2. Statistika za poslovnu analitiku

Dobrodošli u svijet poslovne statistike, gdje se podaci pretvaraju u značajne uvide, usmjeravajući donošenje odluka i otkrivajući skrivene istine. U ovom sveobuhvatnom istraživanju krećemo na putovanje kako bismo demistificirali bitne statističke koncepte i tehnike koji podupiru rigoroznu analizu poslovnih podataka. Od razumijevanja zamršenosti distribucija do primjene testiranja hipoteza i konstruiranja intervala pouzdanosti, svako poglavlje otkriva novi aspekt statističke pismenosti.

U srcu statističke analize leži normalna distribucija, krivulja u obliku zvona koja prožima bezbrojne pojave u prirodi i ljudskom ponašanju. U ovom dijelu ulazimo u bit normalne distribucije, razotkrivajući njena svojstva i značaj u statističkom zaključivanju. Kroz vizualizaciju primjera iz stvarnog svijeta, rasvjetljavamo sveprisutnost ove temeljne distribucije i njezinu ulogu kao kamena temeljca statističke teorije.

Standardna devijacija služi kao kompas u statističkom krajoliku, vodeći nas kroz varijabilnost svojstvenu skupovima podataka. U ovom poglavlju rastavljamo koncept standardne devijacije, otkrivajući njezinu važnost u kvantificiranju disperzije i procjeni raspršenosti podataka. Opremljeni dubljim razumijevanjem standardnih odstupanja, kretat ćete se podacima s povjerenjem, precizno uočavajući uzorke i netipične vrijednosti.

Varijable čine građevne blokove statističke analize, a svaka posjeduje različite karakteristike i implikacije. Ovo poglavlje pojašnjava dihotomiju između kontinuiranih i diskretnih varijabli, bacajući svjetlo na njihovu ulogu u modeliranju i interpretaciji podataka. Shvaćanjem nijansi tipova varijabli, iskoristit ćete puni potencijal statističkih tehnika prilagođenih različitim strukturama podataka.

Sampling-distribucija služi kao temelj statističkog zaključivanja, premošćujući jaz između promatranja uzorka i parametara populacije. U ovom poglavlju razotkrivamo koncept sampling-distribucije, razjašnjavajući njegovu relevantnost u izradi vjerojatnosti o karakteristikama populacije. Kroz konkretne primjere razvit ćete intuitivno razumijevanje uloge sampling-distribucije uzorkovanja u robusnoj statističkoj analizi.

Centralni granični teorem je ključni koncept u statistici koji nam pomaže da shvatimo nesigurnost. Ovo poglavlje objašnjava centralni granični teorem na jednostavan način,

pokazujući kako čini prosjeke uzorka predvidljivijima i pomaže u testiranju hipoteza. Razumijevanjem ovog koncepta moći ćete izvući smislene zaključke iz podataka.

Razumijevanje testiranja hipoteza bitno je za donošenje odluka na temelju podataka. Omogućuje nam da utvrdimo jesu li uočeni obrasci u podacima smisleni ili su jednostavno slučajni. Primjenom testiranja hipoteza možemo procijeniti pretpostavke, usporediti grupe i procijeniti statistički značaj rezultata, što ga čini vitalnim alatom u znanstvenom istraživanju, poslovnoj analizi i mnogim drugim područjima.

Z-standardizirana vrijednost i z-tablice služe kao navigacijska pomoć u moru standardne normalne distribucije, olakšavajući standardizirane usporedbe i izračune vjerojatnosti. Ovo poglavlje pojašnjava zamršenost z-standardiziranih vrijednosti, osnažujući vas da tumačite standardizirane rezultate i koristite Z-tablice za statističku analizu. Uz vještinu o z-standardiziranim vrijednostima, kretat ćete se ogromnim prostranstvom normalne distribucije s povjerenjem i preciznošću.

U situacijama kada su veličine uzorka male ili su standardne devijacije populacije nepoznate, t-rezultati i t-tablice pojavljuju se kao nezamjenjivi alati za statističku analizu. Ovo poglavlje razotkriva misterije t-rezultata, vodeći vas kroz njihov izračun i tumačenje pomoću t-tablica. Naoružani ovim znanjem, lako će te se snalaziti u nijansama t-distribucija, osiguravajući zaključivanje u različitim statističkim scenarijima.

Normalna i t-distribucija predstavljaju stupove teorije vjerojatnosti, a svaka posjeduje jedinstvene karakteristike i primjene. U ovom poglavlju razjašnjavamo razlike između ovih distribucija, omogućujući vam da razlučite kada svaku od njih upotrijebiti u statističkoj analizi. Kroz praktične primjere i komparativne analize, razvit ćete razumijevanje normalne i t-distribucije, obogaćujući svoj skup statističkih alata.

Intervali pouzdanosti pružaju uvid u neizvjesnost oko parametara populacije, omogućujući nam da kvantificiramo preciznost naših procjena. U ovom poglavlju istražujemo konstrukciju intervala pouzdanosti za srednje vrijednosti i proporcija, razotkrivajući metodologiju i tumačenje iza ovih bitnih statističkih alata. Savladavanjem intervala pouzdanosti, transparentno i kritički ćete prenijeti neizvjesnost koja je svojstvena vašim nalazima.

Dok p-vrijednosti nude pristup statističkim zaključivanjima, njihovo pogrešno tumačenje može dovesti do pogrešnih zaključaka i pogrešno informiranih odluka. Ovo poglavlje ispituje

potencijalne zamke pretjeranog oslanjanja na p-vrijednosti, naglašavajući važnost konteksta i veličine učinka u statističkoj analizi. Kroz kritičko ispitivanje i praktične uvide, pažljivo ćete se kretati kroz složenost p-vrijednosti, osiguravajući integritet svojih statističkih zaključaka.

Unutar ovih stranica leže ključevi za otključavanje misterija statističke analize, što vam omogućuje da pouzdano i precizno upravljate složenošću podataka. Dok zajedno krećemo na ovo putovanje, neka nam znatiželja bude kompas, a istraživanje naše svjetlo vodilja, osvjetljavajući put prema dubljem razumijevanju i djelotvornim uvidima.

2.1 Normalna distribucija

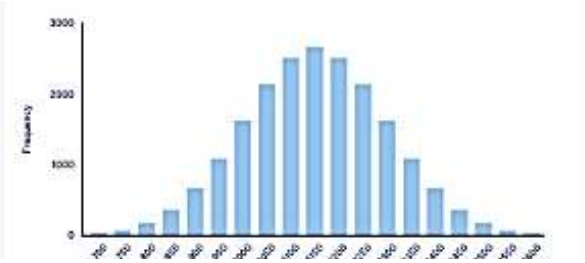
U središtu statističke analize nalazi se normalna distribucija, sveprisutna distribucija vjerojatnosti koja služi kao mjerilo za mnoge statističke tehnike. Udubit ćemo se u njezine karakteristike, njezinu simetričnu krivulju u obliku zvona i značaj u razumijevanju distribucije podataka.

Normalna distribucija nalazi primjenu u raznim područjima, uključujući financije, psihologiju, inženjerstvo i biologiju. Od modeliranja cijena dionica do razumijevanja distribucije ljudske visine, normalna distribucija služi kao svestran alat za analizu i tumačenje podataka.

Kroz ovo poglavlje zadubit ćemo se u matematička svojstva normalne distribucije, istražujući kako izračunati vjerojatnosti, percentile i z-središnje vrijednosti. Raspravljat ćemo o praktičnim tehnikama za vizualizaciju i interpretaciju normalnih distribucija pomoću histograma, dijagrama gustoće i funkcija kumulativne distribucije.

Do kraja ovog poglavlja duboko ćete cijeniti normalnu distribuciju i njen značaj u statističkoj analizi. Bit ćete dobro opremljeni za rješavanje naprednijih statističkih koncepata i njihovu primjenu na skupove podataka u stvarnom svijetu. Krenimo na ovo putovanje kako bismo zajedno razotkrili misterije normalne distribucije.

Normalna distribucija, također poznata kao Gaussova distribucija ili zvonolika krivulja, pokazuje simetričnu distribuciju podataka bez asimetrije. Kada su grafički prikazani, podaci tvore krivulju u obliku zvona, s većinom vrijednosti koje se skupljaju oko središta i smanjuju kako se udaljavaju od njega.

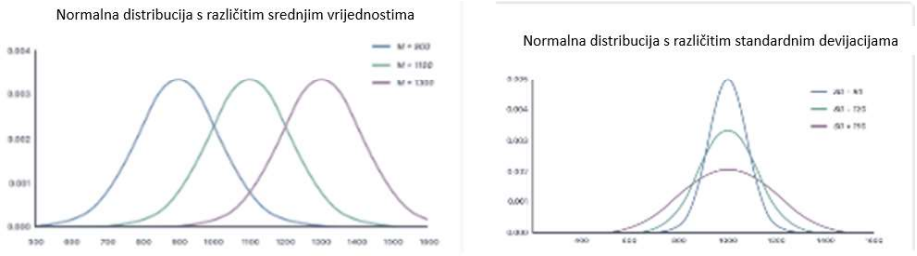


Slika 2.1 Primjer Gaussove distribucije ili zvonolike krivulje

Različite varijable u prirodnim i društvenim znanostima obično pokazuju normalnu distribuciju ili su joj blizu. Primjeri uključuju visinu, porođajnu težinu, sposobnost čitanja, zadovoljstvo poslom i SAT rezultate. Zbog učestalosti normalno raspodijeljenih varijabli, brojni statistički testovi prilagođeni su takvim populacijama. Vještina u razumijevanju karakteristika normalne distribucije osnažuje pojedince da koriste inferencijalnu statistiku za usporedbu grupa i generiranje procjena populacije iz uzoraka.

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

- Srednja vrijednost, medijan i mod su potpuno isti.
- Distribucija je simetrična u odnosu na srednju vrijednost - polovica vrijednosti nalazi se ispod, a polovica iznad srednje vrijednosti.
- Distribucija se može opisati s dvije vrijednosti: srednjom vrijednošću i standardnom devijacijom.



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

Srednja vrijednost služi kao lokacijski parametar koji diktira središte vrha krivulje. Podešavanje srednje vrijednosti pomiče krivulju u skladu s tim: povećanje pomiče krivulju udesno, dok smanjenje pomiče krivulju ulijevo. U međuvremenu, standardna devijacija funkcionira kao parametar razmjera, utječući na širenje ili širinu krivulje.

Standardna devijacija rasteže ili stišće krivulju. Mala standardna devijacija rezultira uskom krivuljom, dok velika standardna devijacija dovodi do široke krivulje.

2.2 Empirijsko pravilo

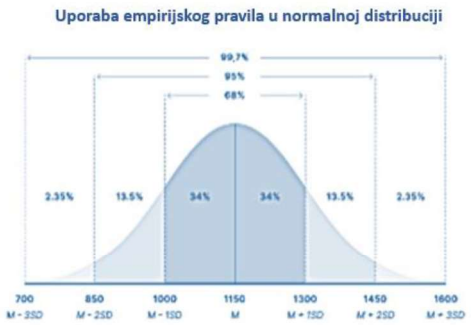
Empirijsko pravilo, također poznato kao pravilo 68-95-99.7, daje uvid u distribuciju vrijednosti unutar normalne distribucije:

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

Na primjer, razmotrite scenarij u kojem se prikupljaju rezultati SAT-a od učenika u novom tečaju pripreme za ispit, a podaci su u skladu s normalnom distribucijom sa srednjom ocjenom (M) od 1150 i standardnom devijacijom (SD) od 150.

Primjena empirijskog pravila daje sljedeće uvide:

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



Slika 2.54 Empirijsko pravilo u normalnoj distribuciji.

Empirijsko pravilo nudi brzu metodu za procjenu podataka, omogućujući otkrivanje outliera ili netipičnih vrijednosti 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. Identificiranje distribucije varijable omogućuje primjenu relevantnih statističkih testova.

2.3 Formula normalne krivulje

Za konstruiranje normalne krivulje na temelju dane srednje vrijednosti i standardne devijacije, može se upotrijebiti funkcija gustoće vjerojatnosti, čime se točno predstavlja distribucija podataka.

Slika 2.74 Normalna krivulja prilagođena podacima SAT rezultata.

Unutar funkcije gustoće vjerojatnosti, područje ispod krivulje predstavlja vjerojatnost. S obzirom da normalna distribucija služi kao distribucija vjerojatnosti, kumulativna površina ispod krivulje uvijek iznosi 1 ili 100%. Iako se formula za normalnu funkciju gustoće vjerojatnosti može činiti zamršenom, njeno korištenje samo zahtijeva poznavanje srednje vrijednosti populacije i standardne devijacije. Zamjenom ovih parametara u formulu, može se odrediti gustoća vjerojatnosti povezana s bilo kojom danom vrijednošću x.

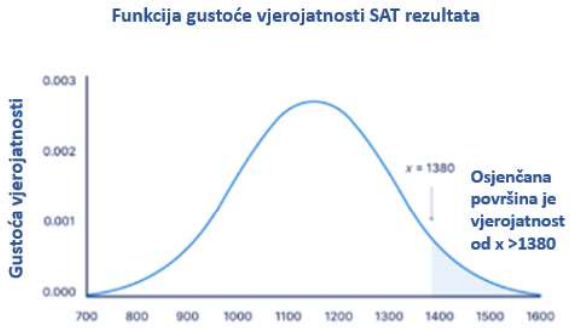
- $f(x)$ = vjerojatnost
- x = vrijednost varijable
- μ = srednja vrijednost
- σ = standardna devijacija
- σ^2 = varijanca

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

Primjer:

Koristeći funkciju gustoće vjerojatnosti, želite znati vjerojatnost da SAT rezultati u vašem uzorku premašuju 1380.

Na vašem grafikonu funkcije gustoće vjerojatnosti, vjerojatnost je osjenčano područje ispod krivulje koja se nalazi desno od mjesta gdje je vaš SAT rezultat jednak 1380.



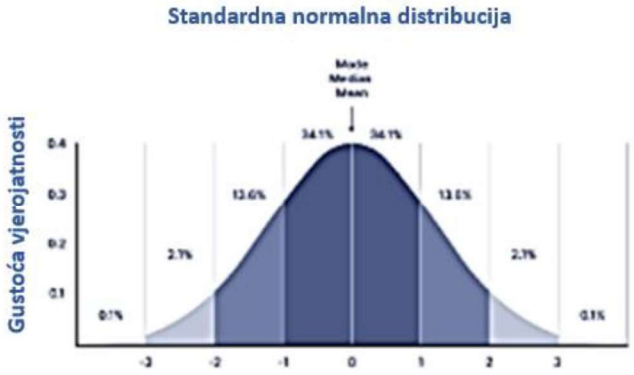
Slika 2.100 Grafikon funkcije gustoće vjerojatnosti SAT rezultata.

Vrijednost vjerojatnosti 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 vrijednost od 0 i standardnu devijaciju od 1. Svaka normalna distribucija može se promatrati kao transformacija standardne normalne distribucije, koja prolazi kroz prilagodbe u mjerilu, položaju ili oba.

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



Slika 2.120 Grafikon standardne normalne distribucije.

Trebate znati samo srednju vrijednost i standardnu devijaciju vaše distribucije da biste pronašli z-skor vrijednosti.

Objašnjenje formule z-skora

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

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

Normalne distribucije pretvaramo u standardnu normalnu distribuciju iz nekoliko razloga:

- kako bismo pronašli vjerojatnost opažanja u distribuciji koja pada iznad ili ispod zadane vrijednosti;
- kako bismo pronašli vjerojatnost da se srednja vrijednost uzorka značajno razlikuje od poznate srednje vrijednosti populacije.
- za usporedbu rezultata na različitim distribucijama s različitim srednjim vrijednostima i standardnim odstupanjima.

2.5 Određivanje vjerojatnosti korištenjem z-distribucije

Svaki z-rezultat odgovara vjerojatnosti, koja se često naziva p-vrijednost, koja ukazuje na vjerojatnost opažanja vrijednosti ispod tog specifičnog z-skora. Transformacijom pojedinačne

vrijednosti u z-skor, može se odrediti vjerojatnost da se sve vrijednosti do te točke pojave unutar normalne distribucije.

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

Izračun formule

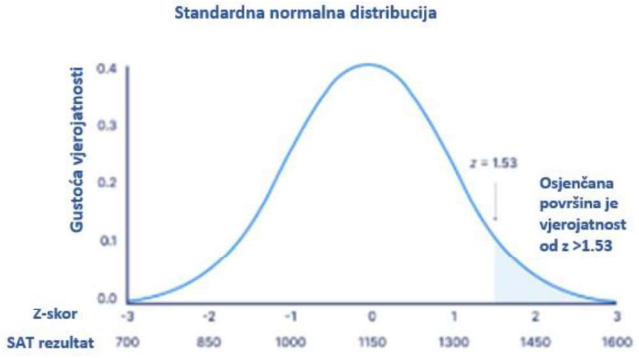
$$z = \frac{x - \mu}{\sigma} = \frac{1380 - 1150}{150} = 1.53$$

Za z -skor od 1,53, p -vrijednost je 0,937. Ovo je vjerojatnost da će SAT rezultati biti 1380 ili manje (93,7%), a to je područje ispod krivulje lijevo od osjenčanog područja.

Da biste pronašli osjenčano područje, oduzmite 0,937 od 1, što je ukupna površina ispod krivulje.

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

To znači da je vjerojatno 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, omogućujući nam izvođenje zaključaka o populacijama na temelju podataka uzorka. Udubit ćemo se u zamršenost

sampling-distribucija, razumijevajući kako odražavaju varijabilnost statistike uzorka i njihovu ključnu ulogu u testiranju hipoteza.

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

2.7 Centralni granični teorem i sampling-distribucija

Centralni granični teorem (engl. Central Limit Theorem - CLT) temeljni je koncept u statistici koji podupire ponašanje sampling-distribucije. Navodi se da se sampling-distribucija srednje vrijednosti uzorka približava normalnoj distribuciji kako se veličina uzorka povećava, bez obzira na oblik distribucije populacije. Ovaj teorem nam omogućuje da napravimo čvrste zaključke o parametrima populacije iz uzorka podataka.

Središnji granični teorem služi kao kamen temeljac razumijevanja normalnih distribucija u statistici. U uvjetima istraživanja, dobivanje točne procjene srednje vrijednosti populacije često uključuje prikupljanje podataka iz brojnih nasumičnih uzoraka unutar populacije. Te pojedinačne srednje vrijednosti uzoraka zajedno tvore ono što je poznato kao sampling-distribucija srednje vrijednosti.

Centralni granični teorem ocrtava dva ključna principa:

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

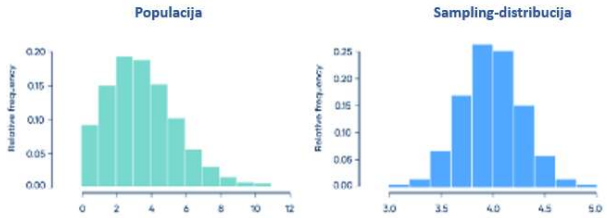
Parametarski statistički testovi konvencionalno pretpostavljaju da su uzorci izvedeni iz normalno distribuiranih populacija. Međutim, centralni granični teorem uklanja nužnost ove pretpostavke za dovoljno velike uzorke. S velikim uzorcima, parametarski testovi mogu se primijeniti bez obzira na distribuciju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

Nasuprot tome, za male uzorke, osiguravanje pretpostavke normalnosti je ključno zbog nesigurnosti koja okružuje sampling-distribuciju srednje vrijednosti. Točni rezultati zahtijevaju

potvrdu da se populacija pridržava normalne distribucije prije korištenja parametarskih testova s malim uzorcima.

Ilustrativno, centralni granični teorem tvrdi da će dobivanjem dovoljno velikih uzoraka iz populacije, srednje vrijednosti tih uzoraka pokazati normalnu distribuciju, čak i ako temeljna distribucija populacije odstupa od normalnosti.

Primjer: Razmotrite populaciju prema Poissonovoj distribuciji (prikazano na lijevoj slici). Nakon izvlačenja 10 000 uzoraka iz ove populacije, od kojih se svaki sastoji od 50 opažanja, distribucija srednjih vrijednosti uzorka blisko je usklađena s normalnom distribucijom, u skladu s centralnim graničnim teoremom (kao što je ilustrirano na desnoj slici).



Slika 2.155 Primjer populacije u Poissonovoj distribuciji i normalnoj distribuciji.

Centralni granični teorem ovisi o pojmu sampling-distribucije, koja predstavlja distribuciju vjerojatnosti statistike izračunate iz brojnih uzoraka izvučenih iz populacije.

Konceptualizacija eksperimenta može pomoći u shvaćanju sampling-distribucije:

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

Združivanje ovih srednjih vrijednosti uzoraka predstavlja primjer sampling-distribucije. Prema centralnom graničnom teoremu, sampling-distribucija srednje vrijednosti teži prema 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 vrijednosti pokazuje normalnost.

Srećom, ne treba opetovano uzorkovati populaciju da bi se razaznao oblik sampling-distribucije. Umjesto toga, parametri sampling-distribucije srednje vrijednosti ovise o parametrima same populacije.

- Srednja vrijednost sampling-distribucije je srednja vrijednost populacije.

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

- Standardna devijacija sampling-distribucije je standardna devijacija populacije podijeljena s kvadratnim korijenom veličine uzorka.

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

Sampling-distribuciju srednje vrijednosti možemo opisati pomoću ove oznake:

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

gdje:

- $\bar{X}$ je sampling-distribucija srednjih vrijednosti uzorka
- ~ znači "slijedi distribuciju"
- *N* je normalna distribucija
- μ je srednja vrijednost 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 utječe na sampling-distribuciju srednje vrijednosti 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 vrijednosti može odstupati od normalnosti. Ovo odstupanje nastaje jer valjanost centralnog graničnog teorema ovisi o "dovoljno velikoj" veličini uzorka.
- Uobičajeno, veličina uzorka od 30 ili više smatra se "dovoljno velikim".
- Kada je n<30, centralni granični teorem se ne primjenjuje, a sampling-distribucija odražava distribuciju populacije. Stoga je sampling-distribucija normalna samo ako je distribucija populacije normalna.
- Nasuprot tome, kada je n≥30, centralni granični teorem vrijedi, a sampling-distribucija približava se normalnoj distribuciji.

2. Veličina uzorka i standardna devijacija:

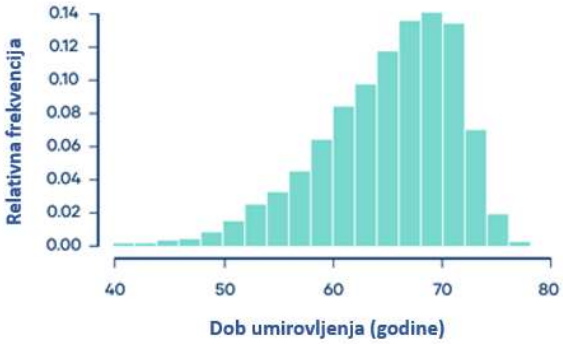
- Veličina uzorka izravno utječe na standardnu devijaciju sampling-distribucije, odražavajući varijabilnost ili raspršenost distribucije.
- S manjim uzorcima, standardna devijacija obično je viša, što ukazuje na veću varijabilnost među srednjim vrijednostima uzorka zbog njihove neprecizne procjene srednje vrijednosti populacije.
- Suprotno tome, veći uzorci odgovaraju nižim standardnim devijacijama, što ukazuje na manju varijabilnost među srednjim vrijednostima uzorka zahvaljujući njihovoj točnijoj procjeni srednje vrijednosti populacije.

Važnost centralnog graničnog teorema:

Parametarski testovi kao što su t-testovi, ANOVA i linearna regresija imaju veću statističku snagu u usporedbi s većinom neparametarskih testova. Ova povećana statistička snaga proizlazi iz pretpostavki o distribuciji populacija, koje su utemeljene na centralnom graničnom teoremu.

Kontinuirana distribucija

Razmotrimo dob za odlazak u mirovinu pojedinaca u Sjedinjenim Američkim Državama. Stanovništvo se sastoji od svih umirovljenih Amerikanaca, a distribucija ovog stanovništva može se predstaviti na sljedeći način:



Slika 2.182 Grafikon kontinuirane

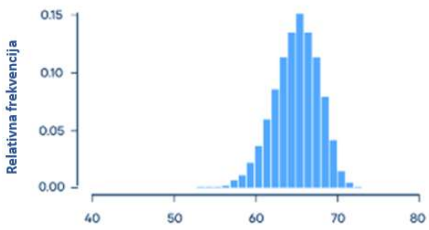
Distribucija dobi za umirovljenje iskrivljena je ulijevo, pri čemu većina odlazi u mirovinu unutar približno pet godina od prosječne dobi za umirovljenje od 65 godina. Međutim, postoji prošireni rep pojedinaca koji odlaze u mirovinu puno ranije, primjerice s 50 ili čak 40 godina. Populacija pokazuje standardnu devijaciju od 6 godina.

Zamislite provođenje malog uzorkovanja ove populacije. Nasumično se odabire pet umirovljenika i bilježi se njihova dob za odlazak u mirovinu. Na primjer: 68, 73, 70, 62, 63.

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

Sada pretpostavimo da se ovaj proces uzorkovanja ponovi 10 puta, a svaki uzorak uključuje pet umirovljenika. Izračunava se srednja vrijednost svakog uzorka, što rezultira distribucijom poznatom kao sampling-distribucija srednje vrijednosti. Na primjer: 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 vrijednosti ovih uzoraka približno će odgovarati normalnoj distribuciji.



Slika 2.208 Normalna distribucija srednjih vrijednosti

Unatoč tome što sampling-distribucija pokazuje nešto normalniji oblik u usporedbi s populacijom, još uvijek zadržava blagi zaokret ulijevo. Osim toga, evidentno je da je varijabilnost u sampling-distribuciji uža od one populacije.

Prema centralnom graničnom teoremu, sampling-distribucija srednje vrijednosti nastoji se približiti normalnoj distribuciji kako se veličina uzorka povećava. Međutim, trenutna sampling-distribucija srednje vrijednosti odstupa od normalne zbog relativno male veličine uzorka.

2.8 Testna statistika

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

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

Ali što točno čini testnu statistiku?

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

Testna statistika sažima vaše opažene podatke u jedinstvenu brojku, koristeći mjere 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), podijeljenih s varijancom podataka (tj. standardnom devijacijom).

Razmotrite ovaj primjer:

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

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

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

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

2.9 Vrste testne statistike

U nastavku je prikazan sinopsis prevladavajućih 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, temeljne hipoteze i tumačenja testne statistike ostaju dosljedni.

Testna statistika	Nulta i alternativna hipoteza	Statistički testovi
***t* vrijednost**	**Nulta:** Srednje vrijednosti dviju grupa su jednake. **Alternativna:** Srednje vrijednosti dviju skupina nisu jednake.	• *T* test • Regresijski testovi
***z* vrijednost**	**Nulta:** Srednje vrijednosti dviju grupa su jednake. **Alternativna:** Srednje vrijednosti dviju skupina nisu jednake.	• *Z* test
***F* vrijednost**	**Nulta:** Varijacija između dvije ili više grupa veća je ili jednaka varijaciji između grupa. **Alternativna:** Varijacije između dvije ili više grupa su manje od varijacija između grupa.	• ANOVA • ANCOVA • MANOVA
***X*²-vrijednost**	**Nulta:** Dva su uzorka neovisna. **Alternativna:** Dva uzorka nisu neovisna (tj. koreliraju).	• Hi-kvadrat test • Neparametarski korelacijski testovi

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

Na primjer, u testiranju vaše hipoteze o temperaturi i datumima cvjetanja, provodite regresijsku analizu. Regresijski test daje: • regresijski koeficijent od 0,36 • t-vrijednost koja uspoređuje ovaj koeficijent s očekivanim rasponom regresijskih koeficijenata pod nultom hipotezom nepostojanja veze.

Rezultirajuća t-vrijednost iz regresijskog testa od 2,36 predstavlja vašu testnu statistiku.

2.10 Standardna pogreška

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

Dok je standardna pogreška srednje vrijednosti najčešće citirani oblik standardne pogreške, slične mjere postoje za druge statističke parametre kao što su medijan ili proporcije. Standardna pogreška funkcionira kao prevladavajuća mjera pogreške uzorkovanja, prikazujući nejednakost između parametra populacije i statistike uzorka.

Kako bi se ublažila standardna pogreška, preporučuje se povećanje veličine uzorka. Korištenje velikog, nasumičnog uzorka služi kao najučinkovitija strategija za smanjenje pristranosti uzorkovanja i povećanje pouzdanosti nalaza.

Standardna pogreška i standardna devijacija mjere su varijabilnosti:

- **Standardna devijacija** opisuje varijabilnost **unutar jednog uzorka**.
- **Standardna pogreška** procjenjuje varijabilnost **u višestrukim uzorcima** populacije.

Standardna devijacija služi kao deskriptivna statistika izvedena izravno iz podataka uzorka, dok standardna pogreška predstavlja inferencijalnu statistiku, obično procijenjenu, osim ako nije poznat točan parametar populacije.

2.11 Formula standardne pogreške

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

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

Kada su poznati parametri populacije

Kada je poznata standardna devijacija populacije, možete je koristiti u donjoj formuli za točan izračun standardne pogreške.

Formula	Obrazloženje
$SE = \frac{\sigma}{\sqrt{n}}$	• *SE* je standardna pogreška • σ je standardna devijacija populacije • *n* je broj elemenata u uzorku

Kada su parametri populacije nepoznati

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

Primjer: Korištenje formule standardne pogreške za procjenu standardne pogreške za rezultate SAT-a iz matematike. Slijedite sljedeća dva koraka.

Najprije pronađite kvadratni korijen veličine uzorka (*n*).

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

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

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

Standardna pogreška rezultata SAT iz matematike je 12,8.

Možete predstaviti standardnu pogrešku uz srednju vrijednost ili je uključiti u interval pouzdanosti kako biste prenijeli nesigurnost koja okružuje srednju vrijednost.

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

Izvještavanje o standardnoj pogrešci unutar intervala pouzdanosti je poželjno jer eliminira potrebu čitatelja za izvođenje dodatnih izračuna kako bi dobili smisleni raspon.

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

Na razini pouzdanosti od 95%, očekuje se da će 95% svih srednjih vrijednosti uzorka pasti unutar intervala pouzdanosti koji obuhvaća ± 1,96 standardnih pogrešaka srednje vrijednosti uzorka. Ovaj interval služi kao procjena unutar koje se vjeruje da se stvarni parametar populacije nalazi unutar 95% pouzdanosti.

Na primjer: Konstruiranje intervala pouzdanosti od 95% Vi konstruirate interval pouzdanosti od 95% (CI) da biste procijenili srednju vrijednost matematičke SAT ocjene populacije. S obzirom na normalno raspodijeljenu karakteristiku kao što su SAT rezultati, otprilike 95% svih srednjih vrijednosti uzorka pada unutar približno 4 standardne pogreške srednje vrijednosti uzorka.

Formula intervala pouzdanosti

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

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

SE = standardna pogreška = 12,8

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

S nasumičnim uzorkovanjem, 95% CI [525 575] govori vam da postoji vjerojatnost od 0,95 da je srednja vrijednost matematičkog SAT rezultata populacije između 525 i 575.

Literatura 2. poglavlja

- *Introductory Statistics.* Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l
- 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