STATISTIKA

KONCEPTI : POPULACIJA i UZORAK

Primjer: svi glasači, samo neki glasači

populacija uključuje sve podatke, a uzorak samo dio, slučajno izabranih

kako procjeniti reprezentativni element?

MJERE CENT. TENDENCIJE

SREDNJA VRIJEDNOST

MOD

MEDIAN

$niz \ \{x_1, x_2, x_3, ..., x_T\} = (x_1 + x_2 + ...x_T)/n$

srednja vrijednost

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

PRIMJER

$2, 4, 4, 5, 3, 5, 2, 4$

sortirano $2, 2, 3, 4, 4, 4, 5, 5$

srednja vrijednost

$\overline{x} = (2 + 4 + 4 + 5 + 3 + 5 + 2 + 4)/8 = 3.625$

mod, podatak koji se najčešće ponavlja

4

median

4

P PERCENTIL

P percentil je vrijednost ispod koje se nalazi P % svih elemenata grupe

Pozicija P percentila je dana kao (n+1) P / 100

naš primjer

50 percentil je na poziciji 4.5 — 4

KVARTIL

podjela podataka na četvrtine

donji, srednji = median , gornji

MJERE VARIJABILNOSTI

$3, 3, 3, 3, 3, 3, 3$

$0, 1, 2, 3, 4, 5, 6$

VARIJANCA I STANDARDNA DEVIJACIJA

VARIJANCA:prosječni kvadrat odstupanja

VARIJANCA:UZORAK

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{2}$$

VARIJANCA:POPULACIJA

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N} \tag{3}$$

μ :srednja vrijednost populacije

STANDARDNA DEVIJACIJA

UZORAK

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{4}$$

GRUPIRANJE PODATAKA:HISTOGRAM

KLASA:skup podataka unutar nekih granica

distribucija frekvencija: histogram

relativne i apsolutne frekvencije

SKEWNESS I KURTOSIS

skewness. mjera asimetrije podataka

kurtosis. mjera spljoštenosti

METODE PRIKAZIVANJA PODATAKA

PIE CHART: postotak površine odredjuje postotak varijable

POLIGON RELATIVNIH FREKVENCIJA, započinje i završava nulom kumulativne relativne frekvencije

STANDARDIZIRANA NORMALNA DISTRIBUCIJA

Neka je X iz $N(\mu, \sigma)$. Tada se može pokazati da je varijabla

$$Z = \frac{X - \mu}{\sigma} \tag{5}$$

iz distribucije $N(0, 1)$. Koja je korist ove transformacije? ZADATAK: Neka je srednja vrijednost visina medju studentima $\mu = 64.3$ incha i $\sigma^2 = 9.8$. Kolika je vjerojatnost medju studentima da je visina veća od 69 incha?

Bez standardizirane normalne distribucije morali bi integrirati numerički normalnu distribuciju. Medjutim vrijedi, $Pr(X > 69) = Pr(\frac{X-64.3}{3.13} > \frac{69-64.3}{3.13})$. Vrijedi, $Pr(Z > 1.5) = 0.5 - A = 0.5 - 0.4332$.

Poznavanjem uzorka dobivamo ocjenu srednje vrijednosti populacije medjutim kako vrijedi varijabilnost uzoraka (neki drugi dao bi drugačiju vrijednost) zanima nas i pouzdanost ocjene. Kako je srednja vrijednost iz distribucije $N(\mu, \sigma/\sqrt{n})$, iz nje definirajmo standardiziranu varijablu $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Iz tablica dobivamo $Pr(-1.96 < Z < 1.96) = 0.95$. $Pr(\overline{X} + 1.96\frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$.

Dobili smo interval za koji sa 95% sigurnošću sadrži srednju vrijednost populacije.

STATISTIČKO ZAKLJUČIVANJE

kako na osnovu uzorka zaključiti o populaciji

kolike su srednja vrijednost i varijanca populacije?

PRIMJER:

Slučajni uzorak 40 odvjetničkih ureda pokazao je da su pravne usluge po satu dane sa srednjom vrijednošću 25 dolara a sa stdev $s = 3.7$ dolara. Nadjite s 95 postotni interval pouzdanosti za cijelu pravnu profesiju.

Neka je zadana slučajna varijabla X dana normalnom distribucijom $N(\mu, \sigma)$. Uzmimo uzorak i izračunajmo srednju vrijednost. Jasno, različiti uzorci daju različite srednje vrijednosti, dakle očita je uzoračka varijabilnost. U praksi ne znamo koliki su populacijski parametri. Njih iz uzorka želimo procijeniti. Uzmimo veliki broj uzoraka i za svaki izračunajmo srednju vrijednost. Jasno iz tih vrijednosti izgradimo distribuciju srednjih vrijednosti. Ako je X iz $N(\mu, \sigma)$ tada je **distribucija srednjih vrijednosti ima srednju vrijednost μ i standardnu devijaciju $\sigma/\sqrt{n}$.** Veći uzorak, bolja procijena populacijske srednje vrijednosti.

STUPNJEVI SLOBODE

Neka su zadane n varijable i f restrikcija (relacija, jednjadžbi) medju njima. Tada kažemo da ima $n - f$ stupnjeva slobode. Npr, neka su zadane

varijable X, Y i Z. Neka varijable nisu nezavisne jer zadovoljavaju slijedeću jednadžbu: $X+Y+2Z=5$. Primjetimo da bilo koje dvije varijable možemo slobodno mijenjati, ali treću ne možemo jer vrijedi ranija jednadžba. Koliko varijabli možemo slobodno mijenjati toliko je i stupnjeva slobode (degrees of freedom d.f.).

SVRHA STUDENTOVE T- i CHI2- DISTRIBUCIJE

Imaš uzorak, a nemaš populaciju. Iz uzorka želimo procijeniti srednju vrijednost μ ili standardnu devijaciju σ populacije. Za procijenu μ koristimo t distribuciju, a za procijenu σ koristimo χ^2 distribuciju.

JEDAN REP DVA REPA DISTRIBUCIJE

STUDENTOVA T-DISTRIBUCIJA

U praksi najčešće ne poznamo parametre populacije, ni srednju vrijednost ni varijancu. Jedini čime raspolažemo je uzorak. **Ova distribucija pomaže nam da procijenimo populacijsku srednju vrijednost odredjivanjem srednje vrijednosti i standardne devijacije uzorka.** Ako je populacija normalno distribuirana, standardizirana statistika

$$t=\frac{\overline{X}-\mu}{S/\sqrt{n}} \tag{6}$$

slijedi $t-$distribuciju s $n-1$ stupnjeva slobode $(d.f.)$. Stupnjevi slobode odgovaraju srupnjevima povezanima sa standardnom devijacijom uzorka.

Svaka t distribucija u tablicama dana je preko d.f. (broj po-

dataka -1)

U središnjem dijelu, t distribucija nalikuje standardiziranoj normalnoj distribuciji. Obje su simetrične (skewness nula). Ipak, od potonje se razlikuje u repovima (dalje od središnjeg dijela). Što je d.f. veći to t-distribucija više nalikuje normalnoj distribuciji.

Primjena distribucije: Ma koju statističku varijablu imali (visina ljudi, rok trajanja auto-guma, ...) možemo je standardizirati na raniji način i stoga koristiti tablice te standardizirane varijable. Sad želimo odrediti poznavanjem srednje vrijednosti uzorka gdje se nalazi srednja vrijednost populacije. Poznavanjem $\overline{X}$ i S ne možmo točno odrediti (jednim brojem) μ, ali možemo odrediti s nekom vjerojatnošću interval u kojem se nalazi μ populacije. Veća vjerojatnost, veći interval a time i naša nesigurnost u odredjivanju položaja μ. Dakle, unaprijed moramo odrediti koliko želimo biti nesigurni. Pretpostavimo da želimo 95% interval pouzdanosti za odredjivanje μ primjenom t distribucije. U tablicama uočimo iz zadnjeg reda kako područje izmedju $t = -1.96$ i $t = 1.96$ pokriva 0.95% vjerojatnosti. Koristimo distribuciju s 2 repa jer se μ može nalaziti i lijevo i desno od $\overline{X}$. Kako se zadnji red odnosi na $d.f. = \infty$, u tablici putujemo prema gore dok ne nadjemo red s našim stupnjem slobode. Ako je npr d.f. $= 20$ tada je $t_{0.025} = 2.086$.

Dakle 95% interval pouzdanosti za traženje μ iznosi

$$\overline{X} \pm t_{0.025}\frac{s}{\sqrt{n}} \tag{7}$$

PRIMJER:

1) Proizvodjač guma želi ocijeniti prosječan broj milja koji predju neke gume prije no što se potroše. Slučajan uzorak od 32 gume (u tisućama milja) dao je vrijednosti:
32, 33, 28, 37, 29, 30, 25, 27, 39, 40, 26, 26, 27, 30, 25, 30, 31, 29, 24, 36, 25, 37, 37, 20, 22, 35, 23, 28, 30, 36, 40, 41
Nadjite 99% interval za prosječan broj milja gume.

2) U svrhu odredjivanja potrošačkih navika turista, državna agencija želi ocijeniti koliko u prosjeku turisti troše na nekom području. Slučajan uzorak od 56 turista daje 258 dolara i $s = 85$ dolara. Nadjite 95 postotni interval pouzdanosti za prosječni potrošeni novac na danom području.

3) Visa daje kartičarima poseban bonus u dolarima koji se mogu upotrijebiti pri kupnji poklona. Kompanija želi ocijeniti koliki je prosječan iznos plačen na taj način po kartičaru. Uzorak ima 225 člana. Pokazalo se da je za uzorak $\overline{x} = 259.60$ $s = 52$. Dajte 95 % interval pouzdanosti za prosječan iznos bonus dolara.

χ^2 DISTRIBUCIJA

Problem u kojem se primjenjuje:
Kompanija planira kupiti žarulje pri čemu je nužno da žarulje budu uniformne (što sličnije u svojstvima). Iz iskustva se zna kako varijanca života

žarulja ne smije prelaziti $300 sati^2$. Slučajni uzorak 25 žarulja pokazao je kako je varijanca $340 sati^2$. Pokažite uz nivo značajnosti 0.05 da li kupljene žarulje zadovoljavaju dane zahtjeve.

Neka je zadano n neovisno distribuiranih varijabli iz normalne distribucije $N(0,1)$. Tada je suma kvadrata

$$\Sigma_i^n Z_i^2 = Z_1^2 + Z_2^2 Z_3^2 + ... + Z_n^2 \tag{8}$$

χ_n^2 distribucija s n stupnjeva slobode.

Varijablu χ_n^2 dobivamo sumiranjem pozitivnih realnih brojeva pa je distribucija zadana samo za pozitivne realne brojeve. Kao i za svaku distribuciju, površina ispod krivulje distribucije je jednaka 1. S porastom n vrijednosti, maksimum distribucije pomjera se udesno (srednja vrijednost χ^2 distribucije jednaka je n).

Pretpostavimo da imamo uzorak od n mjerenja: X_1, X_2, ..., X_n pri čemu su vrijednosti varijabli dobivene iz $N(\mu, \sigma)$ distribucije. Tada slijedi da je varijabla

$$\frac{X_i - \mu}{\sigma} \tag{9}$$

iz standardizirane normalne distribucije $N(0,1)$. Iz definicije varijance uzorka dobivamo:

$$\frac{s^2(n-1)}{\sigma^2} = \frac{\Sigma_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \tag{10}$$

Očito je desna strana definicija χ_n^2 s $n-1$ d.f.

Vrijedi da je srednja vrijednost χ^2 jednaka d.f. a varijanca iznosi 2 d.f.

Iz definicije χ^2 vidimo da ako želimo odrediti recimo 95% interval pouzdanosti pri ocjeni varijance populacije moramo pronaći dvije kritične vrijednosti na nasuprotnim stranama distribucije izmedju kojih je integral 0.95. Dakle, iz tablica se nadje $\chi_{0.025}$ i $\chi_{0.975}$. Vrijedi

$$\chi^2_{0.975} < \chi^2 < \chi^2_{0.025} \tag{11}$$

iz čega slijedi da je σ^2 iz intervala

$$[\frac{(n-1)s^2}{\chi^2_{0.025}}, \frac{(n-1)s^2}{\chi^2_{0.975}}] \tag{12}$$

Riješeni primjer:

1) Pokazalo se kako je kompanija koristila proizvodni proces u kojem je standardna devijacija jednaka 35 s. Takva varijabilnost izazivala je prekide u proizvodnji. Inženjeri tvrde da su našli novu metodu s manjom standardnom devijacijom. Dvadeset radnika pokazalo je kako je stdev vremena novog procesa $s = 28s$. Možemo li zaključiti da je novi proces uveo poboljšanje?

$$H_0 : \sigma^2 = 1225 \tag{13}$$

$$H_A : \sigma^2 < 1225 \tag{14}$$

$$TS = \frac{s^2(n-1)}{1225} \; je \; \chi^2 \; s \; n-1 \; d.f \tag{15}$$

2) U problemu 1 iz poglavlja s t-distribucijom nadjite 99% interval pouzdanosti za varijancu broja milja koju predje guma prije no što se potroši.

F DISTRIBUCIJA

Distribucija se koristi za usporedjivanje varijanci dviju populacija.
F distribucija je distribucija kvocijenata dviju χ^2 slučajnih neovisnih varijabli, pri čemu je u definiciji svaka χ^2 podijeljena sa svojim stupnjem slobode. Neka je χ_1^2 slučajna varijabla s k_1 stupnja slobode, a χ_2^2 druga slučajna varijabla, neovisna o prvoj, s k_2 stupnja slobode. Tada varijabla

$$F_{k_1,k_2} = \frac{\chi_1^2/k_1}{\chi_2^2/k_2} \tag{16}$$

zadovoljava F distribuciju s k_1 i k_2 stupnja slobode. Dakle, za razliku od Student t distribucije i χ^2 distribucije koje ovise o jednom parametru (stupnju slobode), F distribucija ovisi o 2 stupnja slobode što je jasno jer joj je varijabla kvocijent dviju neovisnih χ^2 varijabli gdje je svaka dana svojim stupnjem slobode. k_1 (k_2) je stupanj slobode χ^2 varijable u brojniku (nazivniku) gornje jednadžbe.

Kao i svaka distribucija vjerojatnosti, **normirana je na 1**, i nas u praksi zanima kako pronaći kritične vrijednosti takve da su integrali **desno** od njih jednaki npr 0.05 (0.01).

F distribuciju koristimo najčešće pri testiranju jednakosti dviju populacijskih varijanci. Prisjetimo se kako je definirana χ^2 slučajna varijabla:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \tag{17}$$

gdje je S^2 uzoračka varijanca dobivena iz normalno distribuirane populacije. Uzorak ima n elemenata pa je broj stupnjeva slobode za jedan manji. Pretpostavimo da imamo dva neovisna slučajna uzorka dobivena iz dviju normalno distribuiranih populacija. Za svaki od dvaju uzoraka izračunamo odgovarajuće varijance S_1^2 (s brojem stupnjeva slobode $n_1 - 1$), odnosno S_2^2 (s brojem stupnjeva slobode $n_2 - 1$). Definirajmo slučajnu varijablu:

$$\frac{S_1^2}{S_2^2} = \frac{\chi_1^2\sigma_1^2/(n_1-1)}{\chi_2^2\sigma_2^2/(n_2-1)} \tag{18}$$

Kad su populacijske varijance σ_1^2 i σ_2^2 jednake, gornja jednadžba postaje jednaka jednadžbi za slučajnu varijablu F distribucije s n_1-1 i n_2-1 stupnjeva slobode.

Logičan test za ispitivanje jednakosti varijanci koje pripadaju dvjema normalno distribuiranim populacijama:

$$F_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2} \tag{19}$$

Test jednakosti dviju populacija

Dva su uzorka slučajno izabrana. Testiramo (i zanima nas pokazati) da li ta 2 uzorka pripadaju različitim ili jednakim populacijama. Moguće hipoteze koje testiramo su slijedeće: Test s 2 repa:

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Test primjenjemo kad želimo ustanoviti da li su populacije različite pri čemu varijanca populacije II može biti bilo manja bilo veća od populacije I.

Test s jednim repom:

$H_0 : \sigma_1^2 \leq \sigma_2^2$

$H_1 : \sigma_1^2 > \sigma_2^2$

Ovaj test koristimo kad želimo testirati da li je σ_1^2 veće od σ_2^2. Jasno u tom slučaju podrazumjeva se i da je $S_1^2 > S_2^2$. Jasno i da je odbacivanje nulte hipoteze moguće samo na desnom repu distribucije gdje su veće vrijednosti. Logično da je nužan uvjet da ćemo prihvatiti hipotezu $\sigma_1^2 > \sigma_2^2$ ako je $S_1^2 > S_2^2$. Medjutim, nije dovoljno da je S_1^2 samo veće od S_2^2. Mora biti i **dovoljno veće**. Koliko veće? To nam otkrivaju tablice F distribucije.

U tablicama su dane kritične vrijednsti za nivoe značajnosti $\alpha = 0.05$ i $\alpha = 0.01$.

Primjer 1:

Standardna devijacija se uzima kao mjera rizika koji pridružujemo cijenama dionica. Financijski analitičar želi testirati hipotezu da dionica A ima veći rizik od dionice B. Slučajan uzorak od 25 dnevnih cijena dionice A daje $s_1^2 = 6.52$, a slučajni uzorak 22 cijene dionice B daje $s_2^2 = 3.47$. Testirajte uz $\alpha = 0.01$. A kakav je zaključak ako uzmemo $\alpha = 0.05$?

Primjer 2:

Testirajmo jednakost dviju populacija. Velika robna kuća želi testirati da li je varijanca vremena čekanja u redu približno jednaka. Dva slučajna uzorka s $n_1 = 14$ i $s_1 = 0.12$ i $n_2 = 9$ i $s_1 = 0.11$ testiraju pretpostavku jednakih populacijskih varijanci.

Iz tablice iščitamo desnu kritičnu vrijednost za distribuciju F_{k_1,k_2} uz nivo značajnosti ili $\alpha = 0.05$ ili $\alpha = 0.01$. Lijeve kritične vrijednosti za distribuciju F_{k_1,k_2} sami izračunamo kao:

$$\frac{1}{F_{k_2,k_1}} \tag{20}$$

Uočimo da u tom slučaju područje izmedju lijeve i desne kritične vrijednosti pokriva $\alpha = 0.10 = 2(0.05)$ jer iammo tabele samo za $\alpha = 0.05$ i $\alpha = 0.01$. Da bi smo mogli odrediti područje F distribucije s dva repa za $\alpha = 0.05$, morali bi koristiti tabele F distibucije za $\alpha = 0.025$. Takve najćešće nemamo.

$F_{13,8} = 1.19$

Test razlika dviju populacijskih srednjih vrijednosti

Želimo recimo procijeniti na osnovu izračunatih srednjih vrijednosti da li dva uzorka pripadaju istoj populaciji ili različitima. Test koristimo za velike uzorke (broj elemenata veći od 30 za oba uzorka). $H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

mogući, ali rjedje korišten test je

$H_0 : \mu_1 - \mu_2 \leq D$

$H_1 : \mu_1 - \mu_2 > D$

Odgovarajuća statistika je:

$$z = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{21}$$

gdje je $(\mu_1 - \mu_2)$ ralika populacijskih srdnjih vrijednosti prema nultoj hipotezi.

PRIMJER:

Analitičar u modnoj industriji želi procijeniti tvdnju da modeli koji nose Peru Kardena (PC) zaradjuju u prosjeku više od modela koji nose Calvin Klein (CK). Za dano vremensko razdoblje 32 PC modela zaradila su u prosjeku $4,238$ dolara sa $s = 1,000.50$ dolara. Slučajan uzorak 37 CK modela zaradio je u prosjeku $3,888$ dolara sa $s = 876.05$ dolara. Što ste zaključili?

JEDNOSTAVNA REGRESIJA (DVIJE VARIJABLE)

Uočavamo ponekad da postoje veze izmedju dviju ekonomskih varijabli. Na primjer, jasno je da obitelji koje više zaradjuju u prosjeku više i troše, a obitelji koje manje zaradjuju manje troše. **U PROSJEKU**, jer uvijek su moguće i iznimke. Zanima nas uspostaviti vezu (matematičku funkciju) medju danim varijablama. Kakva je korist u slučaju zarade i potrošnje? Jasno, različite obitelji s istom zaradom različito troše, ali nas zanima kako bi mogli predvidjati, kolika je očekivana (prosječna) potrošnja za neku obitelj koja ima odredjenu zaradu.

REGRESIJSKA ANALIZA: uspostavlja veze medju promatranim varijablama

Prvi je korak uočiti uzročnu vezu medju varijablama. Jasno da nema potrošnje bez primanja, dakle, zarade odredjuje potrošnju, pa primanja uzimamo kao objasnidbenu (nezavisnu) varijablu X, a potrošnju kao zavisnu Y. Pretpostavimo da imamo sve podatke (primanja i potrošnju) neke zemlje, i da smo našli vezu:

$$E(Y) = \alpha + \beta X \tag{22}$$

gdje su α i β populacijski parametri. Uočimo da smo pretpostavili **linearnu** vezu izmedju X i Y, ali nismo rekli kako smo došli do te linearne veze. O tom poslije. Parametri populacije α i β su u praksi nepoznati jer nikad nemamo informaciju o cijeloj populaciji. Naime ta je informacija vrlo vrlo skupa.

$E(Y)$ se čita očekivana potrošnja, jer kao što smo rekli, obitelji istih primanja X nemaju istu potrošnju. Zato $E(Y)$ predstavlja prosječnu potrošnju za primanja X. Gornju jednadžbu nazivamo **populacijska regresijska jednadžba**.

Jasno, stvarnja potrošnja neke obitelji primanja X razlikuje se od prosječne (očekivane) za obitelji tih primanja. Razlika je dana kao

$$Y = E(Y) + \epsilon \tag{23}$$

odnosno

$$Y = \alpha + \beta X + \epsilon \tag{24}$$

gdje se ϵ u statistici zove smetnja, a predstavlja razliku stvarne i prosječne potrošnje za obitelj danih primanja. Kako u praksi uvijek imamo konačno mnogo podataka (informacija) za dani slučaj recimo raspolažemo s n obitelji s dakle n različitih primanja i potrošnji. Stoga X_i i Y_i dobivaju indeks. X_1 su primanja prve obitelji a Y_8 je potrošnja osme obitelji. Ranija jednadžba sada glasi:

$$Y_i = \alpha + \beta X_i + \epsilon \tag{25}$$

Prisjetimo se kako smo statističke distribucije (F, t i χ^2) uveli da bi procijenili populacijske parametre μ i σ. Sada je pitanje kako procijeniti populacijske parametre α i β. Istraživanja na populacijama su ponekad nemoguća ponekad preskupa pa definiramo slučajne uzorke (od recimo n elemenata). Uzorak od

n parova primanja i potrošnje slučajno izaberemo iz populacije. Na uzorku definiramo regresijsku jednadžbu uzorka:

$$Y_i = \tilde{\alpha} + \tilde{\beta} X_i \tag{26}$$

i ide od 1 do n, jer imamo n parova (primanje, potrošnja) u uzorku. Jasno, i na uzorku postoji razlika izmedju stvarne i očekivane potrošnje za dana primanja X_i

$$Y_i = \tilde{Y}_i + e_i = \tilde{\alpha} + \tilde{\beta} X_i + e_i \tag{27}$$

gdke s e_i označavamo rezidual, kako zovemo za uzorak razliku stvarne i očekivane varijable Y.

Uočimo, **različiti uzorci daju različite parametre $\tilde{\alpha}$ i $\tilde{\beta}$**. Dakle, različiti uzorci daju različite procjene populacijskih parametara α i β.

METODA NAJMANJIH KVADRATA

Ranije smo spomenuli kako smo pretpostavili linearnu vezu izmedju varijabli X i Y. A kako možemo izračunati parametre α i β. Kad grafički predočimo parove (X_i, Y_i) jasno je ukoliko pretpostavimo pravac kao najbolju funkciju prilagodbe medju varijablama, da postoji neki od njih beskonačno, koji se najbolje slaže s podacima X_i i Y_i. Za svaki par (X_i, Y_i) možemo izračunati rezidual e_i, razliku izmedju stvarne Y_i i očekivane vrijednosti $E(Y_i)$. Najbolji je onaj pravac za koji je suma kvadrata e_i najmanja. Dakle ova metoda traži minimum za

$$\Sigma_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} X_i)^2 \tag{28}$$

Pri minimiziranju dolazimo do slijedećih izraza:

$$\overline{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i; \qquad \overline{Y} = \frac{1}{n}\Sigma_{i=1}^{n} Y_i \tag{29}$$

$$S_{xx} = \Sigma_{i=1}^{n}(X_i - \overline{X})^2 = \Sigma_{i=1}^{n} X_i^2 - \frac{(\Sigma_{i=1}^{n} X_i)^2}{n} \tag{30}$$

$$S_{yy} = \Sigma_{i=1}^{n}(Y_i - \overline{Y})^2 = \Sigma_{i=1}^{n} Y_i^2 - \frac{(\Sigma_{i=1}^{n} Y_i)^2}{n} \tag{31}$$

$$S_{xy} = \Sigma_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \Sigma_{i=1}^{n} X_i Y_i - \frac{(\Sigma_{i=1}^{n} X_i)(\Sigma_{i=1}^{n} Y_i)}{n} \tag{32}$$

PROCJENE β i α

$$\tilde{\beta} = \frac{S_{xy}}{S_{xx}} \tag{33}$$

$$\tilde{\alpha} = \overline{Y} - \tilde{\beta}\,\overline{X} \tag{34}$$

Pretpostavka je kako smetnje ϵ zadovoljavaju normalnu distribuciju čiju σ^2 moramo procijeniti:

$$S^2 = \frac{SSE}{n-2} \tag{35}$$

gdje je

$$SSE = \Sigma_{i=1}^{n} e^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \tag{36}$$

Sa SSE označili smo sumu kvadrata. Dijelimo s n-2 jer krećemo s n podataka, ali koliko je stupnjeva slobode. Kako metoda procjenjuje 2 parametra (koji su dani podacima), dakle imamo dvije relacije medju podacima i stoga $n-2$ stupnja slobode.

TABLES

TABLE I. Primjer

X	Y	X^2	Y^2	xy	$\tilde{\alpha} + \tilde{\beta}X$
3	9	9	81	27	7.15
3	5	9	25	15	7.15
4	12	16	144	48	9.89
5	9	25	81	45	12.63
6	14	36	196	84	15.37
6	16	36	256	96	15.37
7	22	49	484	154	18.11
8	18	64	324	144	20.85
8	24	64	576	192	20.85
9	22	81	484	198	23.59

Izračunajte S_{xx}, S_{yy}, S_{xy}, SSE, β i α.

PRIMJER 2:

Zadani su X i Y.

X 0 1 6 3 5

Y 4 3 0 2 1

Izračunajte S_{xx}, S_{yy}, S_{xy}, SSE, β i α. Grafički prikažite podatke.

Kvaliteta regresije (fita)

$$Y_i - \overline{Y} = \tilde{Y}_i - \overline{Y} + e_i \tag{37}$$

Kvadriranjem

$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\tilde{Y}_i - \overline{Y})^2 + \Sigma e_i^2 \tag{38}$$

da se vidjeti da mješoviti član iščezava

SST = SSE + SSR

total sum of square = explained sum of squares + sum of squares of residuals

mjera kvalitete fita: što je manji član, bolji fit

$$R^2 = \frac{SSE}{SST} \tag{39}$$

R može biti izmedju 0 i 1.

Procjena regresijskih parametara

Razumljivo je da različiti uzorci daju različie procjene parametara. Kad bi napravili beskonačno uzoraka i za svaki izračunali parametre $\tilde{\alpha}$ i $\tilde{\beta}$ jasno da bi za svaki parametar mogli definirati distribuciju. Da se vidjeti da su distribucije

$$\frac{\tilde{\beta} - \beta}{\sigma_{\tilde{\beta}}} \tag{40}$$

$$\frac{\tilde{\alpha} - \alpha}{\sigma_{\tilde{\alpha}}} \tag{41}$$

standardizirane normalne distribucije $N(0, 1)$. Vrijedi:

$$\sigma_{\tilde{\beta}}^2 = \frac{s^2}{\Sigma_{i=1}^2 (X_i - \overline{X})^2} \tag{42}$$

gdje je

$$S^2 = \frac{\Sigma_{i=1}^2 e_i^2}{n - 2} \tag{43}$$