OCENJEVANJE PARAMETROV POPULACIJE

Slučajna spremenljivka, ki zavzame vrednosti izračunane iz vzorca, se imenuje statistika

Najpogosteje jih uporabljamo za ocenjevanje parametrov populacije iz katere je izbran slučajni vzorec

Statistiko, ki jo uporabljamo za ocenjevanje parametra populacije imenujemo cenilko.

Vrednost cenilke je ocena parametra populacije

Cenilka je postopek po katerem iz izbranega vzorca izračunamo oceno parametra.

Centralni limitni izrek

Če iz populacije z aritmetično sredino μ in standardnim odklonom σ izberemo slučajni vzorec velikosti n, katerega aritmetična sredina je $\overline{x}$

potem je veličina

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

tudi slučajna spremenljivka in teži k standardizirani normalni, ko velikost vzorca raste.

Praktično to pomeni, da je aritmetična sredina vzorca $\overline{X}$ porazdeljena normalno z matematičnim upanjem μ in standardnim odklonom $\sigma/\sqrt{n}$

Za $n \geq 30$ zelo dobri rezultati

Zaradi centralnega limitnega izreka velja

$$M(\overline{X}) = \mu$$

Lastnost cenilke, da je njeno matematično upanje enako parametru populacije, ki ga s cenilko ocenjujemo, imenujemo nepristranskost

Sami cenilki pa pravimo nepristranska cenilka

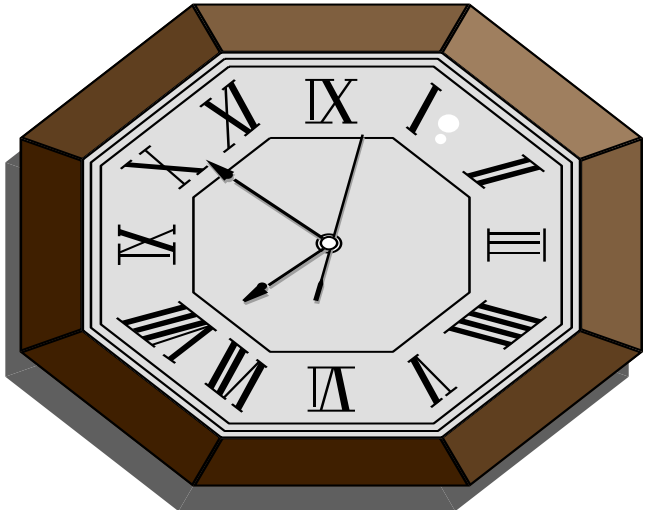
Tudi za varianco vzorca velja

$$M(S^2) = \sigma^2$$

Vrednost cenilke s katero ocenjujemo neki parameter populacije imenujemo točkasta ocena tega parametra

Interval na katerem z določeno verjetnostjo leži neznani parameter, imenujemo interval zaupanja tega parametra, verjetnosti pa pravimo koeficient zaupanja.

Neznani parameter tako ocenimo z nekim intervalom, ki mu pravimo intervalska ocena parametra.



Ocenjevanje aritmetične sredine populacije

Vzemimo normalno porazdeljeno populacijo $N(\mu,\sigma)$ z neznanim matematičnim upanjem in znanim standardnim odklonom σ

Izberimo iz nje slučajni vzorec velikosti n :

$$\{x_1, x_2, \ldots, x_n\}$$

Aritmetična sredina vzorca

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

je normalna slučajna spremenljivka z istim matematičnim upanjem kot populacija iz katere smo izbrali vzorec in standardnim odklonom

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

torej

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

Za izbrani vzorec je vrednost aritmetične sredine vzorca točkovna ocena za matematično upanje μ populacije iz katere izhaja vzorec.

Z vpeljavo standardizirane slučajne spremenljivke

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}, \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

in predpisane verjetnosti α obstoja natanko določeno realno število k, da velja:

$$P\left(-k \le \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \le k\right) = 1 - \alpha$$

Iz te enačbe dobimo intervalsko oceno za μ

$$\bar{x} - k\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + k\frac{\sigma}{\sqrt{n}}$$

temu intervalu pravimo *interval zaupanja* aritmetične sredine

Verjetnost α pri kateri smo določili realno število k, imenujemo koeficient tveganja ali tudi napaka I.vrste, medtem ko verjetnost 1-α imenujemo koeficient zaupanja.

Če predpišemo napako vzorca $|\bar{x} - \mu| \leq d$

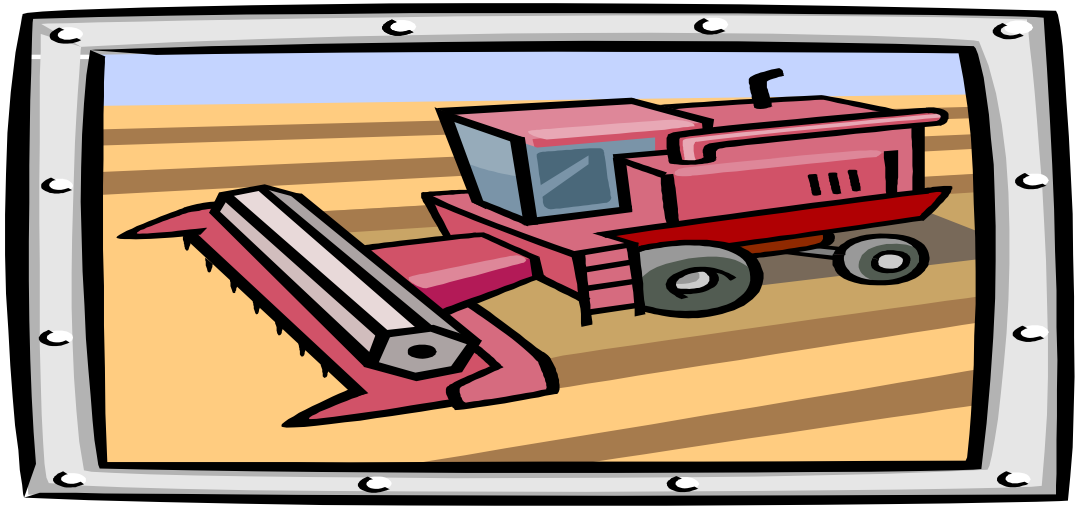
potem je velikost vzorca določena z

$$n \geq \left(\frac{k.\sigma}{d} \right)^2$$

Testiranje hipotez

Postavljanje odločitev o parametrih populacije imenujemo testiranje hipotez

Testi so parametrični kadar se nanašajo na kvantitativne parametre populacije, če pa se nanašajo na nekvantitativne parametre, pa so testi neparametrični



Ničelna hipoteza H_0

je predpostavka o vrednosti nekega parametra populacije.

Alternativna ali nasprotna hipoteza H_1

je predpostavka,da izbrani parameter nima vrednosti,kot jo predpostavlja ničelna hipoteza.

Odločitev o sprejetju ali zavrnitvi hipoteze je zasnovana na statistiki, ki jo imenujemo test hipoteze in jo izračunamo iz podatkov vzorca.

Interval sprejemanja hipoteze določa vrednosti testa hipoteze pri katerem ničelno hipotezo sprejmemo

Zavrnitev ničelne hipoteze, če je ta pravilna imenujemo **napaka I.vrste**

Verjetnost, da to napako naredimo označimo z α

Sprejetje ničelne hipoteze, če je ta napačna imenujemo **napaka II.vrste**. Verjetnost, da to napako napravimo, označimo z β.

Postopek testiranja hipotez izvedemo v naslednjih korakih:

1.Postavimo H_0 proti H_1 in določimo α

2.Izberemo primerno statistiko testa hipoteze in določimo območje sprejemanja ničelne hipoteze velikosti α

3.Izračunamo vrednost statistike testa hipoteze iz podatkov vzorca

4. Preverimo,če vrednost statistike testa hipoteze pade v območje sprejemanja ničelne hipoteze in skladno s tem ničelno hipotezo sprejmemo ali pa jo ne sprejmemo

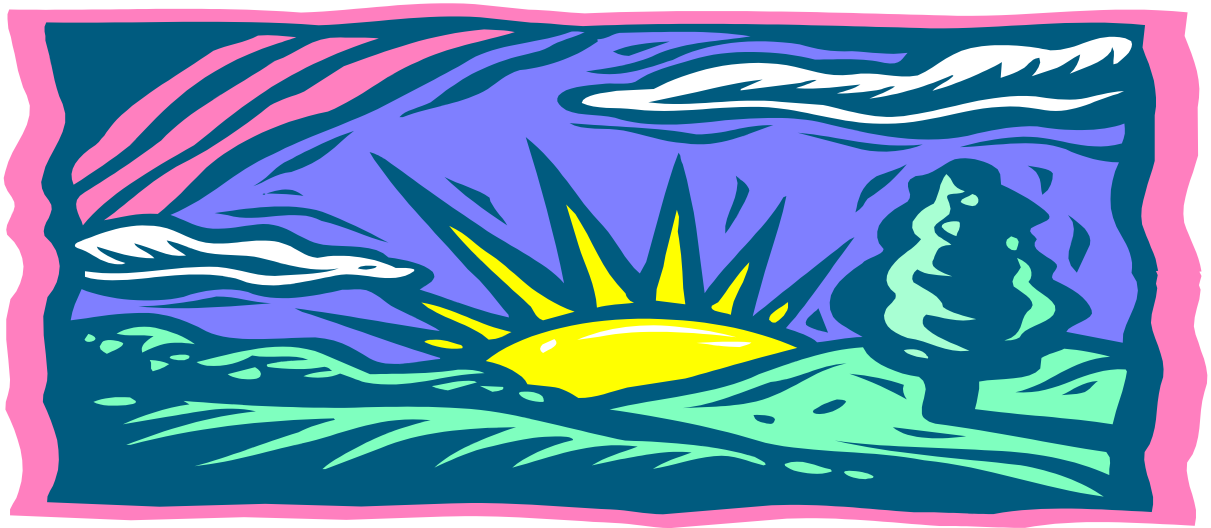
Test aritmetične sredine populacije

V populaciji N(μ,σ) je neznano aritmetična sredina in poznan standardni odklon.

Postavimo ničelno hipotezo

$$H_0 : \mu = A \qquad H_1 : \mu \neq A$$

kjer je A izbrano realno število



Statistika

$$t = \frac{\bar{x} - A}{\sigma}\sqrt{n}$$

je za n>30 standardizirana normalna slučajna spremenljivka N(0,1).

Pri napaki I.vrste z verjetnostjo α je interval sprejemanja ničelne hipoteze množica realnih števil $t \in (-k,k)$, kjer je število k določeno z zvezo:

$$P\left(-k \le \frac{\bar{x} - A}{\sigma}\sqrt{n} \le k\right) = 1 - \alpha$$

Intervalu $(-k,k)$ pravimo tudi interval sprejemanja ničelne hipoteze.

REGRESIJA IN KORELACIJA

Regresija

O regresiji govorimo,kadar sta dva ali več pojavov(količin) v medsebojni odvisnosti.
Regresija je *enostavna* kadar nastopata v medsebojni odvisnosti samo dva pojava(količini).
Naloga regresije je, poiskati tako funkcijo , ki najbolje podaja medsebojno odvisnost količin.

Odvisnost je *enostranska* , kadar je količina X vzrok, količina Y pa posledica.

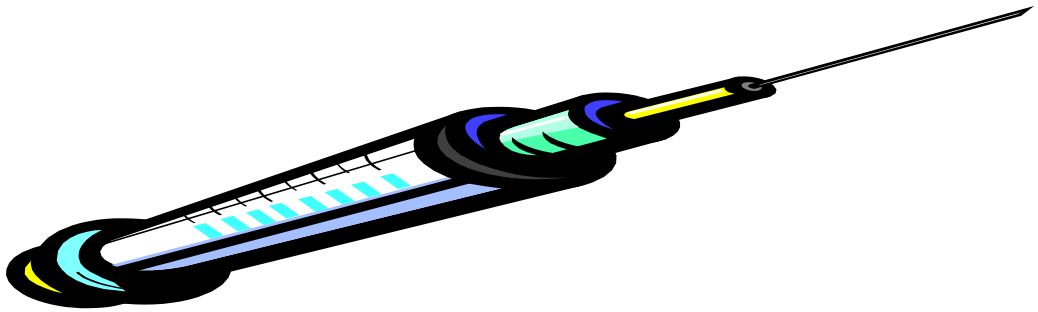
Odvisnost je *dvostranska* , kadar ni možno določiti, kaj je vzrok in kaj posledica.

Kadar je količina Y slučajna spremenljivka, njeno vrednost ne moremo natanko vnaprej predvideti, ko je vrednost neodvisne spremenljivke x znana, to je količina X zavzame vrednost x

Določimo lahko matematično upanje slučajne spremenljivke Y, dejanske vrednosti pa "nihajo" okrog matematičnega upanja v skladu Z porazdelitvenim zakonom slučajne spremenljivke Y.

Lahko zapišemo le zvezo

$$\mathbf{M(Y|x) = f(x)}$$



Če je ε odmik realizirane vrednosti slučajne spremenljivke(pojava) Y od matematičnega upanja $M(Y|x)$

$$\varepsilon = Y - M(Y|x)$$

lahko zapišemo model

$$Y = M(Y|x) + \varepsilon = f(x) + \varepsilon$$

Količina ε je slučajna spremenljivka in se imenuje napaka, modelu pa pravimo regresijski model

Kadar iščemo odvisnost v obliki linearne funkcije

$$M(Y|x) = \alpha + \beta x$$

govorimo o *linearni regresiji*

Sam regresijski model pa zapišemo v obliki

$$Y = \alpha + \beta x + \varepsilon$$

Parametra α in β ocenimo na naslednji način

Pri predpostavki,da je količina X neodvisna Y pa odvisna spremenljivka in sta količini podani vsaka z slučajnim vzorcem velikosti n :

$$X \square \{x_1, x_2, \ldots x_n\}$$

$$Y \square \{y_1, y_2, \ldots y_n\}$$

izračunamo točkovni oceni a in b za parametra α in β.

Po metodi najmanjših kvadratov jih dobimo kot minimum funkcije dveh spremenljivk a in b :

$$F = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2 \rightarrow \min\text{imum}$$

Iz tega pogoja dobimo za oceni a in b formuli

$$a = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

Vpeljimo naslednje oznake

$$S_{xx} = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = \sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right) = \sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y} \,.$$

S temi oznakami oceno b koeficienta β lahko zapišemo

$$b = \frac{S_{xy}}{S_{xx}}$$

Ocenjeno regresijsko premico zapišemo

$$y = a + bx$$

in če vzamemo povprečne vrednosti v obeh vzorcih, dobimo

$$\bar{y} = a + b\bar{x}$$

Odtod dobimo oceno za α

$$a = \bar{y} - b.\bar{x}$$

Na ta način pridemo po enostavnejši poti do ocene za parameter α

Varianco σ^2 količine(v pojavu) Y imenujemo

skupna ali začetna varianca

Njena točkovna ocena za izbrani vzorec

$$\{y_1, y_2, \cdots y_n\}$$

je

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

ki jo lahko tudi zapišemo krajše

$$S_Y^2 = \frac{S_{yy}}{n-1}$$

Delimo jo v dva dela : na *nepojasnjeno varianco* s_e^2

katere ocena za izbrani vzorec je

$$\mathbf{s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2}$$

kar lahko tudi zapišemo

$$s_e^2 = \frac{1}{n-2} . \left(S_{yy} - b . S_{xy} \right)$$

in *pojasnjeno varianco*

$$s^2_{XY} = s^2_Y - s^2_e$$

Korelacijska analiza proučuje, kako dobra je matematična povezava med količinama X in Y , ki ju povezuje regresijska premica

Determinacijski koeficient (koeficient določenosti) nam v primeru enostranske odvisnosti meri linearno povezavo med vzrokom X in posledico Y in ga ocenimo :

$$D = \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Mejni primeri

a./ D=1 med količino X in količino Y obstoja popolna matematična povezava v obliki linearne funkcije (napaka ε v modelu je 0)

b/ D=0 med količinama X in Y ni nobene linearne odvisnosti

c./ 0 < D <1 med X in Y obstoja verjetna linearna povezava.

Koeficient korelacije ρ

meri linearno odvisnost med dvostransko odvisnima količinama X in Y, njegova točkovna ocena je

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

ki jo lahko tudi zapišemo na naslednji način

$$r = \frac{S_{xy}}{\sqrt{S_{yy}.S_{xx}}}$$

Vrednosti koeficienta korelacije leže med -1 in +1

Za vrednost -1 imamo strogo obratno linearno odvisnost in za +1 strogo premo linearno odvisnost

V primeru,ko je koeficient korelacije 0 ni nobene linearne odvisnosti.

Koeficient določenosti D in koeficient korelacije r sta povezana z naslednjo zvezo :

$$D=1-\frac{n-1}{n-2}\left(1-r^2\right)$$

Koeficient korelacije ranžirnih vrst

meri odvisnost dveh rangirnih vrst.

Za dve vrsti

$\{v_1, v_2, \ldots v_n\}$ **in** $\{w_1, w_2, \ldots, w_n\}$

je določen s formulo

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (w_i - v_i)^2$$

ANALIZA ČASOVNIH VRST

O časovni vrsti govorimo, kadar sta v medsebojni odvisnosti čas t in neka količina Y.

Pomembna uporaba pri prognoziranju

Prognoza je napoved, da bo količina Y imela v nekem prihodnjem času neko določeno vrednost

Pri prognozah ekstrapoliramo stanja časovne vrste v prihodnost.

Označimo $\{y_1, y_2, \ldots, y_t\}$ **stanja časovne vrste**

$\{y_1', y_2' \ldots, y_t'\}$ **napovdi**

Metoda drsečih sredin

y_t **dejanska vrednost količinne Y v času t**

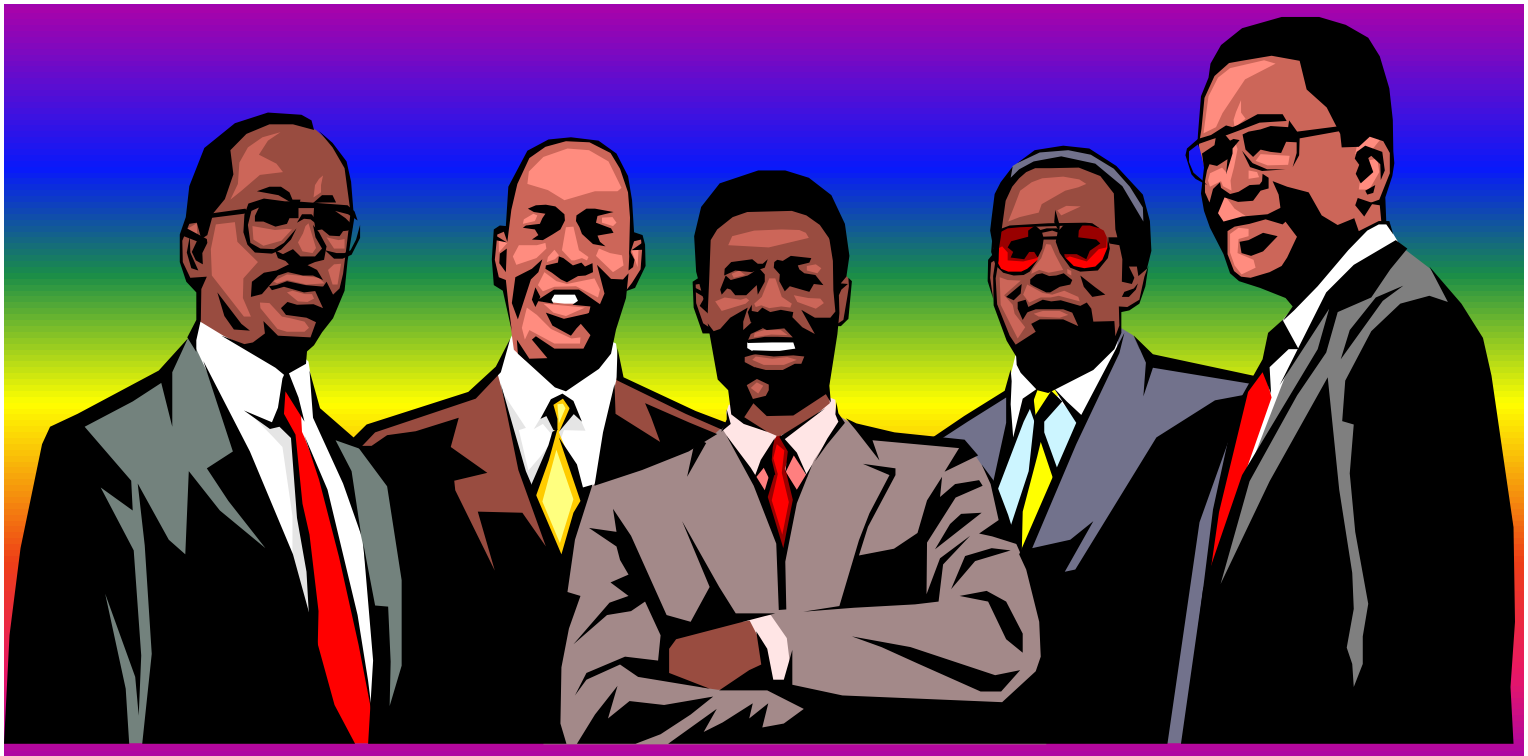
y'_t **napovedana vrednost v času t in m korak ali periodo drsenja**

Vrednost napovedi pojava Y v času t+1 izračunamo po obrazcu

$$y'_{t+1} = y'_t + \frac{y_t - y_{t-m}}{m}$$

Napoved v času t = m+1
(začetno napoved za periodo m) pa dobimo

$$y'_{m+1} = \frac{y_1 + y_2 + \ldots y_m}{m}$$



Metoda eksponentnega glajenja

Pri tej metodi z ocenjenim faktorjem glajenja

$$0 \leq g \leq 1$$

Napovedano vrednost s preteklimi vrednostmi eksponentno večamo ali manjšamo po formuli :

$$y'_{t+1} = y'_t + g\left(y_t - y'_t\right)$$

Za začetno napoved vzamemo,da je

$$y'_1 = y_1$$

Model časovne vrste

je neka funkcija , odvisna od časa, ki jo določimo z metodo najmanjših kvadratov in ji rečemo trend časovne vrste

Problem določanja funkcije je zato enak kot pri regresiji.

Kazalci dinamike časovne vrste

Variabilnost časovne vrste merimo s parametri, ki jim pravimo kazalci dinamike časovne vrste.

Tempo rasti pokaže relativno razliko med dvema zaporednima členoma

$$T_k = \frac{y_k - y_{k-1}}{y_{k-1}} 100\%$$

Koeficient dinamike K_k

kaže relativno spremembo dveh zaporednih členov

$$K_k = \frac{y_k}{y_{k-1}}$$

Verižni indeks I_k

je v procentih izražen koeficient dinamike

$$I_k = \frac{y_k}{y_{k-1}} 100\%$$

Bazni indeks I_{k0}

dobimo, če v verižnem indeksu vzamemo za y_{k-1} nek karakteristični člen y_0

$$I_{k0} = \frac{y_k}{y_0} 100\%$$