Predstavljanje i obrada podataka

Načini prikazivanja podataka

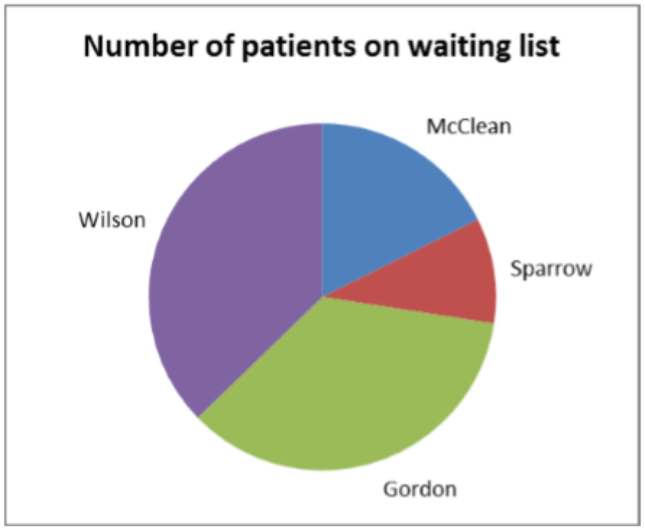
- Podaci koji se prikupljaju u postupku merenja mogu se prikazati:
  - tabelarno,
  - grafički.

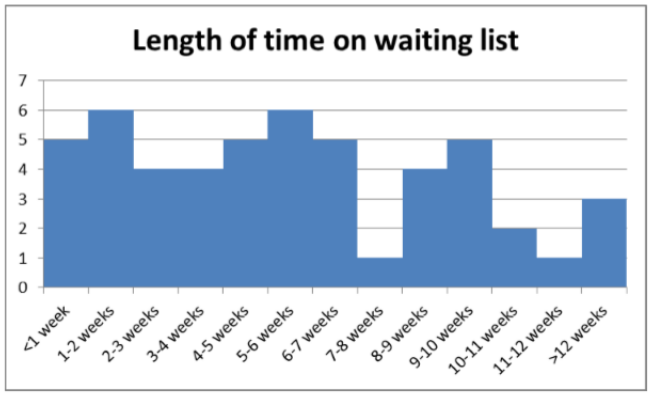
- Tabelarni prikaz:
  - uglavnom prirodno sakupljamo podatke ovim putem,
  - sadrži sve podatke,
  - tačan je,
  - pogodan je za manje skupove podataka,
  - težak je za interpretaciju podataka i određivanje pravilnosti u podacima.

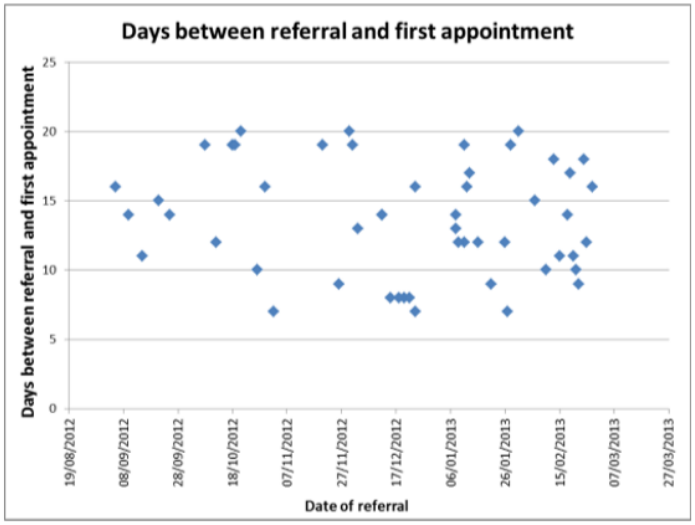
- Grafički prikaz:
  - postoji više opcija: histogram, kružni grafikon *(pie chart),* korelacioni dijagram/dijagram rasipanja (*scatter plot*), linijski dijagram (*line chart*), ...
  - olakšava interpretaciju na račun izgubljene tačnosti,
  - svaka od opcija pogodna je za određene situacije, ali ima i nedostatke.

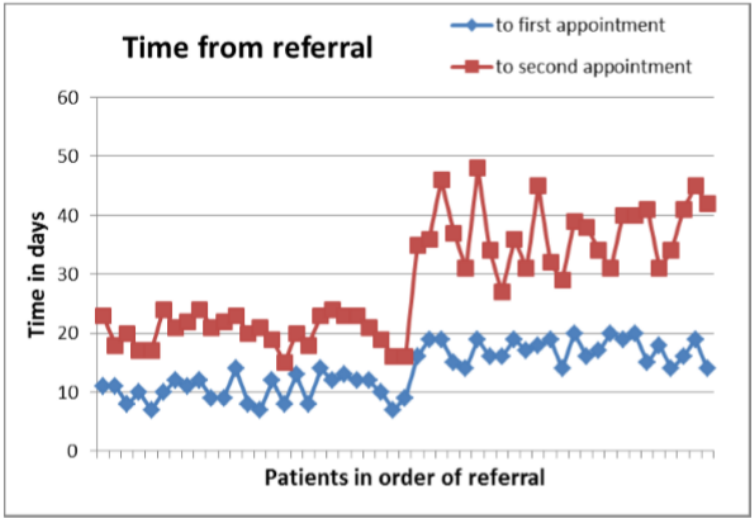
Primeri

GP of referral	Number of patients referred
10577	1
15498	4
13457	9
15468	4
41324	2
15846	1
44532	5
12546	4
12489	8
25432	2
68542	5
54798	6
	51









** ilustracije preuzete sa: www.qihub.scot.nhs.uk/media/530244/data%20presentation%20types.pdf*

Uprošćavanje podataka

- Prilikom vršenja merenja nad istim (stacionarnim ili dinamičkim) procesom dobija se velika količina podataka.

- Uprošćavanje podataka omogućava lakšu interpretaciju i razumevanje izmerenih vrednosti.

- **Opseg** izmerenih vrednosti – raspon između najveće i najmanje izmerene vrednosti:
  - opseg se može podeliti na manje intervale,
  - svaka izmerena vrednost dodeljuje se jednom od intervala,
  - interval sa najvećim brojem izmerenih vrednosti naziva se **modalni** interval,
  - intervali sa obe strane modalnog intervala sadrže manji broj vrednosti,
  - po pravilu, u praksi, broj izmerenih vrednosti opada ka krajevima opsega.

- Broj izmerenih vrednosti u manjim intervalima opsega može se prikazati na histogramu:
  - u zavisnosti od merenja, ovaj histogram može da ukaže na raspodelu grešaka pri merenju i prisustvo **sistematskih** grešaka.

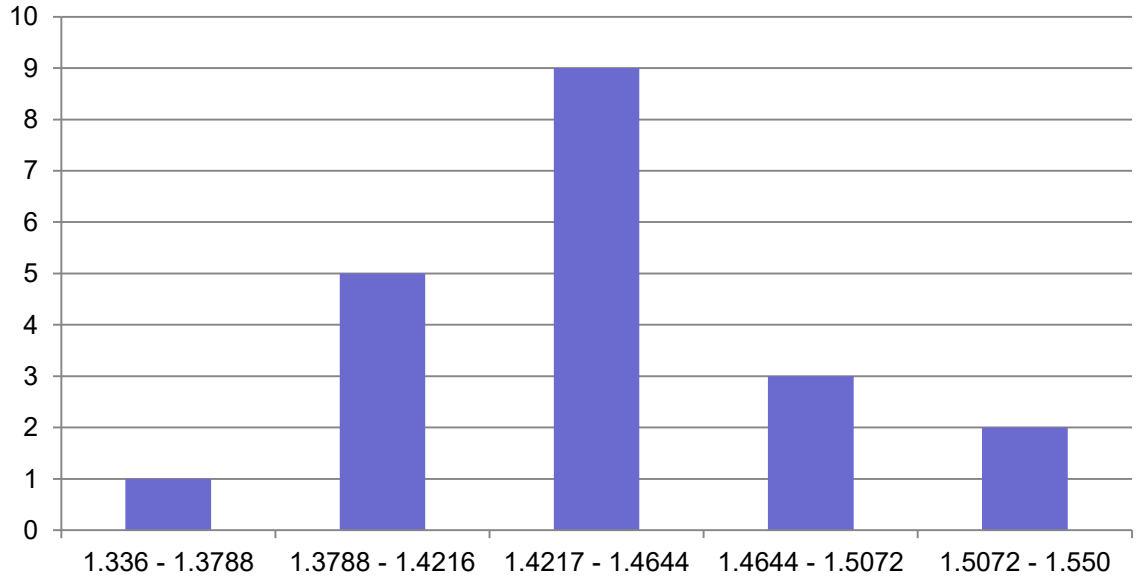
Primer – opseg, histogram

i	*T*(*i*)
1	1.4285
2	1.499
3	1.55
4	1.4385
5	1.4215
6	1.426
7	1.451
8	1.444
9	1.336
10	1.4565
11	1.4525
12	1.498
13	1.412
14	1.447
15	1.512
16	1.385
17	1.396
18	1.412
19	1.445
20	1.504

max	1.55
min	1.336

intervali	broj ponavljanja
1.336 - 1.3788	1
1.3788 - 1.4216	5
1.4217 - 1.4644	9
1.4644 - 1.5072	3
1.5072 - 1.550	2

modalni interval



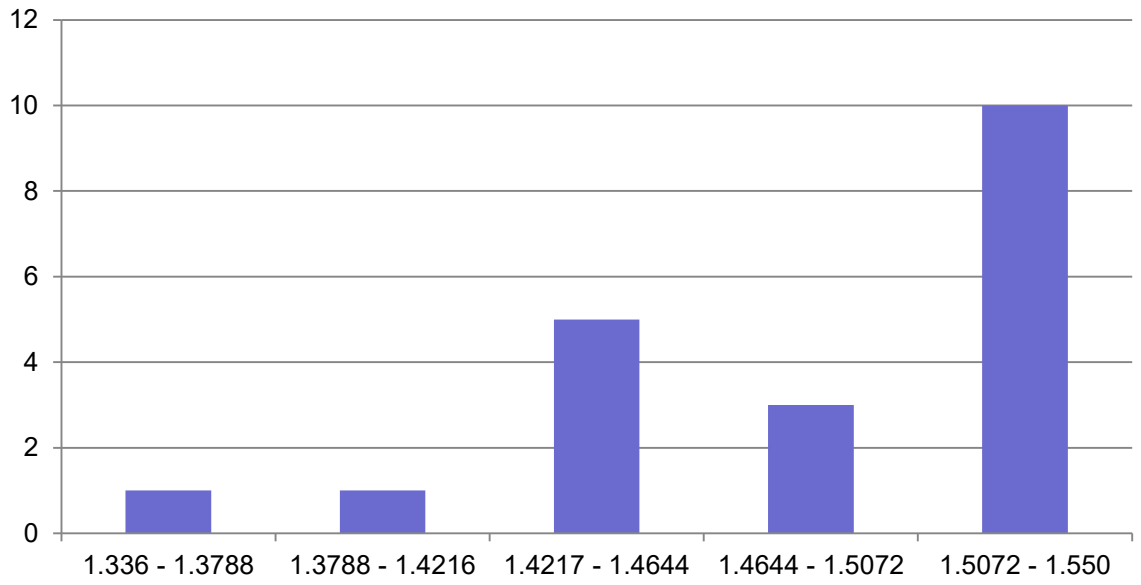
Primer 2 – opseg, histogram

i	*T*(*i*)
1	1.504
2	1.545
3	1.512
4	1.485
5	1.396
6	1.512
7	1.447
8	1.512
9	1.498
10	1.4525
11	1.4565
12	1.336
13	1.544
14	1.541
15	1.426
16	1.5215
17	1.4385
18	1.55
19	1.517
20	1.539

max	1.55
min	1.336

intervali	broj ponavljanja
1.336 - 1.3788	1
1.3788 - 1.4216	1
1.4217 - 1.4644	5
1.4644 - 1.5072	3
1.5072 - 1.550	10

modalni interval



Tipične vrednosti

- **Mere položaja** izmerene veličine mogu biti:
  - srednja vrednost,
  - medijana,
  - modalna vrednost.
- **Srednja vrednost** ($\bar{x}$) definiše se kao:

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

- **Medijana** ($\tilde{x}$) je vrednost za koju važi da je polovina izmerenih vrednosti veća od nje, a druga polovina manja. Ako poređamo izmerene vrednosti od najmanje do najveće (y_n), medijana je:

$$\tilde{x} = \begin{cases} y_{(n+1)/2}, & n \text{ neparno} \\ (y_{n/2} + y_{n/2+1})/2, & n \text{ parno} \end{cases}$$

- **p-ti percentil** skupa izmerenih vrednosti je ona izmerena vrednosti za koju je p% izmerenih vrednosti manje, a prostalih (100-p)% izmerenih vrednosti veće.
- **Modalna vrednost** je ona koja se najčešće pojavljuje u izmerenim vrednostima. Ne mora biti jedinstvena.

Primer – tipične vrednosti

i	*Ts(i)*
1	1.336
2	1.385
3	1.396
4	1.412
5	1.412
6	1.4215
7	1.426
8	1.4285
9	1.4385
10	1.444
11	1.445
12	1.447
13	1.451
14	1.4525
15	1.4565
16	1.498
17	1.499
18	1.504
19	1.512
20	1.55

- srednja vrednost:

$$\overline{T_s} = 1{,}4457$$

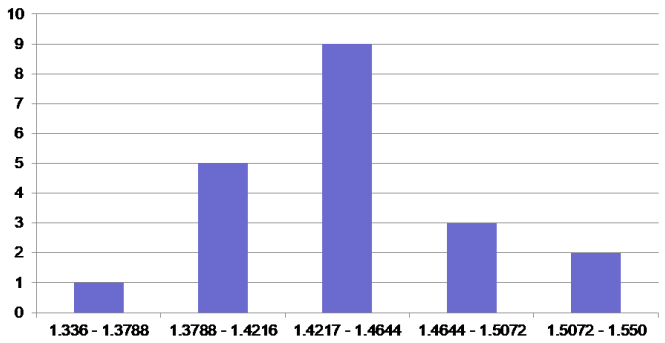
- medijana:

$$\widetilde{T_s} = 1{,}4445$$

- modalna vrednost: 1.412

- 25-ti percentil: 1.4191

- 75-ti percentil: 1.4669



Primer 2 – tipične vrednosti

i	*Ts(i)*
1	1.336
2	1.396
3	1.426
4	1.4385
5	1.447
6	1.4525
7	1.4565
8	1.485
9	1.498
10	1.504
11	1.512
12	1.512
13	1.512
14	1.517
15	1.5215
16	1.539
17	1.541
18	1.544
19	1.545
20	1.55

- srednja vrednost:

$$\overline{T_s} = 1{,}4867$$

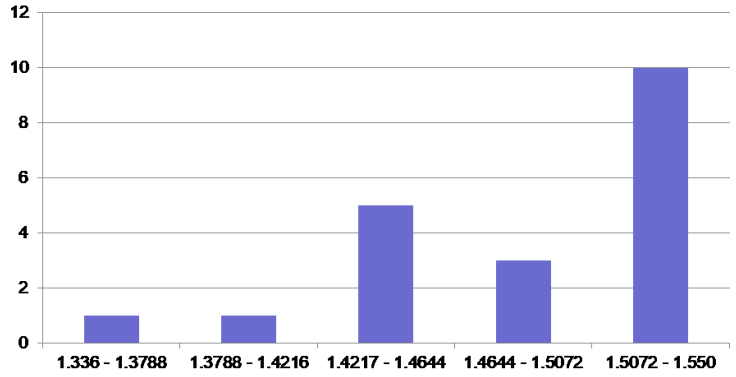
- medijana:

$$\widetilde{T_s} = 1{,}508$$

- modalna vrednost: 1.512

- 25-ti percentil: 1.4484

- 75-ti percentil: 1.5346



Mere razmere

- **Mera razmere** izmerene veličine može biti:
  - varijansa,
  - standardna devijacija,
  - opseg, interkvartalni opseg,
  - srednja apsolutna devijacija i medijana apsolutne devijacije, ...
- **Varijansa** (s^2) definiše se kao:

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

  - varijansa daje veću težinu vrednostima na krajevima opsega,
  - standardna devijacija (s) predstavlja kvadratni koren varijanse,
  - obe veličine ukazuju na to koliko su izmerene vrednosti skoncentrisane oko srednje vrednosti, kao i na **slučajnu grešku** u procesu merenja.
- **Interkvartalni opseg** – raspon između 75-tog i 25-tog percentila.
- Srednja apsolutna devijacija i medijana apsolutne devijacije

$$SAD = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{n}$$

$$MAD = med(|x_i - \bar{x}|)$$

Primer – mere razmere

i	*Ts*(*i*)
1	1.336
2	1.385
3	1.396
4	1.412
5	1.412
6	1.4215
7	1.426
8	1.4285
9	1.4385
10	1.444
11	1.445
12	1.447
13	1.451
14	1.4525
15	1.4565
16	1.498
17	1.499
18	1.504
19	1.512
20	1.55

- varijansa:

$$s^2 = 0.0023$$

- standardna devijacija:

$$s = 0.0481$$

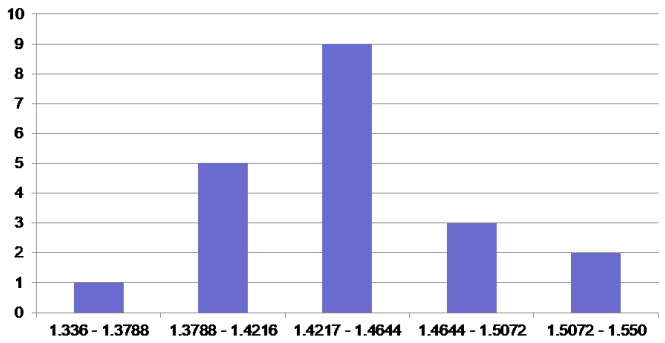
- interkvartalni opseg : (1.4191-1.4669)

- srednja apsolutna devijacija:

$$SAD = 0.0358$$

- medijana apsolutne devijacije:

$$MAD = 0.029$$



Primer 2 – mere razmere

i	*Ts(i)*
1	1.336
2	1.396
3	1.426
4	1.4385
5	1.447
6	1.4525
7	1.4565
8	1.485
9	1.498
10	1.504
11	1.512
12	1.512
13	1.512
14	1.517
15	1.5215
16	1.539
17	1.541
18	1.544
19	1.545
20	1.55

- varijansa:

$$s^2 = 0.0031$$

- standardna devijacija:

$$s = 0.0554$$

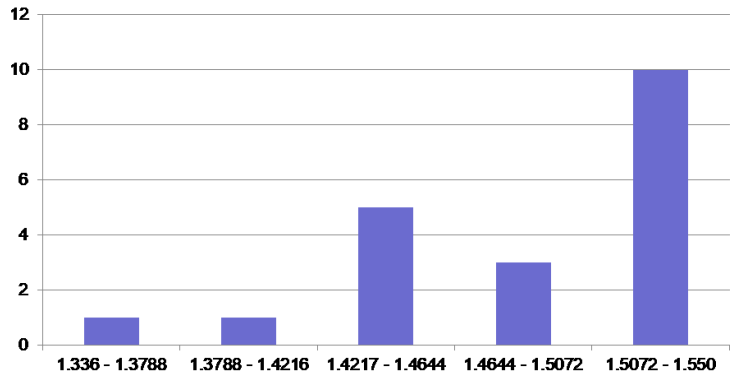
- interkvartalni opseg : (1.4484-1.5346)

- srednja apsolutna devijacija:

$$SAD = 0.0456$$

- medijana apsolutne devijacije:

$$MAD = 0.0373$$



Druge mere

- **Zakrivljenost** (z) definiše se kao:

$$z = \frac{1}{s^3} \cdot \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{n-1}$$

  – predstavlja meru simetrije podataka u odnosu na srednju vrednost,
  – kada je raspodela simetrična, jednaka je nuli.

- **Spljoštenost** (k) definiše se kao:

$$k = \frac{1}{s^4} \cdot \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{n-1}$$

  – spljoštenost veća od tri ukazuje na raspodelu sa „vrhovima“, a manja od tri na „ravnu“ raspodelu.

Primer – druge mere

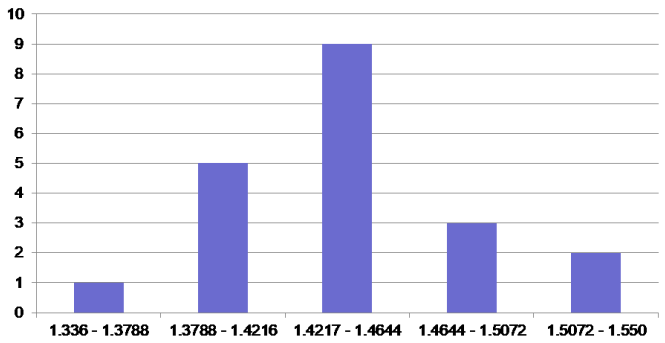
i	*Ts*(*i*)
1	1.336
2	1.385
3	1.396
4	1.412
5	1.412
6	1.4215
7	1.426
8	1.4285
9	1.4385
10	1.444
11	1.445
12	1.447
13	1.451
14	1.4525
15	1.4565
16	1.498
17	1.499
18	1.504
19	1.512
20	1.55

- zakrivljenost:

$$z = 0.0694$$

- spljoštenost:

$$k = 3.2751$$



Primer 2 – druge mere

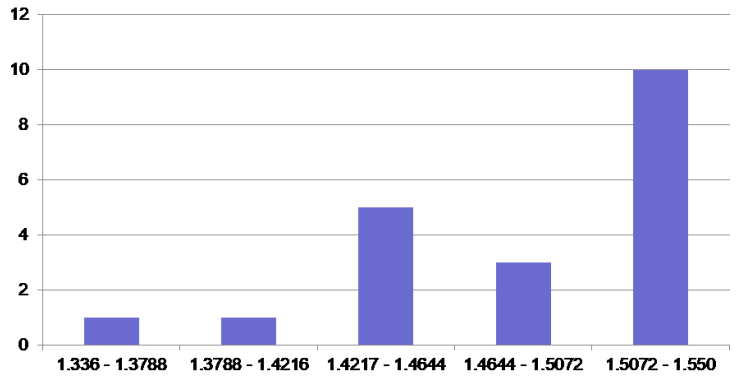
i	*Ts*(*i*)
1	1.336
2	1.396
3	1.426
4	1.4385
5	1.447
6	1.4525
7	1.4565
8	1.485
9	1.498
10	1.504
11	1.512
12	1.512
13	1.512
14	1.517
15	1.5215
16	1.539
17	1.541
18	1.544
19	1.545
20	1.55

- zakrivljenost:

$$z = -1.0991$$

- spljoštenost:

$$k = 3.7047$$



Raspodela

- Raspodela merenih vrednosti definiše se najčešće na osnovu **funkcije gustine raspodele**. Kod kontinualnih raspodela, verovatnoća se najčešće definiše na intervalu kao:

$$P(a \leq x \leq b) = \int_a^b f(x)\mathrm{d}x$$

  - $f(x)$ predstavlja gustinu raspodele, P verovatnoću, x moguću vrednost merene veličine, a a i b granice intervala unutar kojih se računa verovatnoća.
- **Funkcija raspodele** jednaka je verovatnoći da neka veličina bude manja ili jednaka posmatranoj vrednosti:

$$F(x) = \int_{-\infty}^{x} f(\mu)\mathrm{d}\mu$$

- Podaci dobijeni merenjem mogu biti različito raspoređeni. Česti vid raspodele je **Gausova** ili **normalna** raspodela:

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

  - μ je srednja vrednost, a σ standardna devijacija raspodele,
  - za $\mu = 0$ i $\sigma = 1$ reč je o standardnoj normalnoj raspodeli.

Normalna raspodela

- Normalna raspodela ima sledeće osobine:
  - srednja vrednost, medijana i modalna vrednost su međusobno jednake,
  - opseg raspodele je beskonačan,
  - zakrivljenost raspodele je nula,
  - spljoštenost raspodele je tri.

- Često su podaci prikupljeni merenjem normalne raspodele. Ova raspodela se koristi i pri modelovanju podataka.

- **Centralna granična teorema** pruža teorijski osnov za široku primenu normalne raspodele:
  - posmatramo nekoliko nezavisnih skupova merenja jednog procesa,
  - ako je broj skupova dovoljno veliki (u praksi najčešće veći od 10), tada je:
    - raspodela srednje vrednosti normalna, bez obzira na raspodelu osnovnog skupa,
    - srednja vrednost jednaka srednjoj vrednosti osnovnog skupa,
    - standardna devijacija srednje vrednosti jednaka je $\sigma/\sqrt{N}$, gde je N broj skupova.

Modelovanje procesa

- Pri merenju više veličina, javlja se potreba utvrđivanja veze između pojedinih veličina.

- Postupak u kome se uspostavlja veza između nezavisno izmerenih veličina i jedne ili više zavisnih izmerenih naziva se **modelovanje procesa**.

- Najpoznatije metode su **regresije najmanjih kvadrata**
  - linearno prilagođavanje (*linear fit*):

$$f(\vec{x}, \vec{\beta}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots$$

$$f(x, \vec{\beta}) = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$$

$$f(x, \vec{\beta}) = \beta_0 + \beta_1 \cdot \ln(x)$$

$$f(x, \vec{\beta}) = \beta_0 + \beta_1 \cdot \sin(x) + \beta_2 \cdot \sin(2x) + \beta_3 \cdot \sin(3x)$$

  - nelinearno prilagođavanje (*nonlinear fit*):

$$f(x, \vec{\beta}) = \beta_0 + \beta_0 \cdot \beta_1 \cdot x$$

Modelovanje procesa (2)

- **Linearna regresija**:
  - često se koristi jer su procesi ili linearni ili se mogu linearno aproksimirati u nekom manjem opsegu,
  - ograničena upotreba, loša ekstrapolaciona svojstva, osetljivost na krajevima opsega

- **Težinska regresija** – svaki podatak ima svoju težinu (meru uticaja).

- **Model lokalne regresije** (LOESS) – opis funkcije po segmentima.

Procena (estimacija) parametara

- Problem optimizacije – nepoznati parametri se posmatraju kao nepoznate, a podaci kao koeficijenti.

- Zadatak – minimizacija sume:

$$Q = \sum_{i=1}^{n} \left(y_i - f\left(\vec{x}_i, \vec{\hat{\beta}}\right) \right)^2$$

  - y_i su izmerene vrednosti zavisne promenljive, $\vec{x}_i$ izmerene vrednosti nezavisnih promenljivih, a $\vec{\hat{\beta}}$ nepoznate

- Kod linearnih modela, moguće je analitički odrediti rešenje.

- Kod nelinearnih modela, koriste se numerički algoritmi.

Primer – analitičko rešenje

- Posmatramo linijski model:

$$f(x, \vec{\beta}) = \beta_0 + \beta_1 \cdot x$$

- Zadatak – minimizacija sume:

$$Q = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta_0} + \widehat{\beta_1} \cdot x)\right)^2$$

- Dobijene procenjene vrednosti (izjednačavanjem izvoda Q po parametrima sa nulom):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

Kvalitet modelovanja i provera modela

- **Standardna devijacija greške modela**:

$$\hat{\sigma} = \sqrt{\frac{Q}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - f\left(\vec{x}_i, \vec{\hat{\beta}}\right)\right)^2}{n-p}}$$

  - $\hat{\sigma}$ je procenjena standardna devijacija, p je broj parametara u funkciji regresije,
  - šta ako je $n = p$?

- **Koeficijent varijabilnosti podataka** (R^2):

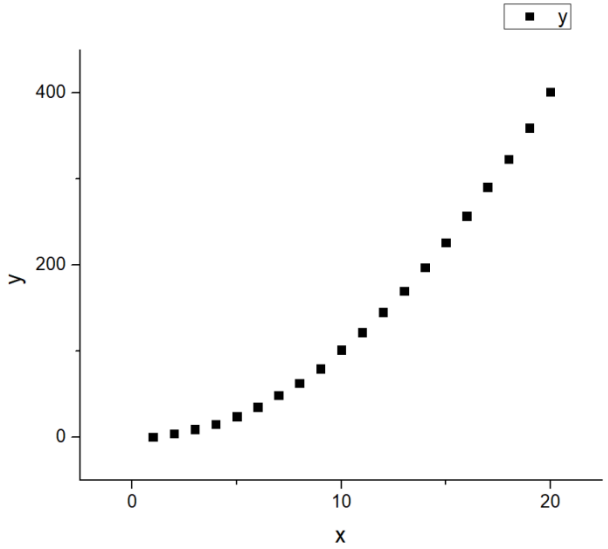
$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - f\left(\vec{x}_i, \vec{\hat{\beta}}\right)\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - mera poklapanja modela i izmerenih podataka – što je bliže jedinici veće je poklapanje,
  - nije dovoljan, zbog čega se često vrši grafička analiza ostataka – ako je model dobar, ostaci će se ponašati kao slučajna greška nulte srednje vrednosti i određene varijanse:

$$e_i = y_i - f\left(\vec{x}_i, \vec{\hat{\beta}}\right)$$

Primer modelovanja

x(i)	*y*(i)
1	-0.11
2	3.84
3	9.16
4	14.89
5	23.73
6	34.57
7	48.36
8	62.41
9	79.41
10	101.11
11	121.80
12	145.11
13	169.80
14	197.11
15	225.95
16	256.95
17	290.27
18	323.05
19	359.41
20	401.27



		Value	Standard Error
y	Intercept	-1.45362	0.79747
	B1	0.21161	0.1749
	B2	0.99398	0.00809

Statistics

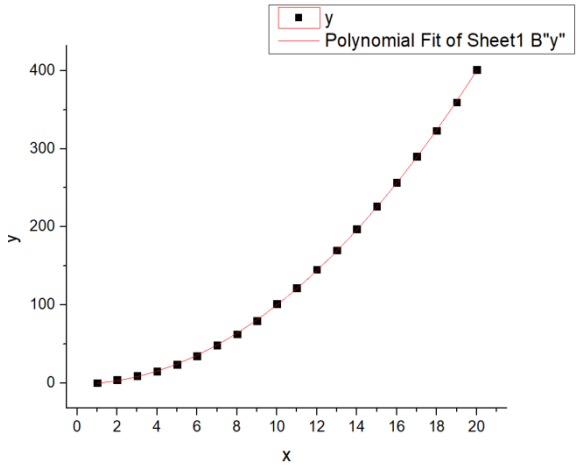
	y
Number of Points	20
Degrees of Freedom	17
Residual Sum of Squares	19.53234
Adj. R-Square	0.99993

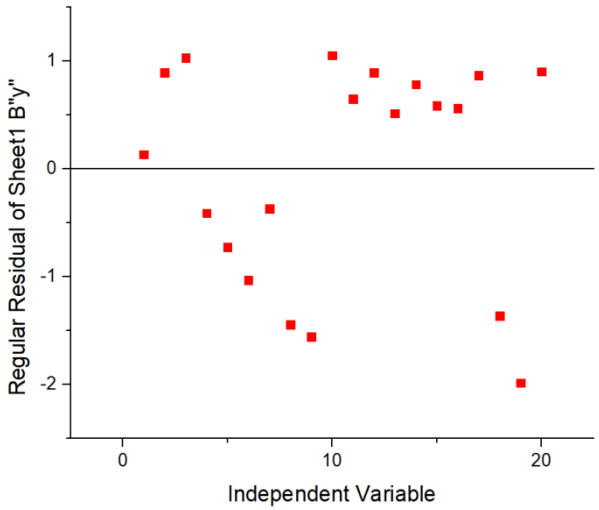
Summary

	Intercept		B1		B2		Statistics
	Value	Standard Error	Value	Standard Error	Value	Standard Error	Adj. R-Square
y	-1.45362	0.79747	0.21161	0.1749	0.99398	0.00809	0.99993

ANOVA

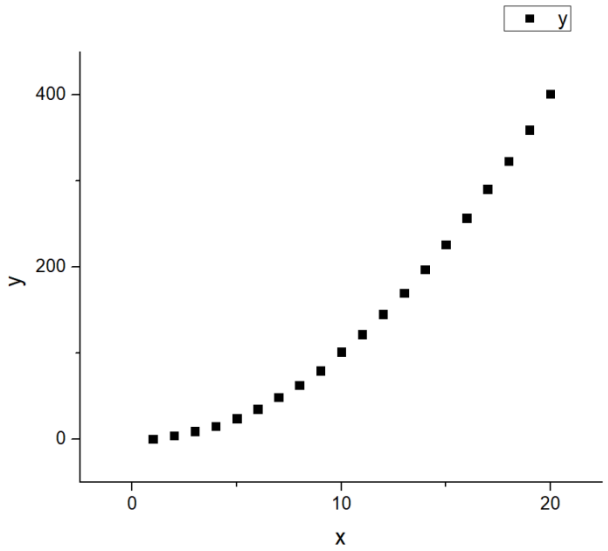
		DF	Sum of Squares	Mean Square	F Value	Prob>F
y	Model	2	312994.79054	156497.39527	136207.73485	0
	Error	17	19.53234	1.14896		
	Total	19	313014.32288			





Primer modelovanja 2

x(i)	*y*(i)
1	-0.11
2	3.84
3	9.16
4	14.89
5	23.73
6	34.57
7	48.36
8	62.41
9	79.41
10	101.11
11	121.80
12	145.11
13	169.80
14	197.11
15	225.95
16	256.95
17	290.27
18	323.05
19	359.41
20	401.27



		Value	Standard Error
y	Intercept	-77.99011	14.42826
	Slope	21.0852	1.20445

Statistics

	y
Number of Points	20
Degrees of Freedom	18
Residual Sum of Squares	17364.81071
Pearson's r	0.97187
Adj. R-Square	0.94144

Summary

	Intercept		Slope		Statistics
	Value	Standard Error	Value	Standard Error	Adj. R-Square
y	-77.99011	14.42826	21.0852	1.20445	0.94144

ANOVA

		DF	Sum of Squares	Mean Square	F Value	Prob>F
y	Model	1	295649.51217	295649.51217	306.46411	9.4702E-13
	Error	18	17364.81071	964.71171		
	Total	19	313014.32288