2. Statistika za poslovnu analitiku

Dobrodošli u svet poslovne statistike, gde se podaci pretvaraju u značajne uvide, usmeravajući donošenje odluka i otkrivajući skrivene istine. U ovom sveobuhvatnom istraživanju krećemo na putovanje kako bismo demistifikovali bitne statističke koncepte i tehnike koji podržavaju rigoroznu analizu poslovnih podataka. Od razumevanja zamršenosti distribucija do primene testiranja hipoteza i kreiranja intervala pouzdanosti, svako poglavlje otkriva novi aspekt statističke pismenosti.

U srcu statističke analize leži normalna distribucija, kriva u obliku zvona koja prožima bezbrojne pojave u prirodi i ljudskom ponašanju. U ovom delu ulazimo u srž normalne distribucije, razotkrivajući njena svojstva i značaj u statističkom zaključivanju. Kroz vizualizaciju primera iz stvarnog sveta, rasvetljavamo sveprisutnost ove temeljne distribucije i njenu ulogu kao kamena temeljca statističke teorije.

Standardna devijacija služi kao kompas u statističkom ambijentu, vodeći nas kroz varijabilnost svojstvenu skupovima podataka. U ovom poglavlju razlažemo koncept standardne devijacije, otkrivajući njenu važnost u kvantifikovanju disperzije i proceni raspršenosti podataka. Opremljeni dubljim razumijevanjem standardnih odstupanja, kretaćete se podacima s poverenjem, precizno uočavajući uzorke i netipične vrednosti.

Varijable čine gradivne blokove statističke analize, a svaka poseduje različite karakteristike i implikacije. Ovo poglavlje pojašnjava dihotomiju između kontinuiranih i diskretnih varijabli, prikazujući njihovu ulogu u modeliranju i interpretaciji podataka. Shvatanjem nijansi tipova varijabli, iskoristićete pun potencijal statističkih tehnika prilagođenih različitim strukturama podataka.

Sampling-distribucija služi kao osnov statističkog zaključivanja, premošćavajući jaz između promatranja uzorka i parametara populacije. U ovom poglavlju razotkrivamo koncept sampling-distribucije, razjašnjavajući njegovu relevantnost u izradi verovatnosti o karakteristikama populacije. Kroz konkretne primere razvićete intuitivno razumevanje uloge sampling-distribucije uzorkovanja u robusnoj statističkoj analizi.

Centralna granična teorema je ključni koncept u statistici koji nam pomaže da shvatimo nesigurnost. Ovo poglavlje objašnjava centralnu graničnu teoremu na jednostavan način, pokazujući kako proseke uzorka čini predvidljivijima i pomaže u testiranju hipoteza. Razumevanjem ovog koncepta moći ćete izvući smislene zaključke iz podataka.

Razumevanje testiranja hipoteza bitno je za donošenje odluka na osnovu podataka. Omogućuje nam da utvrdimo jesu li uočeni obrasci u podacima smisleni ili su jednostavno slučajni. Primenom testiranja hipoteza možemo proceniti pretpostavke, uporediti grupe i proceniti statistički značaj rezultata, što ga čini vitalnim alatom u naučnom istraživanju, poslovnoj analizi i mnogim drugim područjima.

Z-standardizovana vrednost i z-tabele služe kao navigaciona pomoć u moru standardne normalne distribucije, olakšavajući standardizovana poređenja i izračunavanja verovatnoće. Ovo poglavlje pojašnjava zamršenost z-standardizovanih vrednosti, jačajući vas da tumačite standardizovane rezultate i koristite Z-tabele za statističku analizu. Uz veštinu o z-standardizovanim vrednostima, kretaćete se ogromnim prostranstvom normalne distribucije s poverenjem i preciznošću.

U situacijama kada su veličine uzorka male ili su standardne devijacije populacije nepoznate, t-rezultati i t-tabele pojavljuju se kao nezamenjivi alati za statističku analizu. Ovo poglavlje razotkriva misterije t-rezultata, vodeći vas kroz njihov izračunavanje i tumačenje pomoću t-tabela. Opremljeni ovim znanjem, lako ćete se snalaziti u nijansama t-distribucija, osiguravajući zaključivanje u različitim statističkim scenarijima.

Normalna i t-distribucija predstavljaju stubove teorije verovatnoćei, a svaka poseduje jedinstvene karakteristike i primene. U ovom poglavlju razjašnjavamo razlike između ovih distribucija, omogućavajući vam da shvatite kada svaku od njih da upotrebite u statističkoj analizi. Kroz praktične primere i komparativne analize, razvićete razumevanje normalne i t-distribucije, obogaćujući svoj skup statističkih alata.

Intervali pouzdanosti pružaju uvid u neizvesnost oko parametara populacije, omogućavajući nam da kvantifikujemo preciznost naših procena. U ovom poglavlju istražujemo konstrukciju intervala pouzdanosti za srednje vrednosti i proporcije, razotkrivajući metodologiju i tumačenje ovih bitnih statističkih alata. Savladivanjem intervala pouzdanosti, transparentno i kritički ćete preneti neizvesnost koja je svojstvena vašim nalazima.

Dok p-vrednosti nude pristup statističkim zaključivanjima, njihovo pogrešno tumačenje može dovesti do pogrešnih zaključaka i pogrešno informisanih odluka. Ovo poglavlje ispituje potencijalne zamke preteranog oslanjanja na p-vrednosti, naglašavajući važnost konteksta i veličine učinka u statističkoj analizi. Kroz kritičko ispitivanje i praktične uvide, pažljivo ćete se kretati kroz složenost p-vrednosti, osiguravajući integritet svojih statističkih zaključaka.

Unutar ovih stranica leže ključevi za otključavanje misterija statističke analize, što vam omogućava da pouzdano i precizno upravljate složenošću podataka. Dok zajedno krećemo na ovo putovanje, neka nam znatiželja bude kompas, a istraživanje naše svetlo vodilja, osvetljavajući put prema dubljem razumevanju i delotvornim zaključcima.

2.1 Normalna distribucija

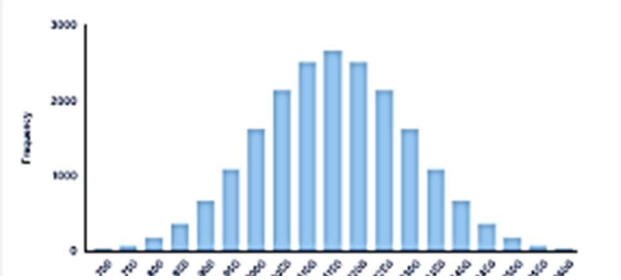
U središtu statističke analize nalazi se normalna distribucija, sveprisutna distribucija verovatnoća koja služi kao merilo za mnoge statističke tehnike. Udubićemo se u njene karakteristike, njenu simetričnu krivu liniju u obliku zvona i značaj u razumevanju distribucije podataka.

Normalna distribucija nalazi primenu u raznim područjima, uključujući finansije, psihologiju, inženjerstvo i biologiju. Od modeliranja cena deonica do razumevanja distribucije visine ljudi, normalna distribucija služi kao svestran alat za analizu i tumačenje podataka.

Kroz ovo poglavlje proučiće se matematička svojstva normalne distribucije, istražujući kako izračunati verovatnoće, percentile i z-središnje vrednosti. Raspravljaćemo o praktičnim tehnikama za vizualizaciju i interpretaciju normalnih distribucija pomoću histograma, dijagrama gustine i funkcija kumulativne distribucije.

Do kraja ovog poglavlja duboko ćete vrednovati normalnu distribuciju i njen značaj u statističkoj analizi. Bićete spremni za rešavanje naprednijih statističkih koncepata i njihovu primenu na skupove podataka u stvarnom svetu. Krenimo na ovo putovanje kako bismo zajedno razotkrili misterije normalne distribucije.

Normalna distribucija, takođe poznata kao Gaussova distribucija ili zvonasta kriva, pokazuje simetričnu distribuciju podataka bez asimetrije. Kada su grafički prikazani, podaci grade krivu liniju u obliku zvona, s većinom vrednosti koje se skupljaju oko središta i smanjuju kako se udaljavaju od njega.

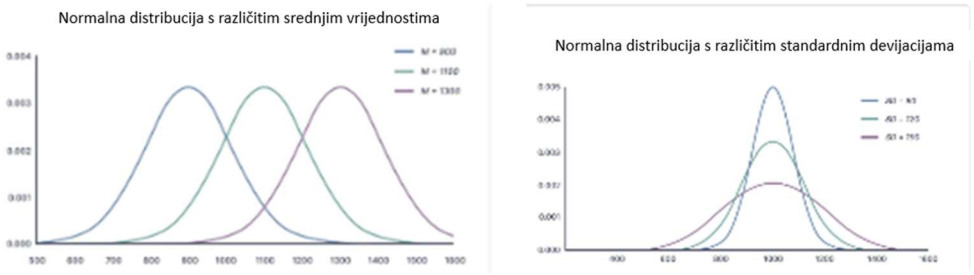


Slika 2.1 Primer Gaussove distribucije ili zvonaste krive

Različite varijable u prirodnim i društvenim naukama obično pokazuju normalnu distribuciju ili su joj blizu. Primeri uključuju visinu, porođajnu težinu, sposobnost čitanja, zadovoljstvo poslom i SAT rezultate. Zbog učestalosti normalno raspodeljenih varijabli, brojni statistički testovi prilagođeni su takvim populacijama. Veština u razumevanju karakteristika normalne distribucije osnažuje pojedince da koriste inferencijalnu statistiku za poređenje grupa i generisanje procena populacije iz uzoraka.

Normalne distribucije imaju ključne karakteristike koje je lako uočiti na grafikonima:

- Srednja vrednost, medijana i mod su potpuno isti.
- Distribucija je simetrična u odnosu na srednju vrednost - polovina vrednosti nalazi se ispod, a polovina iznad srednje vrednosti.
- Distribucija se može opisati s dve vrednosti: srednjom vrednošću i standardnom devijacijom.



Slika 2.28 Normalna distribucija s različitim srednjim vrednostima i različitim standardnim devijacijama.

Srednja vrednost služi kao lokacijski parametar koji diktira središte vrha krive. Podešavanje srednje vrednosti pomera krivu u skladu s tim: povećanje pomera krivu udesno, dok smanjenje pomera krivu ulevo. U međuvremenu, standardna devijacija funkcioniše kao parametar razmera, utičući na širenje ili širinu krive.

Standardna devijacija širi ili sužava krivu. Mala standardna devijacija rezultira uskom krivom linijom, dok velika standardna devijacija dovodi do široke krive.

2.2 Empirijsko pravilo

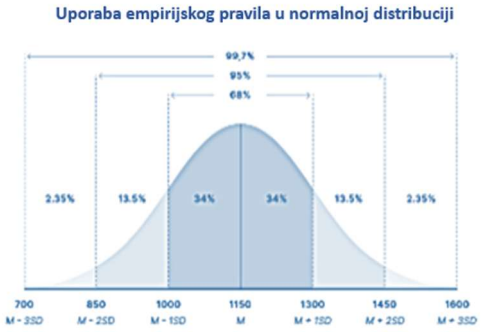
Empirijsko pravilo, takođe poznato kao pravilo 68-95-99.7, daje uvid u raspodelu vrednosti unutar normalne distribucije:

- Otprilike 68% vrednosti pada unutar 1 standardne devijacije od srednje vrednosti.
- Otprilike 95% vrednosti nalazi se unutar 2 standardne devijacije od srednje vrednosti.
- Oko 99,7% vrednosti obuhvaćeno je unutar 3 standardne devijacije od srednje vrednosti.

Na primer, razmotrite scenario u kojem se prikupljaju rezultati SAT-a od učenika na novom kursu pripreme za ispit, a podaci su u skladu s normalnom distribucijom sa srednjom ocenom (M) od 1150 i standardnom devijacijom (SD) od 150.

Primenom empirijskog pravila može se zaključiti:

- Oko 68% rezultata nalazi se u rasponu od 1000 do 1300, što odgovara 1 standardnoj devijaciji iznad i ispod proseka.
- Otprilike 95% rezultata je unutar raspona od 850 do 1450, što predstavlja 2 standardne devijacije iznad i ispod proseka.
- Gotovo svi rezultati, oko 99,7%, leže u rasponu od 700 do 1600, obuhvatajući 3 standardne devijacije iznad i ispod proseka.



Slika 2.54 Empirijsko pravilo u normalnoj distribuciji.

Empirijsko pravilo

nudi brzu metodu za procenu podataka, omogućavajući otkrivanje outliera ili netipičnih vrednosti koje odstupaju od očekivanog obrasca. U slučajevima kada podaci iz malih uzoraka značajno odstupaju od ovog obrasca, alternativne distribucije kao što je t-distribucija mogu biti prikladnije. Identifikovanje distribucije varijable omogućava primenu relevantnih statističkih testova.

2.3 Formula krive linije normalne distribucije

Za konstruiranje krive linije normalne distribucije na osnovu poznate srednje vrednosti i standardne devijacije, može se upotrebiti funkcija gustine verovatnoća, čime se tačno predstavlja distribucija podataka.

Slika 2.74 Kriva normalne distribucije prilagođena podacima SAT rezultata.

Unutar funkcije gustine verovatnoća, područje ispod krive linije predstavlja verovatnoću. S obzirom da normalna distribucija služi kao distribucija verovatnoće, kumulativna površina ispod krive linije uvek iznosi 1 ili 100%. Iako se formula za normalnu funkciju gustine verovatnoća može činiti zamršenom, njeno korišćenje zahteva samo poznavanje srednje vrednosti populacije i standardne devijacije. Zamenom ovih parametara u formuli, može se odrediti gustina verovatnoće povezana s bilo kojom datom vrednošću x.

- $f(x)$ = verovatnoća
- x = vrednost varijable
- μ = srednja vrednost
- σ = standardna devijacija
- σ^2 = varijansa

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

Primer:

Koristeći funkciju gustine verovatnoća, želite znati verovatnoću da SAT rezultati u vašem uzorku premašuju 1380.

Na vašem grafikonu funkcije gustine verovatnoće, verovatnoća je osenčeno područje ispod krive linije koje se nalazi desno od mesta gde je vaš SAT rezultat jednak 1380.

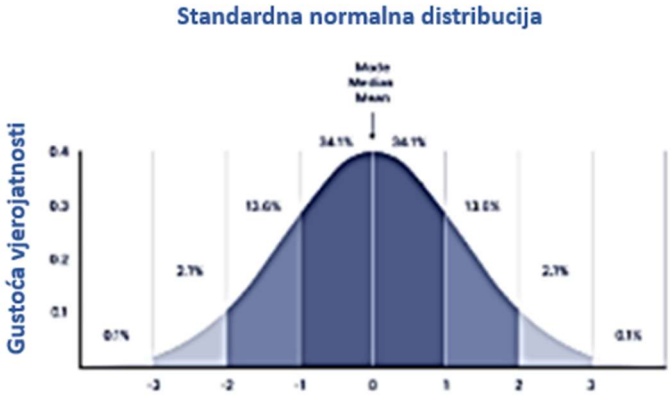
Slika 2.100 Grafikon funkcije gustine verovatnoće SAT rezultata.

Vrednost verovatnoće ovog rezultata možete pronaći pomoću standardne normalne distribucije.

2.4 Standardna normalna distribucija

Standardna normalna distribucija, poznata kao **z-distribucija**, razlikuje se po tome što ima srednju vrednost od 0 i standardnu devijaciju od 1. Svaka normalna distribucija može se promatrati kao transformacija standardne normalne distribucije, koja prolazi kroz prilagođavanja u merama, položaju ili oba.

U kontekstu z-distribucije, pojedinačna opažanja, koja se obično označavaju kao x u normalnim distribucijama, nazivaju se z-standardizovane vrednosti ili z-skorovi. Ovi z-skorovi predstavljaju broj standardnih devijacija za koje svaka vrednost odstupa od srednje vrednosti. Posledično, pretvaranje vrednosti iz bilo koje normalne distribucije u z-skorove olakšava upoređivanje i analizu unutar okvira standardne normalne distribucije.



Slika 2.120 Grafikon standardne normalne distribucije.

Trebate znati samo srednju vrednost i standardnu devijaciju vaše distribucije da biste pronašli *z* -skor vrednosti.

Objašnjenje formule z-skora

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

- x = pojedinačna vrednost
- μ = srednja vrednost
- σ = standardna devijacija

Normalne distribucije pretvaramo u standardnu normalnu distribuciju iz nekoliko razloga:

- kako bismo pronašli verovatnoću opažanja u distribuciji koja pada iznad ili ispod zadate vrednosti;
- kako bismo pronašli verovatnoću da se srednja vrednost uzorka značajno razlikuje od poznate srednje vrednosti populacije.
- za upoređivanje rezultata na različitim distribucijama s različitim srednjim vrednostima i standardnim odstupanjima.

2.5 Određivanje verovatnoće korišćenjem *z* -distribucije

Svaki z-rezultat odgovara verovatnoći, koja se često naziva p-vrednost, koja ukazuje na verovatnoću opažanja vrednosti ispod tog specifičnog z-skora. Transformacijom pojedinačne

vrednosti u z-skor, može se odrediti verovatnoća da se sve vrednosti do te tačke pojave unutar normalne distribucije.

Na primer, razmotrite scenario u kojem želite utvrditi verovatnoću da će SAT rezultati u vašem uzorku premašiti 1380. U početku izračunavate z-skor koristeći srednju vrednost i standardnu devijaciju distribucije. Uz srednju vrednost od 1150 i standardnu devijaciju od 150, z-skor otkriva broj standardnih devijacija za koje 1380 odstupa od srednje vrednosti.

Izračun formule

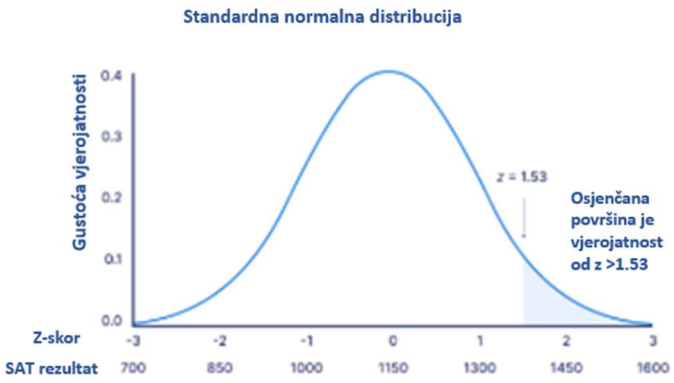
$$z = \frac{x - \mu}{\sigma} = \frac{1380 - 1150}{150} = 1.53$$

Za *z* -skor od 1,53, *p* -vrednost je 0,937. Ovo je verovatnoća da će SAT rezultati biti 1380 ili manje (93,7%), a to je područje ispod krive linije levo od osenčanog područja.

Da biste pronašli osenčano područje, oduzmite 0,937 od 1, što je ukupna površina ispod krive linije.

Verovatnoću x > 1380 = 1 – 0,937 = **0,063**

To znači da je verovatno da samo 6,3% SAT rezultata u vašem uzorku prelazi 1380.



Slika 2.128 Standardna normalna distribucija s naznačenim SAT

2.6 Sampling-distribucija

Sampling-distribucije čine okosnicu statističkog zaključivanja, omogućavajući nam izvođenje zaključaka o populacijama na osnovu podataka iz uzorka. Udubićemo se u zamršenost

sampling-distribucija, da bi razumeli kako se odražava varijabilnost statistike uzorka i njihova ključna uloga u testiranju hipoteza.

Sampling-distribucija odnosi se na distribuciju statističkih podataka, kao što je srednja vrednost uzorka ili proporcija uzorka, dobijenih iz više uzoraka iste veličine izvučenih iz populacije. Pruža uvid u ponašanje statistike uzorka i njihovu varijabilnost u različitim uzorcima.

2.7 Centralna granična teorema i sampling-distribucija

Centralna granična teorema (engl. Central Limit Theorem - CLT) je osnovni koncept u statistici koji podržava ponašanje sampling-distribucije. Navodi se da se sampling-distribucija srednje vrednosti uzorka približava normalnoj distribuciji kako se veličina uzorka povećava, bez obzira na oblik distribucije populacije. Ova teorema nam omogućava da izvedemo čvrste zaključke o parametrima populacije iz uzorka podataka.

Centralna granična teorema služi kao kamen temeljac razumevanja normalnih distribucija u statistici. U uslovima istraživanja, dobijanje tačne procene srednje vrednosti populacije često uključuje prikupljanje podataka iz brojnih slučajnih uzoraka unutar populacije. Te pojedinačne srednje vrednosti uzoraka zajedno čine ono što je poznato kao sampling-distribucija srednje vrednosti.

Centralna granična teorema ističe dva ključna principa:

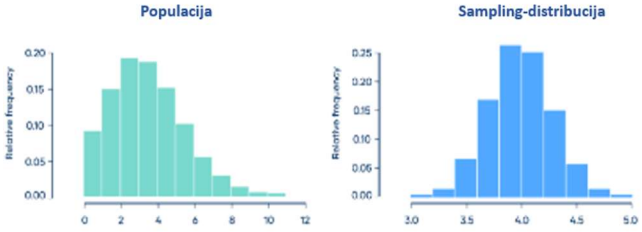
1. **Zakon velikih brojeva:** kako se veličina uzorka ili broj uzoraka povećava, srednja vrednost uzorka nastoji se približiti srednjoj vrednosti populacije.
2. **Normalnost sampling-distribucije:** uprkos izvornoj distribuciji varijable, kada se radi s višestrukim velikim uzorcima, sampling-distribucija srednje vrednosti teži približnoj normalnoj distribuciji.

Parametarski statistički testovi konvencionalno pretpostavljaju da su uzorci izvedeni iz normalno distribuiranih populacija. Međutim, centralna granična teorema uklanja nužnost ove pretpostavke za dovoljno velike uzorke. S velikim uzorcima, parametarski testovi mogu se primeniti bez obzira na distribuciju populacije, pod uslovom da su zadovoljene druge relevantne pretpostavke. Veličina uzorka od 30 ili više obično se smatra dovoljno velikim.

Nasuprot tome, za male uzorke, osiguravanje pretpostavke normalnosti je ključno zbog nesigurnosti koja okružuje sampling-distribuciju srednje vrednosti. Tačni rezultati zahtevaju potvrdu da se populacija pridržava normalne distribucije pre korišćenja parametarskih testova s malim uzorcima.

Ilustrativno, centralna granična teorema tvrdi da će dobijanjem dovoljno velikih uzoraka iz populacije, srednje vrednosti tih uzoraka pokazati normalnu distribuciju, čak i ako osnovna distribucija populacije odstupa od normalnosti.

Primer: Razmotrite populaciju prema Poissonovoj distribuciji (prikazano na levoj slici). Nakon izvlačenja 10 000 uzoraka iz ove populacije, od kojih se svaki sastoji od 50 opažanja, distribucija srednjih vrednosti uzorka blisko je usklađena s normalnom distribucijom, u skladu s centralnom graničnom teoremom (kao što je ilustrovano na desnoj slici).



Slika 2.155 Primer populacije u Poissonovoj distribuciji i normalnoj distribuciji.

Centralna granična teorema zavisi od pojma sampling-distribucije, koja predstavlja distribuciju verovatnoće statistike izračunate iz brojnih uzoraka izvučenih iz populacije.

Konceptualizacija eksperimenta može pomoći u shvatanju sampling-distribucije:

- Zamislimo izvlačenje slučajnog uzorka iz populacije i izračunavanje statistike, kao što je srednja vrednost.
- Nakon toga se izvlači još jedan slučajni uzorak identične veličine, a srednja vrednost se ponovno izračunava.
- Ovaj proces se ponavlja mnogo puta, što rezultuje mnoštvom srednjih vrednosti, od kojih svaka odgovara uzorku.

Združivanje ovih srednjih vrednosti uzoraka predstavlja primer sampling-distribucije. Prema centralnoj graničnoj teoremi, sampling-distribucija srednje vrednosti teži normalnoj

distribuciji kada je veličina uzorka dovoljno velika. Nevjerojatno, bez obzira na distribuciju populacije - bila ona normalna, Poissonova, binomna ili neka druga – sampling-distribucija srednje vrednosti pokazuje normalnost.

Srećom, ne treba više puta uzorkovati populaciju da bi se utvrdio oblik sampling-distribucije. Umesto toga, parametri sampling-distribucije srednje vrednosti zavise od parametara same populacije.

- Srednja vrednost sampling-distribucije je srednja vrednost populacije.

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

- Standardna devijacija sampling-distribucije je standardna devijacija populacije podeljena s kvadratnim korenom veličine uzorka.

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

Sampling-distribuciju srednje vrednosti možemo opisati pomoću ove oznake:

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

gde:

- $\bar{X}$ je sampling-distribucija srednjih vrednosti uzorka
- ~ znači "sledi distribuciju"
- N je normalna distribucija
- μ je srednja vrednost populacije
- σ je standardna devijacija populacije
- n je veličina uzorka.

Veličina uzorka, označena kao n, predstavlja broj opažanja izvučenih iz populacije za svaki uzorak, održavajući ujednačenost u svim uzorcima. Veličina uzorka značajno utiče na sampling-distribuciju srednje vrednosti u dva ključna aspekta.

1. Veličina uzorka i normalnost:

- Veći uzorci obično daju sampling-distribucije koje su bliske normalnoj distribuciji.

- Suprotno tome, s malim uzorcima, sampling-distribucija srednje vrednosti može odstupati od normalnosti. Ovo odstupanje nastaje jer valjanost centralne granične teoreme zavisi od "dovoljno velike" veličine uzorka.
- Uobičajeno, uzorak od 30 ili više smatra se "dovoljno velikim".
- Kada je n<30, centralna granična teorema se ne primenjuje, a sampling-distribucija odražava distribuciju populacije. Stoga je sampling-distribucija normalna samo ako je distribucija populacije normalna.
- Nasuprot tome, kada je n≥30, centralna granična teorema važi, a sampling-distribucija približava se normalnoj distribuciji.

2. Veličina uzorka i standardna devijacija:

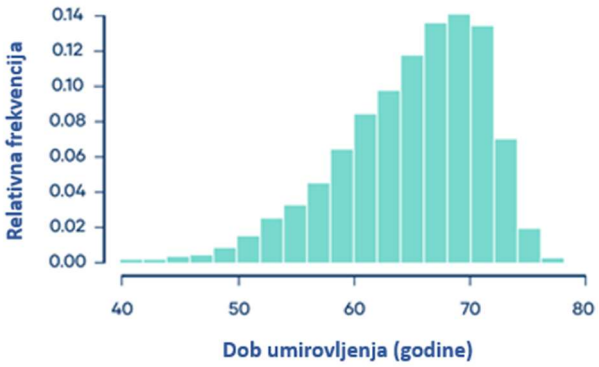
- Veličina uzorka direktno utiče na standardnu devijaciju sampling-distribucije, odražavajući varijabilnost ili raspršenost distribucije.
- S manjim uzorcima, standardna devijacija obično je viša, što ukazuje na veću varijabilnost među srednjim vrednostima uzorka zbog njihove neprecizne procene srednje vrednosti populacije.
- Suprotno tome, veći uzorci odgovaraju nižim standardnim devijacijama, što ukazuje na manju varijabilnost među srednjim vrednostima uzorka zahvaljujući njihovoj tačnijoj proceni srednje vrednosti populacije.

Važnost centralne granične teoreme:

Parametarski testovi kao što su t-testovi, ANOVA i linearna regresija imaju veću statističku snagu u poređenju s većinom neparametarskih testova. Ova povećana statistička snaga proizilazi iz pretpostavki o distribuciji populacija, koje su zasnovane na centralnoj graničnoj teoremi.

Kontinuirana distribucija:

Razmotrimo starost za odlazak u penziju pojedinaca u Sjedinjenim Američkim Državama. Stanovništvo se sastoji od svih penzionisanih Amerikanaca, a distribucija ovog stanovništva može se predstaviti na sledeći način:



Slika 2.182 Grafikon kontinuirane distribucije

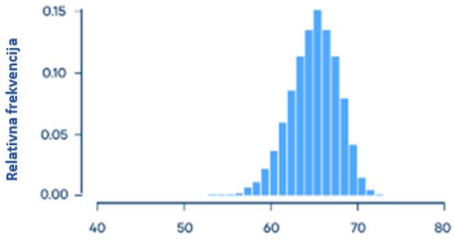
Distribucija starosti za penzionisane iskrivljena je ulevo, pri čemu većina odlazi u penziju unutar približno pet godina od prosečne starosti za penzionisanje od 65 godina. Međutim, postoji prošireni rep pojedinaca koji odlaze u penziju puno ranije, npr. s 50 ili čak 40 godina. Populacija pokazuje standardnu devijaciju od 6 godina.

Zamislite sprovođenje malog uzorkovanja ove populacije. Nasumično se izabere pet penzionera i beleži se njihova starost za odlazak u penziju. Na primer: 68, 73, 70, 62, 63.

Srednja vrednost ovog uzorka služi kao procena srednje vrednosti populacije, iako s ograničenom preciznošću zbog male veličine uzorka od 5. Na primer: Srednja vrednost = (68 + 73 + 70 + 62 + 63) / = 67,2 godine

Sada pretpostavimo da se ovaj proces uzorkovanja ponovi 10 puta, a svaki uzorak uključuje pet penzionera. Izračunava se srednja vrednost svakog uzorka, što rezultuje distribucijom poznatom kao sampling-distribucija srednje vrednosti. Na primer: 60,8; 57,8; 62,2; 68,6; 67,4; 67,8; 68,3; 65,6; 66,5; 62,1.

Budući da se ovaj proces ponavlja mnogo puta, histogram koji prikazuje srednje vrednosti ovih uzoraka približno će odgovarati normalnoj distribuciji.



Slika 2.208 Normalna distribucija srednjih vrijednosti

Uprkos tome što sampling-distribucija pokazuje nešto normalniji oblik u poređenju s populacijom, još uvek zadržava blagi zaokret ulevo. Osim toga, evidentno je da je varijabilnost u sampling-distribuciji uža od one populacije.

Prema centralnoj graničnoj teoremi, sampling-distribucija srednje vrednosti nastoji se približiti normalnoj distribuciji kako se veličina uzorka povećava. Međutim, trenutna sampling-distribucija srednje vrednosti odstupa od normalne zbog relativno male veličine uzorka.

2.8 Testna statistika

Testna statistika predstavlja brojnu vrednost izvedenu iz testiranja statističke hipoteze koja ukazuje na stepen usklađenosti između vaših opaženih podataka i očekivane distribucije prema nultoj hipotezi tog testa.

Ova statistika igra ključnu ulogu u izračunavanju p-vrednosti vaših nalaza, olakšavajući odluku o prihvatanju ili odbacivanju vaše nulte hipoteze.

Ali šta tačno čini testnu statistiku?

Testna statistika artikulše sličnost između distribucije vaših podataka i distribucije predviđene prema nultoj hipotezi korišćenog statističkog testa. Distribucija podataka razjašnjava učestalost svakog opažanja, koju karakteriše centralna tendencija i varijabilnost oko nje. Budući da različiti statistički testovi predviđaju različite vrste distribucije, izbor odgovarajućeg testa usklađen je s vašom hipotezom.

Testna statistika sažima vaše opažene podatke u jedinstvenu vrednost, koristeći mere kao što su centralna tendencija, varijabilnost, veličina uzorka i broj varijabli predviđanja u vašem statističkom modelu.

Tipično, testna statistika proizlazi iz vidljivih obrazaca u vašim podacima (npr. korelacije između varijabli ili odstupanja među grupama), podeljenih s varijansom podataka (tj. standardnom devijacijom).

Razmotrite ovaj primer:

Istražujete povezanost između temperature i datuma cvetanja kod određene vrste stabla jabuke. Analizirajući opsežan skup podataka koji obuhvata 25 godina, prateći temperaturu i datume cvetanja nasumičnim uzorkovanjem 100 stabala godišnje s eksperimentalnog polja.

• Nulta hipoteza (H_0): Ne postoji korelacija između temperature i datuma cvetanja.

• Alternativna hipoteza (H_A ili H_1): Postoji korelacija između temperature i datuma cvetanja.

Da biste ispitali ovu hipotezu, sprovodite regresioni test, dajući t-vrednost kao testnu statistiku. Ova t-vrednost suprotstavlja uočenu korelaciju između varijabli naspram nulte hipoteze koja pretpostavlja da nema korelacije.

2.9 Vrste testne statistike

U nastavku je prikazan sinopsis preovlađujućih testnih statistika, zajedno s njihovim odgovarajućim hipotezama i kategorijama statističkih testova u kojima se koriste. Iako različiti statistički testovi mogu koristiti različite metodologije za izračunavanje ovih statistika, osnovne hipoteze i tumačenja testne statistike ostaju dosledni.

Testna statistika	**Nulta i alternativna hipoteza**	**Statistički testovi**
***t* vrednost**	**Nulta:** Srednje vrednosti dve grupe su jednake. **Alternativna:** Srednje vrednosti dve grupe nisu jednake.	• *T* test • Regresijski testovi
***z* vrednost**	**Nulta:** Srednje vrednosti dve grupe su jednake. **Alternativna:** Srednje vrednosti dve grupe nisu jednake.	• *Z* test
***F* vrednost**	**Nulta:** Varijacija između dve ili više grupa veća je ili jednaka varijaciji između grupa. **Alternativna:** Varijacije između dve ili više grupa su manje od varijacija između grupa.	• ANOVA • ANCOVA • MANOVA
***X*²-vrednost**	**Nulta:** Dva su uzorka nezavisna.	• Hi-kvadrat test

Testna statistika	Nulta i alternativna hipoteza	Statistički testovi
	Alternativna: Dva uzorka nisu nezavisna (tj. koreliraju).	• Neparametarski korelacioni testovi

U scenarijima iz stvarnog sveta, obično ćete izračunati svoju testnu statistiku koristeći statistički softverski paket kao što je R, SPSS ili Excel, koji će takođe dati p-vrednost povezanu sa testnom statistikom. Uprkos tome, formule za ručno izračunavanje ovih statistika mogu se pronaći na internetu.

Na primer, u testiranju vaše hipoteze o temperaturi i datumima cvetanja, provodite regresionu analizu. Regresioni test daje: • regresioni koeficijent od 0,36 • t-vrednost koja upoređuje ovaj koeficijent s očekivanim rasponom regresionih koeficijenata pod nultom hipotezom nepostojanja veze.

Rezultujuća t-vrednost iz regresionog testa od 2,36 predstavlja vašu testnu statistiku.

2.10 Standardna greška

Standardna greška srednje vrednosti (engl. *standard error of the mean* - SE ili SEM) služi kao pokazatelj verovatne razlike između srednje vrednosti populacije i srednje vrednosti uzorka. Nudi uvid u stepen varijabilnosti koji bi se očekivao u srednjoj vrednosti uzorka ako bi se studija replicirala koristeći sveže uzorke izvučene iz iste populacije.

Dok je standardna greška srednje vrednosti najčešće citirani oblik standardne greške, slične mere postoje za druge statističke parametre kao što su medijana ili proporcije. Standardna greška funkcionše kao prevladujuća mera greške uzorkovanja, prikazujući nejednakost između parametra populacije i statistike uzorka.

Kako bi se ublažila standardna greška, preporučuje se povećanje veličine uzorka. Korišćenje velikog, slučajnog uzorka služi kao najučinkovitija strategija za smanjenje pristranosti uzorkovanja i povećanje pouzdanosti nalaza.

Standardna greška i standardna devijacija su mere varijabilnosti:

- **Standardna devijacija** opisuje varijabilnost **unutar jednog uzorka**.
- **Standardna greška** procenjuje varijabilnost **u višestrukim uzorcima** populacije.

Standardna devijacija služi kao deskriptivna statistika izvedena direktno iz podataka uzorka, dok standardna greška predstavlja inferencijalnu statistiku, obično procenjenu, osim ako nije poznat tačan parametar populacije.

2.11 Formula standardne greške

Standardna greška srednje vrednosti određena je primenom standardne devijacije uz veličinu uzorka. Kroz formulu postaje očito da su veličina uzorka i standardna greška u obrnutom odnosu. Jednostavnije rečeno, kako se veličina uzorka povećava, standardna greška se smanjuje. Do ovog fenomena dolazi jer veći uzorak ima tendenciju dati statističke podatke uzorka bliže parametru populacije.

Koriste se različite formule na osnovu toga da li je poznata standardna devijacija populacije. Ove formule su primenjive na uzorke koji sadrže više od 20 elemenata (n > 20).

Ako su poznati parametri populacije

Kada je poznata standardna devijacija populacije, možete je koristiti u donjoj formuli za tačno izračunavanje standardne greške.

Formula **Obrazloženje**

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

- SE je standardna greška
- σ je standardna devijacija populacije
- n je broj elemenata u uzorku

Ako su parametri populacije nepoznati

Kada je standardna devijacija populacije nepoznata, možete koristiti donju formulu samo za procenu standardne greške. Ova formula uzima standardnu devijaciju uzorka kao procenu standardne devijacije populacije.

Formula **Obrazloženje**

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

- SE je standardna greška
- s je standardna devijacija uzorka
- n je broj elemenata u uzorku

Primer: Korišćenje formule standardne greške za procenu standardne greške za rezultate SAT-a iz matematike. Sledite sledeća dva koraka.

Prvo pronađite kvadratni koren veličine uzorka (*n*).

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

Zatim podelite standardnu devijaciju uzorka s brojem koji ste pronašli u prvom koraku.

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

Standardna greška rezultata SAT iz matematike je 12,8.

Možete predstaviti standardnu grešku uz srednju vrednost ili je uključiti u interval pouzdanosti kako biste preneli nesigurnost koja okružuje srednju vrednost.

Na primer: Prikaz srednje vrednosti i standardne greške. Srednji rezultat SAT-a iz matematike za slučajni uzorak ispitanika je 550 ± 12,8 (SE).

Izveštavanje o standardnoj grešci unutar intervala pouzdanosti je poželjno jer eliminiše potrebu čitaoca za izvođenje dodatnih izračunavanja kako bi dobili smisleni raspon.

Interval pouzdanosti označava raspon vrednosti gde se očekuje da će nepoznati parametar populacije najčešće biti ako bi se studija ponovila s novim slučajnim uzorcima.

Na nivou pouzdanosti od 95%, očekuje se da će 95% svih srednjih vrednosti uzorka pasti unutar intervala pouzdanosti koji obuhvata ± 1,96 standardnih grešaka srednje vrednosti uzorka. Ovaj interval služi kao procena unutar koje se veruje da se stvarni parametar populacije nalazi unutar 95% pouzdanosti.

Na primer: Konstruisanje intervala pouzdanosti od 95%. Konstruišete interval pouzdanosti od 95% (CI) da biste procenili srednju vrednost matematičke SAT ocene populacije. S obzirom na normalno raspodijeljenu karakteristiku kao što su SAT rezultati, otprilike 95% svih srednjih vrednosti uzorka pada unutar približno 4 standardne greške srednje vrednosti uzorka.

Formula intervala pouzdanosti

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

$\bar{x}$ = srednja vrednost uzorka = 550

SE = standardna greška = 12,8

Donja granica	**Gornja granica**
$\bar{x} - (1{,}96 \times SE)$	$\bar{x} + (1{,}96 \times SE)$
550 − (1,96 × 12,8) = **525**	550 + (1,96 × 12,8) = **575**

S slučajnim uzorkovanjem, 95% CI [525 575] ukazuje da postoji verovatnoća od 0,95 da je srednja vrednost matematičkog SAT rezultata populacije između 525 i 575.

Literatura 2. poglavlja

- *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 poveznice na literaturu i Youtube videozapise 2. poglavlja

- 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