C)2. Za nepristranske cenilke je B(C) = 0 in zato q(C) = D(C). Če pa je disperzija cenilke skoraj 0, je q(C) ~ B(C)2.

Rao-Cramerjeva ocena: Naj bo p gostotna ali verjetnostna funkcija slučajne spremenljivke X in naj bo odvisna še od parametra a, tako da je p(x; a) njena vrednost v točki x. Združeno gostotno ali verjetnostno funkcijo slučajnega vzorca (X1, . . . , Xn) označimo z L in ji pravimo funkcija verjetja (tudi zanesljivosti, angl. likelihood): L(x1,…,xn; a) = p(x1;a) \* … \* p(xn; a). L(X1,…,Xn) je funkcija vzorca – torej slučajna spremenljivka.

Učinkovitost cenilk: Rao-Cramerjeva ocena da absolutno spodnjo mejo disperzije za vse nepristranske cenilke parametra a (v dovolj gladkih porazdelitvah). Ta meja ni nujno dosežena. Cenilka, ki jo doseže, se imenuje najučinkovitejša cenilka parametra a in je ena sama (z verjetnostjo 1).

Naj bo C0 najučinkovitejša cenilka parametra a in C kaka druga nepristranska cenilka. Tedaj je učinkovitost cenilke C določena s predpisom: . Učinkovitost najučinkovitejše cenilke je e(C0) = 1. Če najučinkovitejša cenilka ne obstaja, vzamemo za vrednost D(C0) desno stran v Rao-Cramerjevi oceni.

Metoda momentov: Parametre populacije (ki jih ne poznamo) določimo tako, da so momenti slučajne spremenljivke X (npr. ) enaki ocenam teh momentov, ki jih izračunamo iz vzorca. Če iščemo en parameter (tj. a) in velja μX = f(a) izračunamo: a = f-1(μX) oz. . Cenilke, ki jih dobimo po metodi momentov so dosledne.

Metoda največjega verjetja: V = (X1,…,Xn) je slučajni vzorec in je odvisen od porazdelitve ter od njenih parametrov a1,…,am. Želimo določiti ocene tako, da je verjetnost, da se je zgodil vzorec V, največja. VELJA NEODVISNOST. Funkcija verjetja L(ai) je verjetnost, da se je zgodil nek vzorec, tj. . Ocena parametra (za parameter a) določimo tako, da bo imela funkcija verjetja L(ai) maksimum. Če najučinkovitejša cenilka obstaja, jo dobimo s to metodo.

Slika, ki vsebuje besede besedilo, pisava, posnetek zaslona, bela

Opis je samodejno ustvarjen

Porazdelitev vzorčnih povprečij:

Slika, ki vsebuje besede besedilo, pisava, rokopis, posnetek zaslona

Opis je samodejno ustvarjen

Porazdelitev vzorčnih deležev: denimo, da želimo na populaciji oceniti delež enot ϖ z določeno lastnostjo. Zato na vsakem vzorcu poiščemo vzorčni delež p. Pokazati se da, da je za dovolj velike slučajne vzorce s ponavljanjem vzorčni deleži porazdeljujejo približno normalno s pričakovano vrednostjo vzorčnih deležev enako deležu na populaciji tj. in standardnim odklonom vzorčnih deležev . Za manjše vzorce se vzorčni deleži porazdeljujejo binomsko. Cenilka populacijskega deleža je nepristranska cenilka.

Porazdelitev razlik vzorčnih povprečij: Denimo, da imamo dve populaciji velikosti N1 in N2 in se spremenljivka X na prvi populaciji porazdeljuje normalno N(µ1, σ), na drugi populaciji pa N(µ2, σ) **(standardna odklona sta na obeh populacijah enaka!).** V vsaki od obeh populacij tvorimo neodvisno slučajne vzorce velikosti n1 in n2. Na vsakem vzorcu (s ponavljanjem) prve populacije izračunamo vzorčno povprečje in podobno na vsakem vzorcu druge populacije . Dokazati se da, da je porazdelitev razlik vzorčnih povprečij normalna, kjer je pričakovana vrednost razlik vzorčnih povprečij enaka: in disperzija razlike vzorčnih povprečij pa: .

Porazdelitev razlik vzorčnih deležev: Podobno kot pri porazdelitvi razlik vzorčnih povprečij naj bosta dani dve populaciji velikosti N1 in N2 z deležema enot z neko lastnostjo π1 in π2. Iz prve populacije tvorimo slučajne vzorce velikosti n1 in na vsakem izračunamo delež enot s to lastnostjo p1. Podobno naredimo tudi na drugi populaciji: tvorimo slučajne vzorce velikosti n2 in na njih določimo deleže p2. Pokazati se da, da se za dovolj velike vzorce razlike vzorčnih deležev porazdeljujejo približno normalno z matematičnim upanjem razlik vzorčnih deležev in disperzijo razlik vzorčnih deležev: .

**Intervali zaupanja:**

Denimo, da s slučajnim vzorcem ocenjujemo parameter γ. Poskušamo najti cenilko G, ki je nepristranska, tj. E(G) = γ in se na vseh možnih vzorcih vsaj približno normalno porazdeljuje s standardno napako SE(G). Nato poskušamo najti interval, v katerem se bo z dano gotovostjo (1 – α) nahajal ocenjevalni parameter: P(a < γ < b) = 1 – α. a je spodnja meja zaupanja, b je zgornja meja zaupanja, α verjetnost tveganja oziroma 1 – α verjetnost gotovosti. Ta interval imenujemo interval zaupanja in ga interpretiramo takole: z verjetnostjo tveganja α se parameter γ nahaja v tem intervalu.

Konstrukcija intervala zaupanja: na osnovi predpostavk o porazdelitvi cenilke G zapišemo standardizirano slučajno spremenljivko , ki se porazdeljuje normalno, N(0,1). Tveganje α porazdelimo simetrično polovico na levo in polovico na desno na konce normalne porazdelitve. Ti konci so označeni z oz. poltraka: .