

Osnove statistične obdelave podatkov

Avtor: Martin Horvat

27. april 2008

Povzetek

Besedilo naj služi kot vodilo ali nasvet pri pisanju poročil o meritvah. Napotki so splošno veljavni, vendar so namenjeni predvsem študentom fizike pri predmetu Praktikum 2.

1 Splošni napotki k grafični prezentaciji meritev

Na grafu naj bodo jasno vidne točke meritev. Te točke so lahko povezane tudi s črto. Če imamo nek smiselni model opisa teh točk na graf še narišemo krivuljo, ki se prilega meritvam najboljše. Za slednje sta na razpolago metodi prilagajanja in interpoliranja. Prva je precej bolj praktično uporabna.

Pomni: S polinomom dovolj visokega reda lahko opišemo vse točke danih meritev tudi eksaktno. Vendar takšna krivulja pogosto ni smiselna, saj ne opiše nujno dobro morebitnih dodatnih točk. Model s katerim nekaj opišemo mora nositi neko univerzalnost.

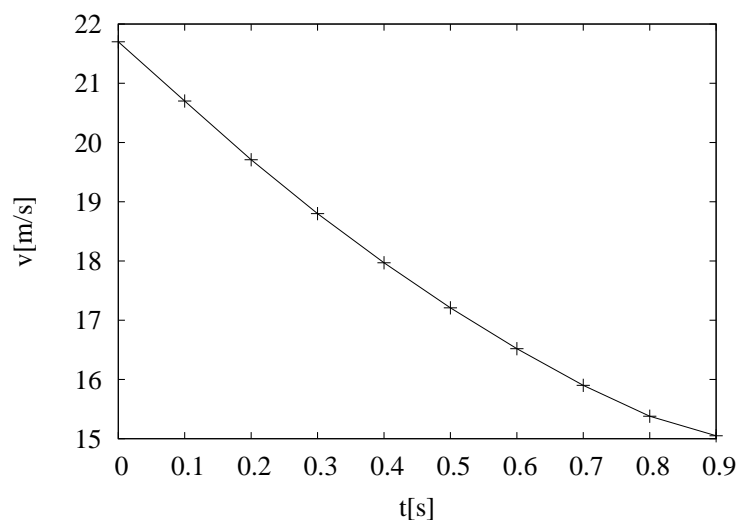
Za primer grafično prezentirajmo podatke navedeni v tabeli 1. Te podatke shranimo v datoteko `podatki.dat`. Sedaj pogledjmo kako podatke iz datoteke prikažemo v programih : Gnuplot in Mathematica:

- z GnuPlot-om narišemo diagram in ga shranimo v PostScript datoteko `podatki.eps`:

```
set term post eps enh "Times-Roman" 26
set out 'podatki.eps'
```

Tabela 1: Simbolične meritve hitrosti v v odvisnosti od časa t , ki vsaka vsebuje neko napako.

$t[s]$	$v[m/s]$	napaka
0	21.70	0.1
0.1	20.70	0.2
0.2	19.71	0.3
0.3	18.80	0.4
0.4	17.97	0.5
0.5	17.21	0.6
0.6	16.52	0.7
0.7	15.90	0.8
0.8	15.38	0.9
0.9	15.05	1



Slika 1: Slika narisana s programom Gnuplot.

```
unset key
set xlabel 't[s]'
set ylabel 'v[m/s]'
plot 'podatki.dat' w lp ps 2
set out
```

in dobimo diagram prikazana na sliki 1.

- v Mathematici narišemo diagram iz podatna naslednji način:

1. prvo podatke preberemo:

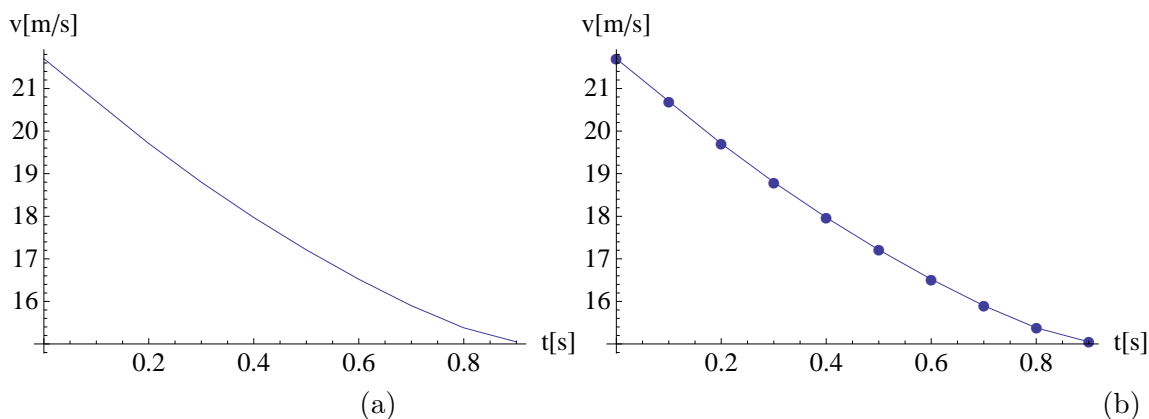
```
m = ReadList["podatki.dat", {Real, Real, Real}];
{x, y, w} = Transpose[m];
```

2. nato pa jih še narišemo: (a) točke meritev se ne vidijo, kar je morda priporočljivo ob veliki gostoti točk:

```
ListPlot[Transpose[{x, y}],
  AxesLabel -> {"t[s]", "v[m/s]"},
  Joined -> True,
  PlotRange -> All,
  LabelStyle -> {FontSize -> 14}]
```

- (b) točke meritev se vidijo:

```
ListPlot[Transpose[{x, y}],
  AxesLabel -> {"t[s]", "v[m/s]"},
  Joined -> True,
  PlotRange -> All,
  PlotMarkers -> Automatic,
  LabelStyle -> {FontSize -> 14}]
```



Slika 2: Slika narisana s programom Mathematica 6.0 po dveh pristopih (a) in (b),

Grafi pridobljeni na tak način so na slik1 1. Tako dobljena slika ni primerna za podatke, kjer je znana natančnost meritev, saj ne vsebuje intervale zaupanj.

2 Povprečna vrednost, deviacija in slučajna napaka

Rezultate meritve neke količine Y podajamo v obliki

$$y \pm \Delta y, \text{ ali } y \left(1 \pm \frac{\Delta y}{y} \right),$$

kjer y pomeni pričakovano vrednost količine Y in Δy efektivno odstopanje meritev od y . Slednje tudi imenujemo napaka meritve.

Naredimo N meritev neke statične fizikalne količine X označene s

$$X_1, \dots, X_N.$$

Te meritve so zaradi zunanjih naključnih vplivov porazdeljene okoli **iskane pričakovane** (povprečne) vrednosti \bar{X} . Razpršenost oz. odstopanje meritev od povprečja \bar{X} imenujemo deviacija σ . Ker je vzorec meritev končen lahko le približno izračunamo omenjeni količini s formulama

$$\tilde{X} = \frac{1}{N} \sum_{n=1}^N X_n \approx \bar{X}$$

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (X_n - \tilde{X})^2 \approx \sigma^2.$$

Pravtako lahko izračunamo koliko smo se zmotili pri oceni pričakovane vrednosti \bar{X} z \tilde{X} in dobimo končni rezultat meritev neke količine X

$$\bar{X} = \tilde{X} \pm \frac{\sigma}{\sqrt{N}}.$$

Pri izračunu lahko odstopanje σ zamenjamo s $\tilde{\sigma}$, saj ne naredimo bistvene napake.

3 Sistematska napaka

Meritve so opravljene s pomočjo instrumentov. Instrumenti prikazujejo merjeno količino do neke natančnosti oz. jo lahko odčitujemo le do neke natančnosti. To napako imenujemo sistematska napaka meritve.

Recimo merimo neko količino x , ki poseduje neko sistematsko napako Δx . Meritev zabeležimo kot

$$x \pm \Delta x$$

S pomočjo te količine želimo izračunamo neko funkcijo $f(x)$. Ta funkcija bo imela tako tudi imela neko napako Δf :

$$x \pm \Delta x \rightarrow f(x \pm \Delta x) \approx f(x) \pm \Delta f, \quad \Delta f = \left| \frac{df}{dx}(x) \right| \Delta x$$

V primeru, da imamo dvo-parametrično funkcijo $f(x, y)$ katere parametra sta zašumljena $x \pm \Delta x$, $y \pm \Delta y$. Napako funkcije izračunamo po naslednji formuli:

$$(\Delta f)^2 = \left(\frac{df(x, y)}{dx} \right)^2 (\Delta x)^2 + \left(\frac{df(x, y)}{dy} \right)^2 (\Delta y)^2$$

To formulo lahko posplošimo na poljubno parametrično funkcijo. V primeru enega parametra pa dobimo nazaj prejšnjo formulo.

Oglejmo si nekaj tipičnih primerov enoparametričnih funkcij in njenih napak, če imamo zašumljen parameter(-e) $x \pm \Delta x$ oz. $y \pm \Delta y$

$f(x)$	$f(x) \pm \Delta f(x)$ oz $f(x) \left(1 \pm \frac{\Delta f(x)}{f(x)} \right)$
$\frac{1}{x}$	$\frac{1}{x} \left(1 \pm \frac{\Delta x}{x} \right)$
$\frac{x}{y}$	$\frac{x}{y} \left(1 \pm \frac{\Delta x}{x} \pm \frac{\Delta y}{y} \right)$
x^n	$x^n \left(1 \pm \frac{\Delta x}{x} \right)$
$\ln(x)$	$\ln(x) \pm \frac{\Delta x}{x}$

4 Linearna regresija - prilagajanje premice podatkom

Večkrat se pojavi problem prilagajanja premice k dvodimenzionalnim podatkom. To lahko naredimo popolnoma ročno tako, da potegnemo premico na grafu približno tako, da le-ta odstopa čim manj od točk podatkov. To navadno zadostuje za delo pri praktikumu. Obstaja pa tudi bolj matematično utemeljena pot. Zaradi njene enostavnosti se le-to splača pogledati.

Imejmo N točk (x_i, y_i) , $i = 1, \dots, N$ v dveh dimenzijah skozi katere želimo potegniti premico oblike $y = Kx + R$. Vsaka točka je v y smeri obremenjena z neko napako w_i . Večja je napaka bolj nesigurna je njena pozicija v y smeri in zato je treba manj upoštevati pri potegovanju premice. Opisano lahko združimo v dve formuli

$$K = \frac{da - eb}{ca - b^2}, \quad R = \frac{ce - db}{ca - b^2}$$

kjer so uvedene konstante a , b , c , d in e definirane kot

$$a = \sum_{n=1}^N \frac{1}{w_i^2}, \quad b = \sum_{n=1}^N \frac{x_i}{w_i^2}, \quad c = \sum_{n=1}^N \frac{x_i^2}{w_i^2}, \quad d = \sum_{n=1}^N \frac{x_i y_i}{w_i^2}, \quad e = \sum_{n=1}^N \frac{y_i}{w_i^2}$$

Naredimi primer uporabe teh formul na podatkih iz tabele 1. Konstante enostano izračunamo v Mathematici z naslednjimi vrsticami:

```
m = ReadList["podatki.dat", {Real, Real, Real}];
{x, y, w} = Transpose[m];

a = Sum[1/w[[i]]^2, {i, Length[w]}];
b = Sum[x[[i]]/w[[i]]^2, {i, Length[w]}];
c = Sum[x[[i]]^2/w[[i]]^2, {i, Length[w]}];
d = Sum[x[[i]]*y[[i]]/w[[i]]^2, {i, Length[w]}];
e = Sum[y[[i]]/w[[i]]^2, {i, Length[w]}];
```

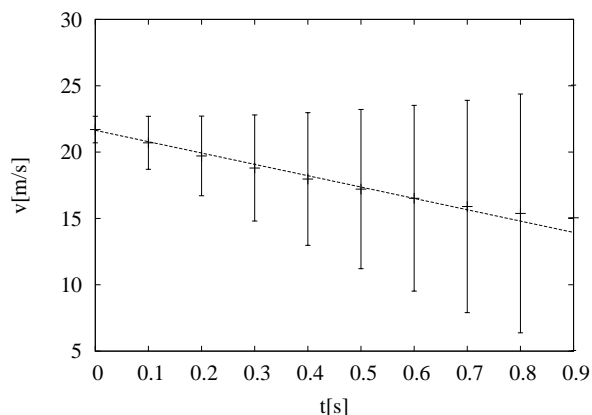
```
K = (d*a - e*b)/(c*a - b^2)
```

```
R = (c*e - d*b)/(c*a - b^2)
```

Z naslednjimi vrsticami pa narišemo premico in podatke:

```
set term post enh eps "Times-Roman" 26
set out 'graf.eps'
unset key
set xlabel 't[s]'
set ylabel 'v[m/s]'
plot 'podatki.dat' w err ps 2, -8.55*x+21.64
set out
```

Rezultat je prikazan na sliki 3. Prilaganje funkcije k podatkom z upoštevanjem sistemske napake



Slika 3: Prikaz podatkov z napakami oz. intervali zaupanja s linijo, ki se prilega se statistično optimalno prilega podatkom.

lahko opravimo prav tako v okviru programa GnuPlot z naslednjima vrsticama:

```
f(x)=K*x+R  
fit f(x) 'podatki.dat' u 1:2:3 via K,R
```