2. Statistika za poslovno analitiko

Dobrodošli v svetu poslovne statistike, kjer se podatki spremenijo v pomembne vpoglede, ki usmerjajo odločanje in odkrivajo skrite resnice. V tem izčrpnem raziskovanju se podajamo na pot, na kateri razkrivamo bistvene statistične koncepte in tehnike, ki so osnova za temeljito analizo poslovnih podatkov. Od razumevanja zapletenosti porazdelitev do uporabe testiranja hipotez in oblikovanja intervalov zaupanja - vsako poglavje razkriva nov vidik statistične pismenosti.

V središču statistične analize je normalna porazdelitev, krivulja v obliki zvona, ki je značilna za številne pojave v naravi in človeškem vedenju. V tem delu se bomo poglobili v bistvo normalne porazdelitve, razkrili njene lastnosti in pomen pri statističnem sklepanju. Z vizualizacijami in primeri iz resničnega sveta osvetlimo vseprisotnost te temeljne porazdelitve in njeno vlogo temeljnega kamna statistične teorije.

Standardni odklon služi kot kompas v statistični pokrajini, ki nas vodi skozi spremenljivost, značilno za zbirke podatkov. V tem poglavju razčlenjujemo koncept standardnega odklona in razkrivamo njegov pomen pri količinskem opredeljevanju razpršenosti in ocenjevanju razširjenosti podatkovnih točk. Oboroženi z globljim razumevanjem standardnih odklonov boste samozavestno krmarili po podatkih ter natančno prepoznavali vzorce in odstopanja.

Spremenljivke so gradniki statistične analize, vsaka od njih pa ima posebne značilnosti in posledice. To poglavje pojasnjuje dihotomijo med zveznimi in diskretnimi spremenljivkami ter osvetljuje njihovo vlogo pri modeliranju in razlagi podatkov. Z razumevanjem nians vrst spremenljivk boste izkoristili celoten potencial statističnih tehnik, prilagojenih različnim podatkovnim strukturam.

Vzorčna porazdelitev je temelj statističnega sklepanja, saj zapolnjuje vrzel med vzorčnimi opazovanji in populacijskimi parametri. V tem poglavju razkrivamo koncept vzorčne porazdelitve in pojasnjujemo njen pomen pri podajanju verjetnostnih izjav o značilnostih populacije. S konkretnimi primeri boste razvili intuitivno razumevanje vloge porazdelitve vzorčenja pri zanesljivi statistični analizi.

Centralna mejna trditev je ključni koncept v statistiki, ki nam pomaga razumeti negotovost. V tem poglavju je na preprost način razložen Centralni mejni teorem, ki pokaže, kako naredi vzorčne sredine bolj predvidljive in pomaga pri preverjanju hipotez. Z razumevanjem tega koncepta boste lahko iz podatkov potegnili smiselne zaključke.

Razumevanje testiranja hipotez je bistveno za sprejemanje odločitev, ki temeljijo na podatkih. Z njim lahko ugotovimo, ali so opaženi vzorci v podatkih smiselni ali zgolj posledica naključja. S testiranjem hipotez lahko ocenimo predpostavke, primerjamo skupine in ocenimo statistično pomembnost rezultatov, zato je to pomembno orodje v znanstvenih raziskavah, poslovnih analizah in na številnih drugih področjih.

Rezultati Z in tabele Z služijo kot navigacijski pripomočki v morju standardne normalne porazdelitve ter olajšajo standardizirane primerjave in izračune verjetnosti. V tem poglavju so pojasnjene zapletene značilnosti Z-skupin, kar vam omogoča interpretacijo standardiziranih rezultatov in uporabo Z-tabel za statistično analizo. Z znanjem o Z-skupinah boste z gotovostjo in natančnostjo krmarili po prostranstvih normalne porazdelitve.

Kadar so vzorci majhni ali standardni odkloni populacije neznani, so t-skupine in t-tabele nepogrešljiva orodja za statistično analizo. To poglavje razkriva skrivnosti t-skupin in vas vodi skozi njihov izračun in interpretacijo s pomočjo t-tabel. Oboroženi s tem znanjem boste spretno krmarili po niansah t-distribucij in si zagotovili zanesljivo sklepanje v različnih statističnih scenarijih.

Normalna in t-distribucija sta stebra teorije verjetnosti, vsaka od njiju pa ima edinstvene značilnosti in uporabo. V tem poglavju pojasnjujemo razlike med tema porazdelitvama in vam omogočamo, da prepoznate, kdaj uporabiti vsako od njiju pri statistični analizi. S praktičnimi primeri in primerjalnimi analizami boste razvili podrobno razumevanje normalnih in t-razdelitev ter tako obogatili svoje statistično orodje.

Intervali zaupanja omogočajo vpogled v negotovost, ki spremlja populacijske parametre, in nam omogočajo količinsko opredelitev natančnosti ocen. V tem poglavju raziskujemo oblikovanje intervalov zaupanja za sredine in deleže ter razkrivamo metodologijo in razlago teh bistvenih statističnih orodij. Z obvladovanjem intervalov zaupanja boste pregledno in strogo izrazili negotovost, ki je neločljivo povezana z vašimi ugotovitvami.

Čeprav p-vrednosti omogočajo statistično sklepanje, lahko njihova napačna razlaga privede do napačnih sklepov in napačnih odločitev. V tem poglavju so obravnavane morebitne pasti prevelikega zanašanja na p-vrednosti, pri čemer je poudarjen pomen konteksta in velikosti učinka pri statistični analizi. S kritičnim pregledom in praktičnimi spoznanji boste previdno krmarili po zapletenosti p-vrednosti in tako zagotovili celovitost svojih statističnih zaključkov.

Na teh straneh se skrivajo ključi za razvozlavanje skrivnosti statistične analize, ki vam omogočajo, da se samozavestno in natančno orientirate po zapletenih podatkih. Ko se bomo skupaj podali na

to potovanje, naj bo radovednost naš kompas, poizvedovanje pa naša vodilna luč, ki nam bo osvetljevala pot do globljega razumevanja in uporabnih vpogledov.

2.1 Normalna porazdelitev

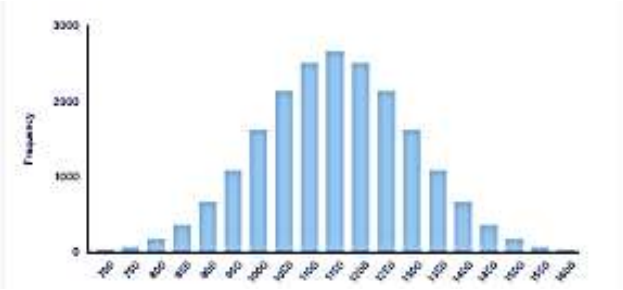
V središču statistične analize je normalna porazdelitev, vseprisotna verjetnostna porazdelitev, ki služi kot merilo za številne statistične tehnike. Spoznali bomo njene značilnosti, simetrično krivuljo v obliki zvona in njen pomen za razumevanje porazdelitve podatkov.

Normalna porazdelitev se uporablja na različnih področjih, vključno s financami, psihologijo, inženiringom in biologijo. Normalna porazdelitev je vsestransko orodje za analizo in razlago podatkov, od modeliranja borznih tečajev do razumevanja porazdelitve človeške višine.

V tem poglavju se bomo poglobili v matematične lastnosti normalne porazdelitve in raziskali, kako izračunati verjetnosti, percentile in z-rezultate. Poleg tega bomo obravnavali praktične tehnike za vizualizacijo in interpretacijo normalne porazdelitve z uporabo histogramov, diagramov gostote in kumulativnih porazdelitvenih funkcij.

Ob koncu tega poglavja boste dobro spoznali normalno porazdelitev in njen pomen v statistični analizi. Oboroženi s tem znanjem boste dobro opremljeni za reševanje naprednejših statističnih konceptov in njihovo uporabo v realnih podatkovnih nizih. Skupaj se podajmo na to potovanje, da razvozlamo skrivnosti normalne porazdelitve.

Normalna porazdelitev, znana tudi kot Gaussova porazdelitev ali krivulja zvona, kaže simetrično porazdelitev podatkov brez poševnosti. Če podatke narišemo na graf, tvorijo krivuljo v obliki zvona, pri čemer je večina vrednosti zbranih okoli središča in se z oddaljevanjem od njega zmanjšuje.



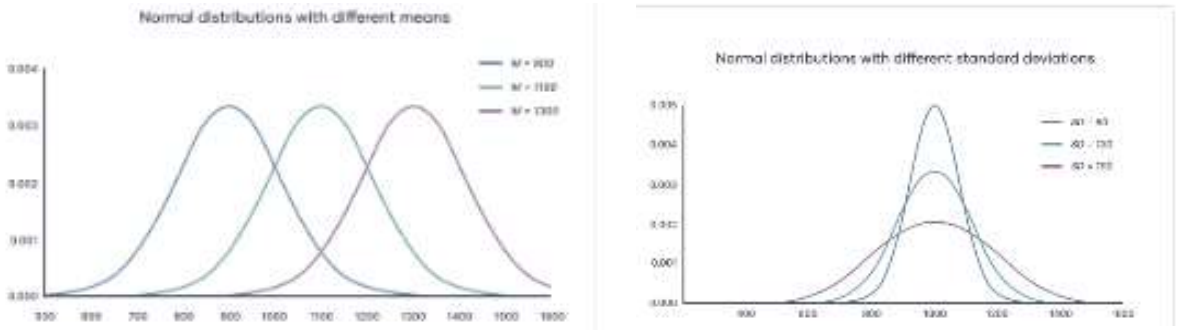
Slika 2.1 Primer Gaussove porazdelitve ali zvonaste krivulje.

Različne spremenljivke v naravoslovju in družboslovju imajo običajno normalno porazdelitev ali njen približek. Primeri so višina, porodna teža, bralne sposobnosti, zadovoljstvo na delovnem mestu in rezultati mature. Zaradi razširjenosti normalno porazdeljenih spremenljivk so številni statistični

testi prilagojeni takšnim populacijam. Spretno razumevanje značilnosti normalnih porazdelitev posameznikom omogoča uporabo inferenčne statistike za primerjavo skupin in izdelavo ocen populacije iz vzorcev.

Normalne porazdelitve imajo ključne značilnosti, ki jih zlahka opazimo v grafih:

- Povprečje, mediana in način so popolnoma enaki.
- Porazdelitev je simetrična glede na srednjo vrednost - polovica vrednosti je pod srednjo vrednostjo, polovica pa nad srednjo vrednostjo.
- Porazdelitev lahko opišemo z dvema vrednostma: srednjo vrednostjo in standardnim odklonom.



Slika 2.2 Normalna porazdelitev z različnimi srednjimi vrednostmi in z različnimi odstopanji.

Srednja vrednost služi kot lokacijski parameter, ki določa središče vrha krivulje. S prilagajanjem sredine se krivulja ustrezno premakne: z njenim povečanjem se krivulja premakne v desno, z njenim zmanjšanjem pa v levo. Medtem standardni odklon deluje kot parameter obsega in vpliva na razpon ali širino krivulje.

Standardni odklon raztegne ali stisne krivuljo. Majhen standardni odklon povzroči ozko krivuljo, velik standardni odklon pa široko krivuljo.

2.2 Empirično pravilo

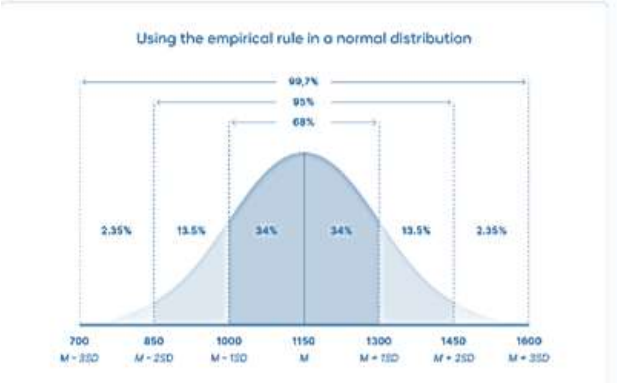
Empirično pravilo, znano tudi kot pravilo 68-95-99,7, omogoča vpogled v porazdelitev vrednosti znotraj normalne porazdelitve:

- Približno 68 % vrednosti je znotraj 1 standardnega odklona od povprečja.
- Približno 95 % vrednosti je znotraj 2 standardnih odklonov od povprečja.
- Približno 99,7 % vrednosti je znotraj 3 standardnih odklonov od povprečja.

Razmislite na primer o scenariju, v katerem so zbrani rezultati izpitov SAT, ki so jih dosegli učenci na novem tečaju za pripravo na izpite, in podatki ustrezajo normalni porazdelitvi s povprečnim rezultatom (M) 1150 in standardnim odklonom (SD) 150.

Z uporabo empiričnega pravila dobimo naslednja spoznanja:

- Približno 68 % rezultatov je v razponu od 1000 do 1300, kar ustreza enemu standardnemu odklonu nad in pod povprečjem.
- Približno 95 % rezultatov je v razponu od 850 do 1450, kar pomeni 2 standardna odklona nad in pod povprečjem.
- Skoraj vsi rezultati, približno 99,7 % so v razponu od 700 do 1600 točk, kar vključuje 3 standardne odklone nad in pod povprečjem.

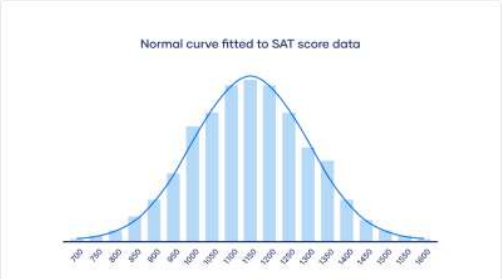


Slika 2.3 Empirično pravilo v normalni porazdelitvi.

Empirično pravilo je hitra metoda za ocenjevanje podatkov, ki omogoča odkrivanje odstopanj ali izjemnih vrednosti, ki odstopajo od pričakovanega vzorca. V primerih, ko podatki iz majhnih vzorcev znatno odstopajo od tega vzorca, so lahko primernejše alternativne porazdelitve, kot je t-razdelitev. Določitev porazdelitve spremenljivke omogoča uporabo ustreznih statističnih testov.

2.3 Formula normalne krivulje

Za konstruiranje normalne krivulje na podlagi dane srednje vrednosti in standardnega odklona lahko uporabimo funkcijo gostote verjetnosti, s čimer natančno predstavimo porazdelitev podatkov.



Slika 2.4 Normalna krivulja, prilagojena podatkom o rezultatih

V funkciji gostote verjetnosti območje pod krivuljo predstavlja verjetnost. Glede na to, da normalna porazdelitev služi kot verjetnostna porazdelitev, je kumulativna površina pod krivuljo vedno enaka 1 ali 100 %. Čeprav se zdi formula za normalno funkcijo gostote verjetnosti zapletena, je za njeno uporabo potrebno le poznavanje populacijske sredine in standardnega odklona. Z nadomestitvijo teh parametrov v formulo lahko določimo gostoto verjetnosti, povezano s katero koli vrednostjo x.

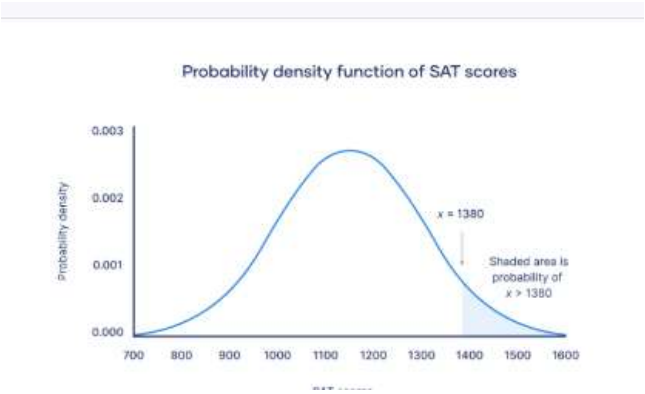
- *f(x)* = verjetnost
- *x* = vrednost spremenljivke
- *μ* = povprečje
- *σ* = standardni *odklon*
- σ^2 = varianta

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Primer:

S pomočjo funkcije gostote verjetnosti želite izvedeti, kolikšna je verjetnost, da bodo rezultati mature SAT v vašem vzorcu presegli 1380 točk.

Na grafu funkcije gostote verjetnosti je verjetnost osenčeno območje pod krivuljo, ki leži desno od točke, kjer je vaš rezultat na testu SAT enak 1380.



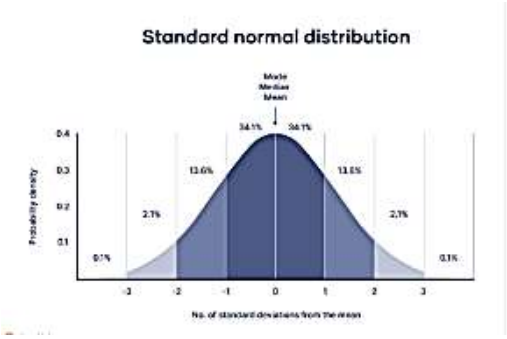
Slika 2.5 Graf funkcije gostote verjetnosti rezultatov na testu

Verjetnostno vrednost tega rezultata lahko določite s standardno normalno porazdelitvijo.

2.4 Standardna normalna porazdelitev

Standardna normalna porazdelitev, znana kot **z-razdelitev**, se razlikuje po tem, da ima srednjo vrednost 0 in standardni odklon 1. Vsako normalno porazdelitev lahko obravnavamo kot transformacijo standardne normalne porazdelitve s prilagoditvami v merilu, položaju ali oboje.

V kontekstu porazdelitve z se posamezna opazovanja, ki so v normalni porazdelitvi običajno označena z x, imenujejo z-vrednost. Te vrednosti z predstavljajo število standardnih odklonov, za katere posamezna vrednost odstopa od povprečja. Zato pretvorba vrednosti iz katere koli normalne porazdelitve v z-lestvice olajša primerjavo in analizo v okviru standardne normalne porazdelitve.



Slika 2.6 Graf standardne normalne porazdelitve.

Za določitev *z-rezultata vrednosti* morate poznati le srednjo vrednost in standardni odklon porazdelitve.

Razlaga formule Z-rezultata

$$z = \frac{x - \mu}{\sigma}$$

- x = individualna vrednost
- μ = povprečje
- σ = standardni odklon

Normalne porazdelitve pretvorimo v standardno normalno porazdelitev iz več razlogov:

- Iskanje verjetnosti, da bodo opazovanja v porazdelitvi padla nad ali pod določeno vrednostjo.
- ugotavljanje verjetnosti, da se povprečje vzorca pomembno razlikuje od znanega povprečja populacije.

- Primerjava rezultatov na različnih porazdelitvah z različnimi srednjimi vrednostmi in standardnimi odkloni.

2.5 Ugotavljanje verjetnosti s pomočjo z-razdelitve

Vsakemu z-rezultatu ustreza verjetnost, pogosto imenovana p-vrednost, ki označuje verjetnost opazovanja vrednosti pod določenim z-rezultatom. S pretvorbo posamezne vrednosti v z-rezultatu lahko določimo verjetnost vseh vrednosti do te točke v okviru normalne porazdelitve.

Na primer, upoštevajte scenarij, v katerem želite ugotoviti, kolikšna je verjetnost, da bodo rezultati mature SAT v vašem vzorcu presegli 1380 točk. Na začetku izračunate z-vrednost s pomočjo povprečja in standardnega odklona porazdelitve. Pri srednji vrednosti 1150 in standardnem odklonu 150 z-vrednost pokaže število standardnih odklonov, za katere 1380 odstopa od povprečja.

Izračun formule

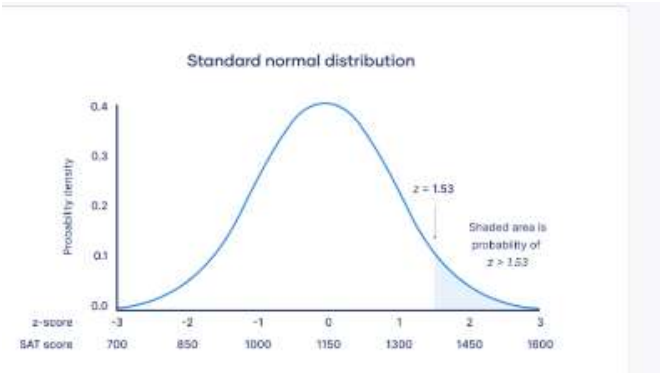
$$z = \frac{x - \mu}{\sigma} = \frac{1380 - 1150}{150} = 1.53$$

Za *z-vrednost* 1,53 je *p-vrednost* 0,937. To je verjetnost, da je rezultat SAT 1380 ali manj (93,7 %), in je površina pod krivuljo levo od osenčenega območja.

Če želite ugotoviti osenčeno površino, od 1 odštejete 0,937, kar je skupna površina pod krivuljo.

Verjetnost x > 1380 = 1 - 0,937 = **0,063**

To pomeni, da je verjetno le 6,3 % rezultatov SAT v vašem vzorcu več kot 1380.



Slika 2.7 Standardna normalna porazdelitev z označenim rezultatom SAT.

2.6 Porazdelitev vzorčenja

Vzorčne porazdelitve so temelj statističnega sklepanja, saj nam omogočajo, da na podlagi vzorčnih podatkov sklepamo o populacijah. Poglobili se bomo v zapletenost vzorčnih porazdelitev, razumeli, kako odražajo variabilnost vzorčnih statistik in njihovo ključno vlogo pri preverjanju hipotez.

Porazdelitev vzorca se nanaša na porazdelitev statistike, kot je vzorčna sredina ali vzorčni delež, pridobljene iz več vzorcev enake velikosti, ki so bili odvzeti iz populacije. Omogoča vpogled v obnašanje vzorčnih statistik in njihovo spremenljivost med različnimi vzorci.

2.7 Centralni mejni teorem in porazdelitev vzorca

Centralni mejni stavek (CLT) je temeljni koncept v statistiki, ki pojasnjuje obnašanje vzorčnih porazdelitev. Trdi, da se vzorčna porazdelitev vzorčne sredine z večanjem velikosti vzorca približuje normalni porazdelitvi, ne glede na obliko populacijske porazdelitve. Ta teorem nam omogoča, da na podlagi vzorčnih podatkov zanesljivo sklepamo o populacijskih parametrih.

Osrednji mejni teorem je temelj razumevanja normalnih porazdelitev v statistiki. V raziskovalnih okoljih je za pridobitev natančne ocene populacijske sredine pogosto treba zbrati podatke iz številnih naključnih vzorcev znotraj populacije. Te posamezne vzorčne sredine skupaj tvorijo tako imenovano vzorčno porazdelitev sredine.

Centralni limitni teorem opredeljuje dve ključni načeli:

1. **Zakon velikih števil:** S povečevanjem velikosti vzorca ali števila vzorcev se povprečje vzorca približuje povprečju populacije.
2. **Normalnost porazdelitve vzorca:** Pri delu z več velikimi vzorci se kljub prvotni porazdelitvi spremenljivke vzorčna porazdelitev povprečja približa normalni porazdelitvi.

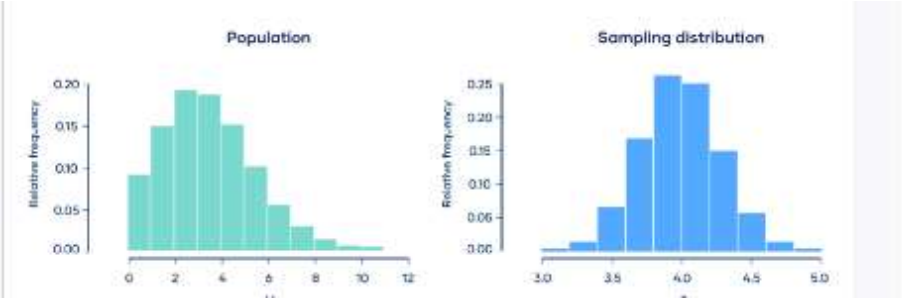
Parametrični statistični testi običajno predpostavljajo, da vzorci izhajajo iz normalno porazdeljenih populacij. Vendar teorem o centralni meji odpravlja potrebo po tej predpostavki za dovolj velike vzorce. Pri velikih vzorcih se parametrični testi lahko uporabljajo ne glede na porazdelitev populacije, če so izpolnjene druge ustrezne predpostavke. Za dovolj velik vzorec se običajno šteje vzorec velikosti 30 ali več.

Nasprotno pa je pri majhnih vzorcih zagotavljanje predpostavke o normalnosti ključnega pomena zaradi negotovosti, povezane z vzorčno porazdelitvijo povprečja. Natančni rezultati zahtevajo

potrditev, da populacija ustreza normalni porazdelitvi, preden se uporabijo parametrični testi z majhnimi vzorci.

Centralni mejni teorem nazorno pravi, da bodo sredine dovolj velikih vzorcev iz populacije pokazale normalno porazdelitev, tudi če osnovna populacijska porazdelitev odstopa od normalnosti.

Primer: Razmislite o populaciji, ki ima Poissonovo porazdelitev (prikazano na levi sliki). Ko iz te populacije vzamemo 10.000 vzorcev, od katerih je vsak sestavljen iz 50 opazovanj, se porazdelitev srednjih vrednosti vzorcev v skladu s centralnim limitnim teoremom natančno ujema z normalno porazdelitvijo (kot je prikazano na desni sliki).



Slika 2.8 Primer populacije v Poissonovi in normalni porazdelitvi.

Osrednji mejni teorem temelji na pojmu vzorčne porazdelitve, ki predstavlja verjetnostno porazdelitev statistike, izračunane iz številnih vzorcev, odvzetih iz populacije.

Konceptualizacija poskusa lahko pomaga pri razumevanju porazdelitev vzorčenja:

- Predstavljajmo si, da izberemo naključni vzorec iz populacije in izračunamo statistiko, na primer srednjo vrednost.
- Nato se izbere še en naključni vzorec enake velikosti in ponovno se izračuna povprečje.
- Ta postopek se večkrat ponovi, tako da dobimo množico srednjih vrednosti, ki ustrezajo posameznemu vzorcu.

Skupek teh vzorčnih povprečij je primer vzorčne porazdelitve. V skladu s centralnim limitnim teoremom se vzorčna porazdelitev povprečja nagiba k normalni porazdelitvi, če je velikost vzorca dovolj velika. Ne glede na porazdelitev populacije - naj bo normalna, Poissonova, binomska ali druga - vzorčna porazdelitev povprečja kaže normalnost.

Na srečo nam ni treba večkrat vzorčiti populacije, da bi ugotovili obliko vzorčne porazdelitve. Namesto tega so parametri vzorčne porazdelitve povprečja odvisni od parametrov same populacije.

- Srednja vrednost vzorčne porazdelitve je srednja vrednost populacije.

$\mu_{\bar{x}} = \mu$

- Standardni odklon vzorčne porazdelitve je standardni odklon populacije, deljen s kvadratnim korenom velikosti vzorca.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

S tem zapisom lahko opišemo porazdelitev vzorčenja povprečja:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Kje:

- $\bar{X}$ je vzorčna porazdelitev vzorčnih sredin
- $\sim$ pomeni "sledi porazdelitvi"
- N je normalna porazdelitev
- μ je srednja vrednost populacije
- σ je standardni odklon populacije
- n je velikost vzorca

Velikost vzorca, označena z n, predstavlja število opazovanj iz populacije za vsak vzorec, pri čemer se ohrani enakomernost vseh vzorcev. Velikost vzorca pomembno vpliva na vzorčno porazdelitev povprečja v dveh ključnih vidikih.

1. Velikost vzorca in normalnost:

- Pri večjih vzorcih so porazdelitve vzorčenja običajno podobne normalni porazdelitvi.
- Nasprotno pa lahko pri majhnih vzorcih vzorčna porazdelitev povprečja odstopa od normalnosti. To odstopanje nastane, ker je veljavnost centralnega limitnega teorema odvisna od "dovolj velike" velikosti vzorca.
- Običajno se za "dovolj velik" vzorec šteje vzorec 30 ali več oseb.
- Kadar je n<30, osrednji mejni teorem ne velja in porazdelitev vzorca odraža porazdelitev populacije. Zato je porazdelitev vzorčenja normalna le, če je populacijska porazdelitev normalna.
- Nasprotno pa pri n≥30 velja osrednji limitni teorem in porazdelitev vzorčenja se približa normalni porazdelitvi.

2. Velikost vzorca in standardni odkloni:

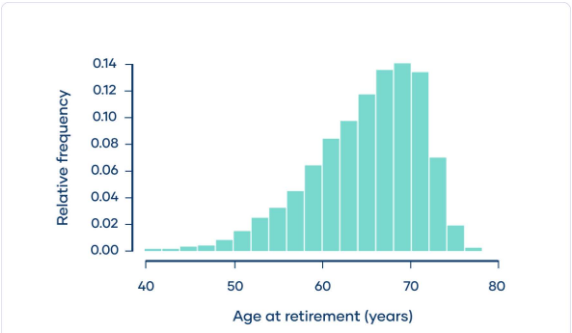
- Velikost vzorca neposredno vpliva na standardni odklon vzorčne porazdelitve, ki odraža variabilnost ali razpršenost porazdelitve.
- Pri manjših vzorcih je standardni odklon običajno večji, kar kaže na večjo variabilnost vzorčnih povprečij zaradi njihove nenatančne ocene populacijskega povprečja.
- Nasprotno pa večje velikosti vzorcev pomenijo manjše standardne odklone, kar kaže na manjšo variabilnost vzorčnih povprečij zaradi natančnejše ocene populacijskega povprečja.

Pomen centralnega limitnega teorema:

Parametrični testi, kot so t-testi, ANOVA in linearna regresija, imajo večjo statistično moč v primerjavi z večino neparametričnih testov. Ta večja statistična moč izhaja iz predpostavk o porazdelitvi populacij, ki temeljijo na osrednjem mejnem teoremu.

Neprekinjena porazdelitev

Poglejmo upokojitveno starost posameznikov v Združenih državah Amerike. Populacijo sestavljajo vsi upokojeni Američani, porazdelitev te populacije pa bi lahko predstavili na naslednji način:



Slika 2.9 Graf zvezne porazdelitve.

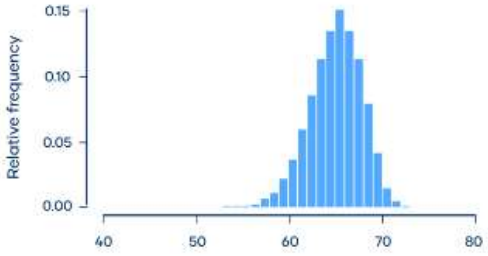
Porazdelitev upokojitvene starosti je nagnjena v levo, saj se večina upokojuje približno pet let pred povprečno upokojitveno starostjo 65 let. Vendar pa obstaja daljši rep posameznikov, ki se upokojijo veliko prej, na primer pri 50 ali celo 40 letih. Standardni odklon populacije je 6 let.

Predstavljajte si, da bi v tej populaciji izvedli manjše vzorčenje. Naključno izberemo pet upokojencev in zabeležimo njihovo upokojitveno starost. Na primer: 68, 73, 70, 62, 63

Povprečje tega vzorca služi kot približek povprečja populacije, čeprav z omejeno natančnostjo zaradi majhnega vzorca 5. Na primer: Povprečje = (68 + 73 + 70 + 62 + 63) / 5 je rezultat povprečje = 67,2 let.

Predpostavimo, da se ta postopek vzorčenja ponovi desetkrat, pri čemer vsak vzorec vključuje pet upokojencev. Izračuna se srednja vrednost vsakega vzorca, s čimer dobimo porazdelitev, znano kot vzorčna porazdelitev srednje vrednosti. Na primer: 60,8; 57,8; 62,2; 68,6; 67,4; 67,8; 68,3; 65,6; 66,5; 62,1.

Ker se ta postopek večkrat ponovi, se histogram, ki prikazuje sredine teh vzorcev, približa normalni porazdelitvi.



Slika 2.10 Normalna porazdelitev srednjih

Kljub temu da je porazdelitev vzorčenja v primerjavi s populacijo nekoliko bolj normalne oblike, je še vedno rahlo nagnjena v levo. Poleg tega je očitno, da je variabilnost vzorčne porazdelitve manjša od variabilnosti populacije.

V skladu s centralnim limitnim teoremom se vzorčna porazdelitev povprečja z večanjem velikosti vzorca približuje normalni porazdelitvi. Vendar trenutna vzorčna porazdelitev povprečja odstopa od normalnosti zaradi relativno majhnega vzorca.

2.8 Testna statistika

Testna statistika je številčna vrednost, pridobljena s testom statistične hipoteze, ki označuje stopnjo usklajenosti med opazovanimi podatki in porazdelitvijo, pričakovano pri ničelni hipotezi tega testa.

Ta statistika ima ključno vlogo pri izračunu p-vrednosti vaših ugotovitev, kar olajša odločitev o sprejetju ali zavrnitvi vaše ničelne hipoteze.

Toda kaj točno je testna statistika?

Testna statistika izraža podobnost med porazdelitvijo vaših podatkov in porazdelitvijo, predvideno pri ničelni hipotezi uporabljenega statističnega testa. Porazdelitev podatkov pojasnjuje pogostost vsakega opazovanja, za katero sta značilni njegova osrednja tendenca in variabilnost okoli nje. Ker različni statistični testi predvidevajo različne tipe porazdelitve, se izbira ustreznega testa uskladi z vašo hipotezo.

Testna statistika strne vaše opazovane podatke v eno samo številko, pri tem pa uporablja merila, kot so centralna tendenca, variabilnost, velikost vzorca in število napovednih spremenljivk v vašem statističnem modelu.

Običajno testna statistika izhaja iz opaznih vzorcev v vaših podatkih (npr. korelacij med spremenljivkami ali razlik med skupinami), deljenih z varianco podatkov (tj. standardnim odklonom).

Oglejte si naslednji primer:

Raziskujete povezavo med temperaturo in datumom cvetenja pri določeni vrsti jablane. Analizirajte obsežen nabor podatkov, ki zajema 25 let, in spremljajte temperaturo in datume cvetenja z naključnim letnim vzorčenjem 100 dreves na poskusnem polju.

- Ničelna hipoteza (H_0): Med temperaturo in datumom cvetenja ni korelacije.

- Nadomestna hipoteza (H_A ali H_1): Obstaja povezava med temperaturo in datumom cvetenja.

Da bi preverili to hipotezo, izvedete regresijski test, pri čemer je testna statistika t-vrednost. Ta t-vrednost primerja ugotovljeno korelacijo med spremenljivkama z ničelno hipotezo o ničelni korelaciji.

2.9 Vrste testne statistike

V nadaljevanju je opisan povzetek prevladujočih statističnih testov skupaj z ustreznimi hipotezami in kategorijami statističnih testov, v katerih se uporabljajo. Čeprav lahko različni statistični testi uporabljajo različne metodologije za izračun teh statistik, temeljne hipoteze in razlage testne statistike ostajajo enake.

Testna statistika	**Ničelna in alternativna hipoteza**	**Statistični testi, ki ga uporabljajo**
vrednost *t*	**Nič:** Srednji vrednosti dveh skupin sta enaki **Druga možnost:** Srednji vrednosti dveh skupin nista enaki	• T-test • Regresijski testi
vrednost *z*	**Nič:** Srednji vrednosti dveh skupin sta enaki **Druga možnost:** Srednji vrednosti dveh skupin nista enaki	• Test *Z*

Testna statistika	Ničelna in alternativna hipoteza	Statistični testi, ki ga uporabljajo
Vrednost *F*	**Nič:** Razlika med dvema ali več skupinami je večja ali enaka razliki med skupinami **Druga možnost:** Razlike med dvema ali več skupinami so manjše od razlik med skupinami	• ANOVA • ANCOVA • MANOVA
***X*-vrednost²**	**Nič:** Dva vzorca sta neodvisna **Druga možnost:** Dva vzorca nista neodvisna (tj. sta korelirana)	• Test Chi-kvadrat • Neparametrični korelacijski testi

V resničnem svetu boste testno statistiko običajno izračunali s statističnim programskim paketom, kot so R, SPSS ali Excel, ki bo tudi podal p-vrednost, povezano s testno statistiko. Kljub temu lahko formule za ročno izračunavanje teh statistik poiščete na spletu.

Pri preverjanju hipoteze o temperaturi in datumu cvetenja na primer izvedete regresijsko analizo. Rezultat regresijskega testa je: - t-vrednost, ki primerja ta koeficient s pričakovanim razponom regresijskih koeficientov pri ničelni hipotezi o odsotnosti povezave.

Rezultat t-vrednosti regresijskega testa, 2,36, predstavlja vašo testno statistiko.

2.10 Standardna napaka

Standardna napaka povprečja (SE ali SEM) služi kot kazalnik verjetne razlike med povprečjem populacije in povprečjem vzorca. Omogoča vpogled v stopnjo variabilnosti, ki bi jo lahko pričakovali v vzorčni sredini, če bi se študija ponovila z uporabo novih vzorcev, odvzetih iz iste populacije.

Standardna napaka povprečja je najpogosteje navedena oblika standardne napake, podobne mere pa obstajajo tudi za druge statistične parametre, kot so mediane ali razmerja. Standardna napaka deluje kot razširjeno merilo napake vzorčenja, ki prikazuje neskladje med populacijskim parametrom in vzorčno statistiko.

Za zmanjšanje standardne napake je priporočljivo povečati velikost vzorca. Uporaba velikega naključnega vzorca je najučinkovitejša strategija za zmanjšanje napake pri vzorčenju in povečanje zanesljivosti ugotovitev.

Standardna napaka in standardni odklon sta merili variabilnosti:

- **Standardni odklon** opisuje variabilnost **znotraj posameznega vzorca.**
- **Standardna napaka** ocenjuje variabilnost **v več vzorcih** populacije.

Standardni odklon je opisna statistika, ki izhaja neposredno iz vzorčnih podatkov, medtem ko je standardna napaka inferenčna statistika, ki je običajno ocenjena, če ni znan natančen parameter populacije.

2.11 Formula za standardno napako

Standardna napaka povprečja se določi tako, da se poleg velikosti vzorca uporabi tudi standardni odklon. S pomočjo formule je razvidno, da sta velikost vzorca in standardna napaka v obratnem sorazmerju. Poenostavljeno povedano, z večanjem velikosti vzorca se standardna napaka zmanjšuje. Do tega pojava pride zato, ker je večji vzorec praviloma bližji vzorčni statistiki populacijskega parametra.

Glede na to, ali je standardni odklon populacije znan, se uporabljajo različne formule. Te formule se uporabljajo za vzorce, ki vsebujejo več kot 20 elementov (n > 20).

Ko so parametri populacije znani

Ko je standardni odklon populacije znan, ga lahko uporabite v spodnji formuli za natančen izračun standardne napake.

Formula

$$SE = \frac{\sigma}{\sqrt{n}}$$

Razlaga:

- *SE* je standardna napaka
- σ je standardni odklon populacije
- *n* je število elementov v vzorcu

Ko so parametri populacije neznani

Kadar standardni odklon populacije ni znan, lahko s spodnjo formulo samo ocenite standardno napako. Ta formula upošteva standardni odklon vzorca kot točkovno oceno standardnega odklona populacije.

Formula

Razlaga:

- *SE* je standardna napaka

Formula **Razlaga:**

$SE = \frac{s}{\sqrt{n}}$

- *s* je standardni odklon vzorca
- *n* je število elementov v vzorcu

Primer: Uporaba formule za standardno napako za oceno standardne napake za matematične rezultate SAT. Izvedite naslednja dva koraka.

Najprej poiščite kvadratni koren velikosti vzorca (*n*).

Formula	**Izračun**
$n = 200$	$\sqrt{n} = \sqrt{200} = 14.1$

Nato standardni odklon vzorca delite s številom, ki ste ga ugotovili v prvem koraku.

Formula	**Izračun**		
$SE = \frac{s}{\sqrt{n}}$	$s = 180$	$\sqrt{n} = 14.1$	$\frac{s}{\sqrt{n}} = \frac{180}{14.1} = 12.8$

Standardna napaka matematičnih rezultatov SAT je 12,8.

Standardno napako lahko predstavite poleg povprečja ali pa jo vključite v interval zaupanja, da izrazite negotovost, ki spremlja povprečje.

Na primer: Povprečni rezultat na naključnem vzorcu udeležencev preizkusa je 550 ± 12,8 (SE).

Poročanje o standardni napaki znotraj intervala zaupanja je bolj priporočljivo, saj tako bralcem ni treba opravljati dodatnih izračunov, da bi dobili smiseln razpon.

Interval zaupanja označuje razpon vrednosti, za katere se pričakuje, da bodo najpogosteje veljale, če bi študijo ponovili z novimi naključnimi vzorci.

Pri 95-odstotni stopnji zaupanja se pričakuje, da bo 95 % vseh vzorčnih povprečij znotraj intervala zaupanja, ki obsega ± 1,96 standardne napake vzorčnega povprečja. Ta interval služi kot ocena, znotraj katere naj bi s 95- odstotnim zaupanjem bil pravi populacijski parameter.

Na primer: Zgradite 95-odstotni interval zaupanja Zgradite 95-odstotni interval zaupanja (CI), da ocenite povprečno oceno populacije za matematični izpit SAT. Pri normalno porazdeljeni lastnosti, kot je ocena SAT, približno 95 % vseh vzorčnih povprečij je znotraj približno 4 standardnih napak vzorčnega povprečja.

Formula za interval zaupanja

$CI = \bar{x} \pm (1{,}96 \times SE)$

$\bar{x}$ =vzorčno povprečje = 550

SE = standardna napaka = 12,8

Spodnja meja	**Zgornja meja**
$\bar{x} - (1{,}96 \times SE)$	$\bar{x} + (1{,}96 \times SE)$
550 - (1.96 × 12.8) = **525**	550 + (1.96 × 12.8) = **575**

Pri naključnem vzorčenju vam 95-odstotni indeks zaupanja [525 575] pove, da obstaja 0,95-odstotna verjetnost, da je povprečni rezultat populacije na testu SAT iz matematike med 525 in 575.

LITERATURA POGLAVJE 2

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/97898151231351230101
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);
- Journal of the Royal Statistical Society 2024, A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011

- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®
- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

DODATNE POVEZAVE DO LITERATURE IN VIDEOPOSNETKOV NA YOUTUBU POGLAVJE 2

- https://open.umn.edu/opentextbooks/textbooks/196
- https://www.scribbr.com/category/statistics/
- https://stats.libretexts.org/Bookshelves/Introductory_Statistics
- https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
- https://saylordotorg.github.io/text_introductory-statistics/
- https://drive.uqu.edu.sa/_/mskhayat/files/MySubjects/20178FS%20Elementary%20Statistics/Introductory%20Statistics%20(7th%20Ed).pdf
- https://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf
- https://www.geeksforgeeks.org/introduction-of-statistics-and-its-types/
- https://onlinestatbook.com/Online_Statistics_Education.pdf
- https://www.researchgate.net/profile/Tareq-Alodat-2/publication/340511098_INTRODUCTION_TO_STATISTICS_MADE_EASY/links/5e8de3dc4585150839c7b58a/INTRODUCTION-TO-STATISTICS-MADE-EASY.pdf
- https://byjus.com/maths/statistics/
- https://www.khanacademy.org/math/statistics-probability