Práce s daty

2. února 2015

V tomto článku si ukážeme statistickou práci v praxi. Setkáme se s mnoha bodovými či intervalovými odhady i s různými testy. Na kraji textu máte vyznačeno, jaké pojmy a znalosti se právě používají. Student na konci předmětu Statistika by měl článku bez problému rozumět.

Zprovoznění

1. Pokud nemáte aktuální statistický balíček pro Scilab, stáhněte si jej.
2. Přihlašte se do Novellu.
3. Vytvořte si adresář na dnešní práci s reálnými daty. Doporučený název: *Tunel*.
4. Do tohoto adresáře stáhněte ze stránek soubory *data.dat* a *Nacti_data.sci*.
5. Otevřete si Scilab.
6. Otevřete si soubor *Nacti_data.sci* a nastavte v něm cesty ke svým adresářům, kde máte uložené soubory *spust_pri_startu.sce* a *data.dat*. Uložte.
7. Otevřete si soubor *spust_pri_startu.sce* a přepište ho. Zadejte absolutní cesty k adresářům *kombinatorika*, *pravděpodobnost*, *statistika*. Doplňte adresář, kde je funkce *Nacti_data.sci*. Výsledek může podle toho, jaké máte adresáře vy, vypadat např. takto:

```
clear;
getd("F:\Scilab\Funkce\kombinatorika");
getd("F:\Scilab\Funkce\pravdepodobnost");
getd("F:\Scilab\Funkce\statistika");
getd("F:\Scilab\Tunel");
```

8. Soubor *spust_pri_startu.sce* pusťte.
9. Otevřete si nový editor a začněte řešit následující úkoly. Ke každému úkolu vytvořte zvláštní program. Ukládejte je do stejného adresáře jako soubor *data.dat*.
10. Na začátku každého programu zavolejte funkci:

```
[I,O,R,H,D]=Nacti_data();
```

Zobrazení veličin

Setkáme-li se s neznámými daty, prvním krokem je seznámení se s nimi. Jednak si přečteme doprovodné informace, jadnak si je zobrazíme v množství grafů, abychom zachytili zajímavé souvislosti.

Doprovodné informace

Veličiny I, O a R znamenají intenzitu, obsazenost a rychlost ve Strahovském tunelu. Měřeny jsou v pětiminutových intervalech po dobu čtyř týdnů. Začátek je v pondělí pět minut po půlnoci.

Veličina I znamená počet projetých aut za pět minut. Je to přirozené číslo.

Veličina R znamená průměrnou rychlost aut, které tunelem projely během pěti minut. Je to reálné číslo v $^{\mathrm{km}}/_{\mathrm{h}}$.

Veličina O znamená obsazenost a je to poměr $O = \frac{I}{R}$. Jednotkou je $\frac{^{\mathrm{ks}}/_{5\,\mathrm{min}}}{^{\mathrm{km}}/_{\mathrm{h}}} = \frac{h}{5\,min} \cdot \frac{ks}{km} = 12 \cdot \frac{ks}{km} = {}^{\mathrm{ks}}/_{83\,\mathrm{m}}$.Veličina tedy znamená průměrný počet aut na úseku délky 83 metrů.

Veličina H znamená hodinu. Je to desetinné číslo z intervalu $(0, 24\rangle$.

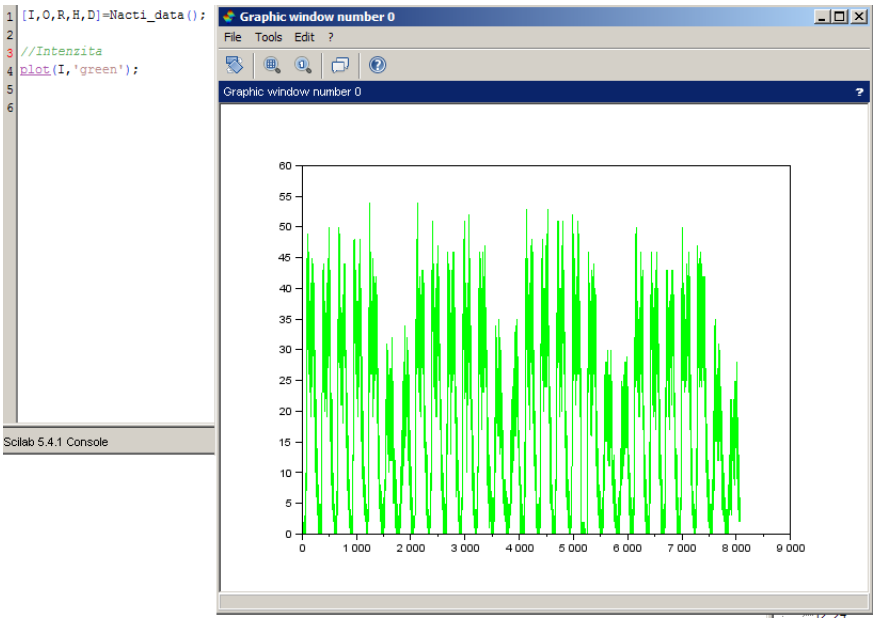
Veličina D označuje den v týdnu. Je to celé číslo od 1 do 7.

Základní grafy

Z množství možných grafů si ukažme tyto:

Intenzita

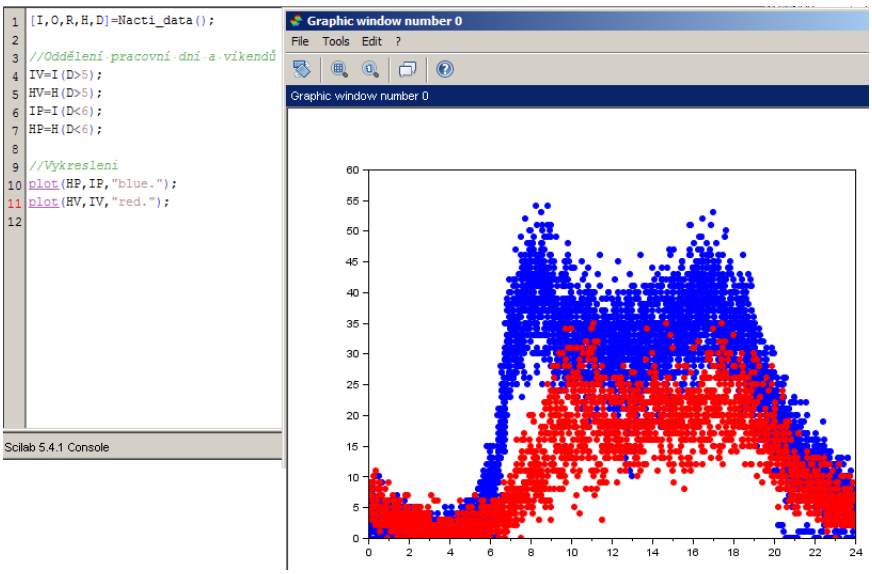
Zobrazíme si intenzitu I. Pokud v příkazu *plot* na ose x nenapíšeme nic, bude na ose x pořadí dat.



Vidíme denní špičky a vidíme rovněž nižší intenzitu během víkendů.

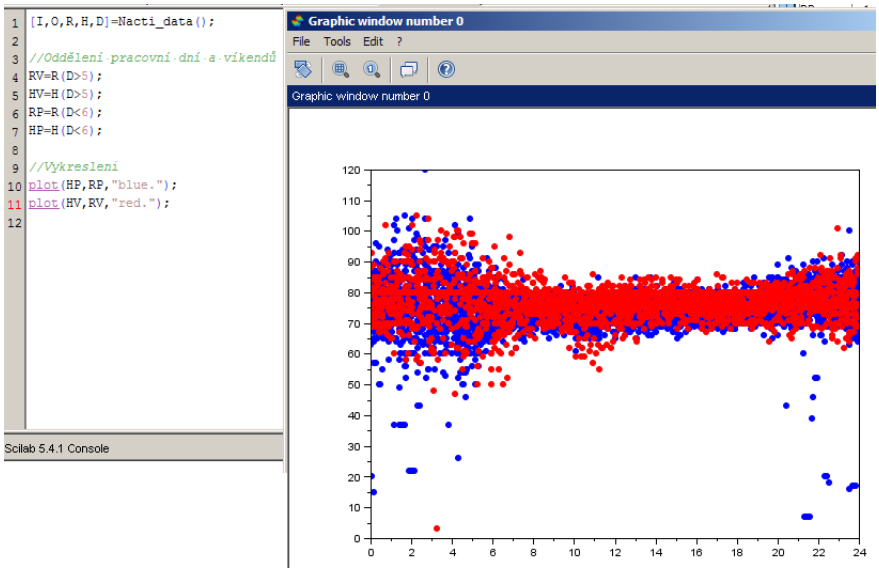
Okamžitě nás může napadnout, zda je doprava během různých pracovních / víkendových dní stejná.

Intenzita proti hodině



Zde opět vidíme, že intenzita má o víkendech jiný vývoj než během pracovních dní. Můžeme si např. všimnout, že ranní špička o víkendu nastupuje později.

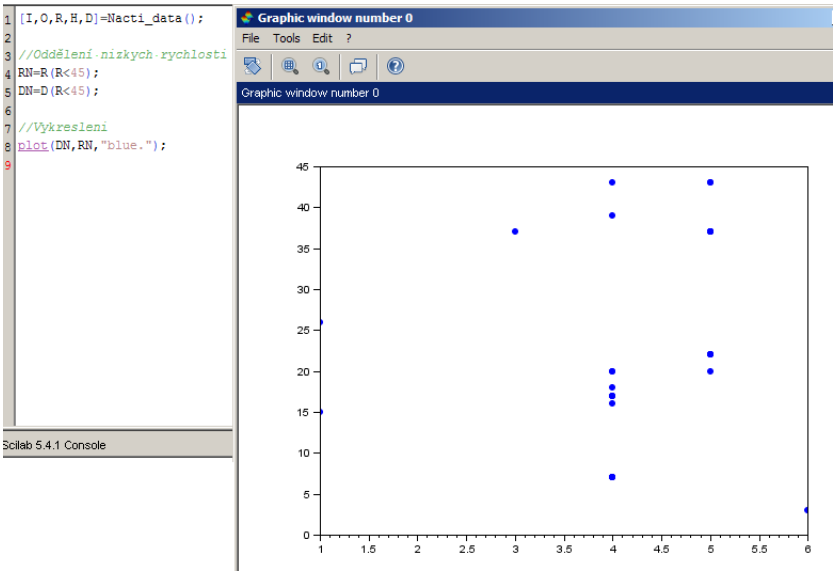
Rychlost proti hodině



Zde můžeme vidět zajímavou věc: V noci mají rychlosti větší rozptyl. Jak tento jev vysvětlit?

Dále pak vidíme, že velmi nízké rychlosti se vyskytují téměř pouze během pracovních dní v noci. Jak vysvětlit tento jev? Vyskytují se tyto velmi nízké rychlosti během všech pracovních dní?

Velmi nízké rychlosti proti dni v týdnu



Vidíme, že nízké rychlosti se vyskytují téměř pouze ve čtvrtek a pátek. To je velmi zajímavé zjištění. Možná se jedná o pravidelnou údržbu?

Podobně bychom mohli pokračovat, abychom objevily zajímavosti, které jsou v datech skryty.

Řídký dopravní proud a Poissonovské rozdělení

Poissonovo rozdělení.

Jistě jste se učili nebo budete učit, že pokud je dopravní proud řídký, tzn. netvoří se žádné kolony ani "buřtíky", měl by mít počet aut za jednotku času (tj. v našem případě za pět minut) Poissonovo rozdělení. Prozkoumejme, zda toto platí pro intenzity mezi druhou a třetí hodinou v noci během pracovních dnů.

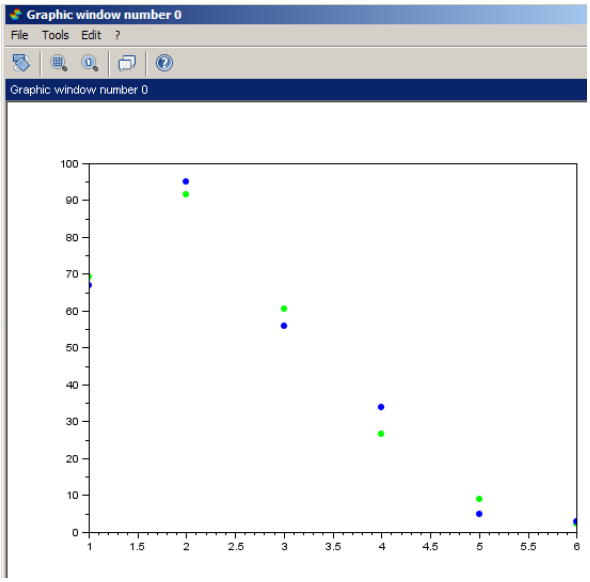
Použijeme Chí-kvadrát test dobré shody.

Chí-kvadrát test dobré shody

Tedy testujeme nulovou hypotézu: Mezi druhou a třetí hodinou v noci pracovního dne má intenzita dopravy ve Smíchovském tunelu Poissonovo rozdělení.

Nulová hypotéza.

```
[I,O,R,H,D]=Nacti_data();

//Vyber dat
I1=I(H>=2 & H<=3 & D<6);

//Bodovy odhad Lambdy
Lambda=mean(I1);
N=length(I1);

//Oi-kolikrat se dana intenzita vyskytovala
Oi=[];
for n=0:5
    In=I1(I1==n);
    K=length(In);
    Oi=[Oi,K];
end

//Ei-kolikrat by se mela vyskytovat dle Poissona
[D,Q]=cdfpoi("PQ",0:5,Lambda*ones(1,6));
D1=[0,D];
D1($)=[];  //Dolar znamena posledni
P=D-D1;
Ei=N*P;

plot(Ei,'g.');
plot(Oi,'b.')

//Chi-kvadrat test dobre shody
[p,stat,kvantil]=chisquare_test(0.01,Oi,Ei);
disp('p-hodnota:');
disp(p);
```



Funkce vyžaduje zadat nějakou hladinu významnosti, ale ve skutečnosti na ní nezáleží, pokud nás zajímá p-hodnota.

Teoretické hodnoty jsou zeleně, skutečné modře. Funkce *cdfpoi* je distribuční funkcí Poissonova rozdělení. Parametr Poissonova rozdělení jsme bodově odhadli: $\lambda = 1,3231$. Vyšla nám krásná p-hodnota $0,4975906$, takže hypotézu o tom, že intenzita v daný čas má Poissonovo rozdělení s danou λ rozhodně zamítat nebudeme.

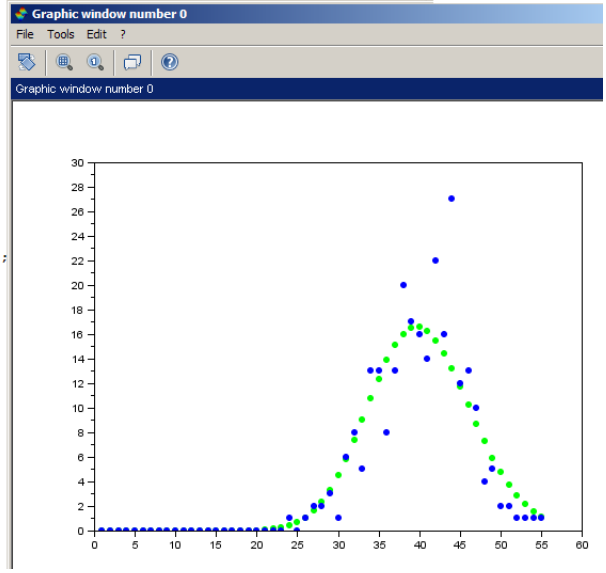
Výpočet pravděpodobnostní funkce z distribuční funkce.

Zkusme stejný test, tentokrát pro čas 7:30 až 8:30 během pracovního dne:

Bodový odhad parametru λ.

P-hodnota

```
[I,O,R,H,D]=Nacti_data();

//Vyber dat
I1=I(H>=7.5 & H<=8.5 & D<6);

//Bodovy odhad Lambdy
Lambda=mean(I1);
N=length(I1);

//Oi-kolikrat se dana intenzita vyskytovala
Oi=[];
for n=0:max(I1)
    In=I1(I1==n);
    K=length(In);
    Oi=[Oi,K];
end

//Ei-kolikrat by se mela vyskytovat dle Poissona
[D,Q]=cdfpoi("PQ",0:max(I1),Lambda*ones(1,max(I1)+1));
D1=[0,D];
D1($)=[];  //Dolar znamena posledni
P=D-D1;
Ei=N*P;

//Vykresleni
plot(Ei,'g.');
plot(Oi,'b.');

//Chi-kvadrat test dobre shody
//Slouceni prvnich 30 a poslednich hodnot
Ei=[sum(Ei(1:30)),Ei(31:49),sum(Ei(50:$))];
Oi=[sum(Oi(1:30)),Oi(31:49),sum(Oi(50:$))];
[p,stat,kvantil]=chisquare_test(0.99,Oi,Ei);
disp('p-hodnota:');
disp(p);
```



Aby byl Chí-kvadrát test kvalitní, doporučuje se, aby teoretická četnost pro každou kolonku byla alespoň pět. Proto jsme tentokrát prvních 30 hodnot a poslední hodnoty od 50 sloučili do jedné.

Bodový odhad parametru $\lambda = 39.1846$, p-hodnota vyšla $0,0474449$. To je mezní hodnota. Může znamenat, že nastal onen jeden případ z 20, kdy p-hodnota bude nižší než 5% i pokud je nulová hypotéza v pořádku, může to ale také znamenat, že během ranních špiček nemůžeme dopravní proud považovat za Poissonovský.

Abychom mezi těmito dvěma alternativami rozhodli, proveďme test ještě pro čas 8:30 až 9:30. P-hodnota tentokrát po sloučení příslušných krajních hodnot vyjde $0,3386562$. Zdá se tedy, že i během ranních špiček můžeme dopravní proud považovat za Poissonovský.

V tunelu se tedy patrně nevyskytují kolony ani při největších dopravních intenzitách. Koneckonců intenzita 60 aut za 5 minut odpovídá v průměru jednomu autu za pět sekund. A to není tolik.Kapacita tunelu je tedy dostatečná.

Je intenzita během různých pracovních dní stejná?

Předpokládejme, že změřená intenzita má během pracovního dne intenzitu $I(t) = f(t) + e(t)$, kde $f(t)$ je stále stejná funkce a $e(t)$ je náhodný šum. O tomto šumu předpokládáme, že má nulovou střední hodnotu a jeho hodnoty jsou pořadově nezávislé.

Pořadová nezávislost dat

Pokud odečteme intenzity dvou různých dní, získáme: $I_1(t) - I_2(t) = f_1(t) + e_1(t) - f_2(t) - e_2(t) = e_1(t) - e_2(t)$, což je šum, který má střední hodnotu také nula a jehož hodnoty jsou také pořadově nezávislé. (Rozptyl je součet rozptylů šumů $e_1(t)$ a $e_2(t)$.)

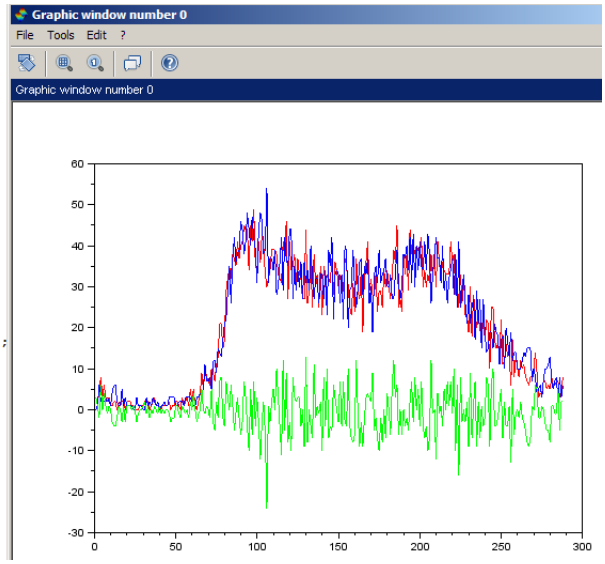
Věta o sčítání náhodných veličin

Testujme tedy, jestli má rozdíl intenzit dvou různých dní nulovou střední hodnotu a jestli je pořadově nezávislý. Nejprve testujme první a druhé pondělí:

Test střední hodnoty

Test pořadové nezávislosti

```
[I,O,R,H,D]=Nacti_data();

//Nacteni dvou dni
I1=I(1:288);
I2=I(2017:2304);

D=I1-I2;

//Test Mi=0
n=length(D);
SH=mean(D);
R=variance(D);
[p,T,kv]=t_test(0.99,0,n,SH,R,'o');
disp('Stredni hodnota - 0 - ---- p-hodnota');
disp(p);

//Test poradove nezavislosti
[p,T,kv]=ordinal_test(0.99,D);
disp('Poradova nezavislost - ---- p-hodnota');
disp(p);

//Vykresleni
plot(I1,"red");
plot(I2,'blue');
plot(D,'green')
```

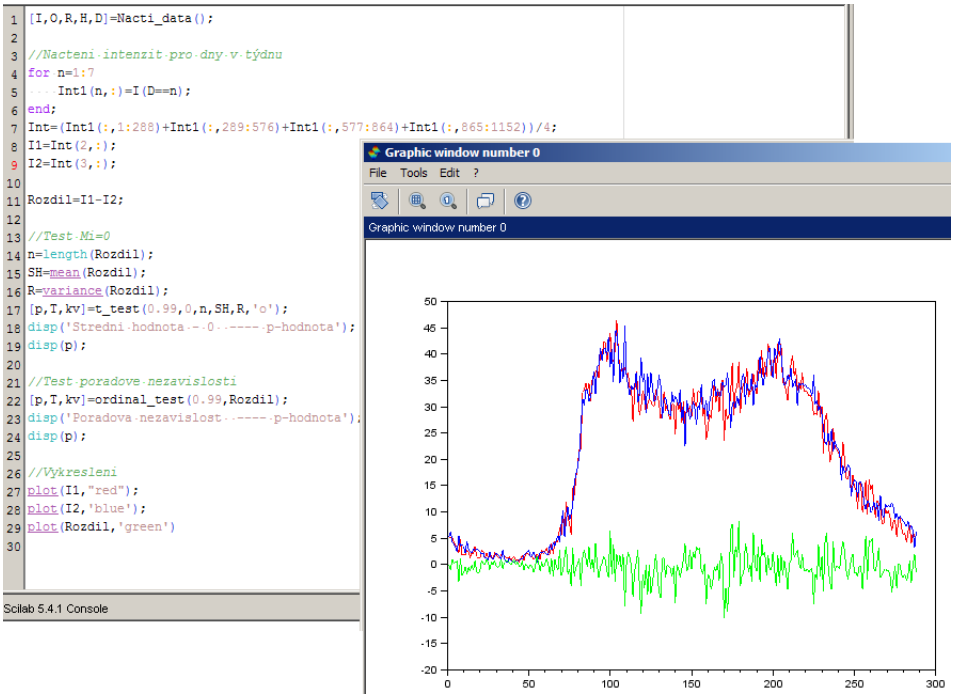


Intenzita je červeně a modře, rozdíl zeleně. Pro test nulové střední hodnoty vyšla p-hodnota 0.0889812, pro test pořadové nezávislosti 0.4066727. První p-hodnota je poměrně nízká, možná to signalizuje, že během jednoho pondělí došlo k nějaké nestandardnosti. Ale i tak oba testy na hladině významnosti 5% prošly.

Šum tedy můžeme považovat za pořadově nezávislý a s nulovou střední hodnotou. Abychom však mohli šum považovat za bílý, měl by být ještě nezávislý na čase. A to není, neboť vidíme, že v noci má menší rozptyl než během dne. Bílý šum

Abychom vyloučili nestandardnosti během jednotlivých dní, budeme pracovat s průměrnými intenzitami za všechny čtyři pondělky, úterky, atd.

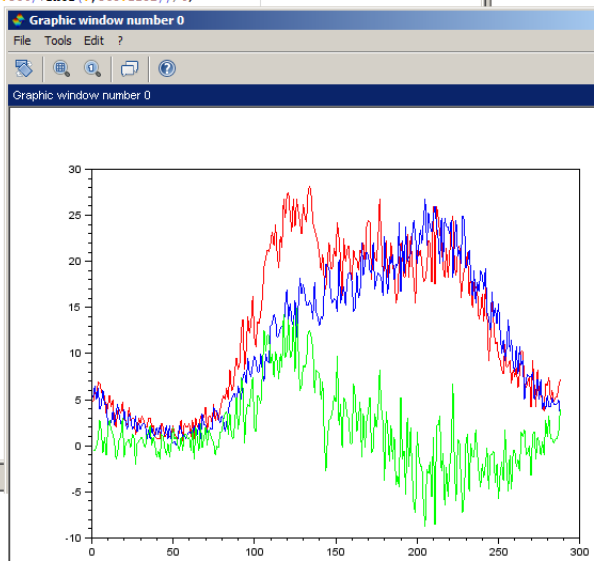
Testujme tedy nulovost střední hodnoty a pořadovou nezávislost rozdílového šumu např. pro úterky a středy:



A zde pro soboty a neděle:

```
[I,O,R,H,D]=Nacti_data();

//Nacteni intenzit pro dny v tydnu
for n=1:7
    Int1(n,:)=I(D==n);
end;
Int=(Int1(:,1:288)+Int1(:,289:576)+Int1(:,577:864)+Int1(:,865:1152))/4;
I1=Int(6,:);
I2=Int(7,:);

Rozdil=I1-I2;

//Test Mi=0
n=length(Rozdil);
SH=mean(Rozdil);
R=variance(Rozdil);
[p,T,kv]=t_test(0.99,0,n,SH,R,'o');
disp('Stredni hodnota - 0 - ---- p-hodnota');
disp(p);

//Test poradove nezavislosti
[p,T,kv]=ordinal_test(0.99,Rozdil);
disp('Poradova nezavislost - ---- p-hodnota');
disp(p);

//Vykresleni
plot(I1,"red");
plot(I2,'blue');
plot(Rozdil,'green')
```

Scilab 5.4.1 Console



Uveďme si tabulku pro p-hodnoty testu nulovosti střední hodnoty a pořadové nezávislosti pro jednotlivé dni v týdnu:

	Po	Út	St	Čt	Pá	So	Ne
Po		0,901	0,028	0,435	0,617	1.115D-41	7.595D-43
		0,016	0,005	0,0001	0,00008	1.557D-37	2.482D-47
Út			0,027	0,331	0,519	3.781D-42	1.751D-45
			0,278	0,017	0,0003	5.956D-35	5.474D-43
St				0,004	0,0095	1.844D-46	5.030D-48
				0,00002	0,0091	1.825D-32	7.922D-49
Čt					0,770	8.811D-40	3.917D-42
					0,00002	5.474D-43	2.482D-47
Pá						3.278D-43	8.109D-44
						3.262D-40	6.831D-52
So							1.156D-09
							1.883D-16

Vidíme, že výrazně jiný průběh má intenzita v pracovní dny, v sobotu a v neděli. Kupodivu testy naznačují, že i mezi jednotlivými pracovními dny jsou statisticky významné rozdíly. Např. pro hladinu významnosti 1% můžeme za stejné považovat průběhy v pondělí a úterý, v úterý a ve středu, v úterý a ve čtvrtek, ale už ne ve středu a čtvrtek. Na hladině významnosti 5% už musíme pro každý den v týdnu uvažovat alespoň trochu jiné rozdělení.

Hladina významnosti

Je rozdělení intenzity opravdu Poissonovské?

V jedné z předchozích kapitolek jsme si ukázali, že můžeme intenzitu v celém rozsahu, od noci až po denní špičky, považovat za Poissonovskou. Rozptyl by tedy měl být, stejně jako střední hodnota, roven λ. A střední hodnota $I(t)$ je v modelu $I(t) = f(t) + e(t)$ rovna $f(t)$, neboť střední hodnota šumu je nulová. Tedy rozptyl šumu $e(t)$ by měl být roven také $f(t)$.

Střední hodnota a rozptyl Poissonova rozdělení

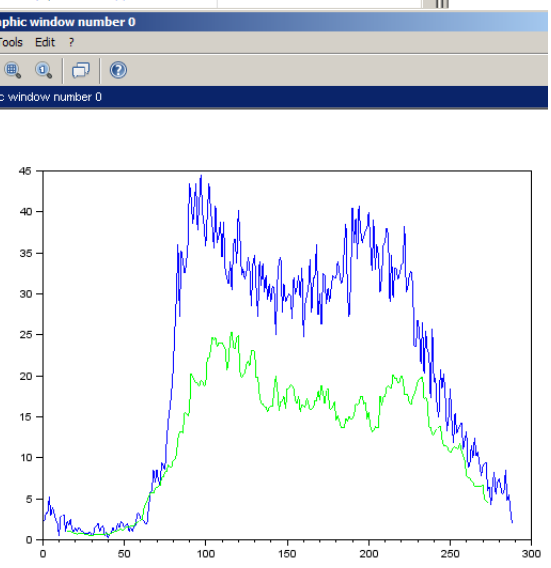
Ověřme si to vykreslením do grafu. Vybereme si pondělní hodnoty a vykreslíme si modře funkci $f(t)$, kterou odhadujeme průměrem ze čtyř pondělků:

$$\widehat{f(t)} = \frac{I_1(t) + I_2(t) + I_3(t) + I_4(t)}{4}.$$

Zeleně vykreslíme rozptyl šumu $e(t) = I_1(t) - f(t)$, kde $I_1(t)$ jsou intenzity za první pondělí. Rozptyl počítáme vždy z třiceti sousedních hodnot:

```
[I,O,R,H,D]=Nacti_data();

//Nacteni intenzit pro dny v týdnu
for n=1:7
    Int1(n,:)=I(D==n);
end;
Int=(Int1(:,1:288)+Int1(:,289:576)+Int1(:,577:864)+Int1(:,865:1152))/4;
I1=Int1(1,1:288);
f=Int(1,:);

e=I1-f;

//Bodovy odhad rozptylu
kolik=30;
for k=1:length(e)-(kolik-1)
    r(k)=variance(e(k:k+(kolik-1)));
end;
rx=fix(kolik/2):length(r)+fix(kolik/2)-1;

//Test Mi=0
n=length(e);
SH=mean(e);
R=variance(e);
[p,T,kv]=t_test(0.99,0,n,SH,R,'o');
disp('Stredni hodnota - 0 ----- p-hodnota');
disp(p);

//Test poradove nezavislosti
[p,T,kv]=ordinal_test(0.99,e);
disp('Poradova nezavislost ----- p-hodnota');
disp(p);

//Vykresleni
plot(f,'blue');
plot(rx,r,'green');
```

Scilab 5.4.1 Console



Co to? Rozptyl je výrazně menší, než bychom očekávali!

Bylo snad něco v našich úvahách špatně?

Ano! Zapomněli jsme na to, že zatímco funkce $f(t)$ je nenáhodná, my ji jen přibližně odhadujeme náhodnou veličinou $\widehat{f(t)} = \frac{I_1(t)+I_2(t)+I_3(t)+I_4(t)}{4}$.

Protože šum

$$e(t) = I_1(t) - \widehat{f(t)} = I_1(t) - \frac{I_1(t) + I_2(t) + I_3(t) + I_4(t)}{4} = \frac{3}{4}I_1(t) - \frac{1}{4}I_2(t) - \frac{1}{4}I_3(t) - \frac{1}{4}I_4(t),$$

měl by být rozptyl:

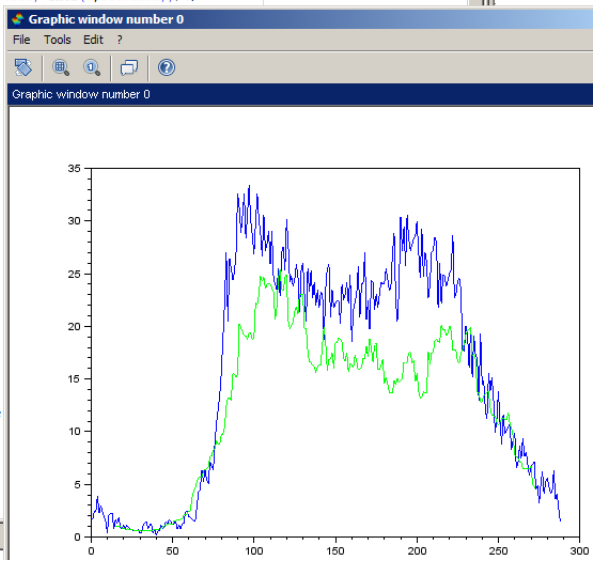
Věty o sčítání náhodných veličin a násobení náhodné veličiny konstantou

$$\sigma^2 = \left(\frac{3}{4}\right)^2 f(t) + \left(\frac{1}{4}\right)^2 f(t) + \left(\frac{1}{4}\right)^2 f(t) + \left(\frac{1}{4}\right)^2 f(t) = \frac{3}{4} f(t).$$

Porovnejme tedy znovu rozptyl nikoliv s $f(t)$, ale s $\frac{3}{4}f(t)$:

```
[I,O,R,H,D]=Nacti_data();

//Nacteni intenzit pro dny v tydnu
for n=1:7
    Int1(n,:)=I(D==n);
end;
Int=(Int1(:,1:288)+Int1(:,289:576)+Int1(:,577:864)+Int1(:,865:1152))/4;
I1=Int1(1,1:288);
f=Int(1,:);

e=I1-f;

//Bodovy odhad rozptylu
kolik=30;
for k=1:length(e)-(kolik-1)
    r(k)=variance(e(k:k+(kolik-1)));
end;
rx=fix(kolik/2):length(r)+fix(kolik/2)-1;

//Test Mi=0
n=length(e);
SH=mean(e);
R=variance(e);
[p,T,kv]=t_test(0.99,0,n,SH,R,'o');
disp('Stredni hodnota - 0 ---- p-hodnota');
disp(p);

//Test poradove nezavislosti
[p,T,kv]=ordinal_test(0.99,e);
disp('Poradova nezavislost ---- p-hodnota');
disp(p);

//Vykresleni
plot(0.75*f,'blue');
plot(rx,r,'green');
```

Scilab 5.4.1 Console

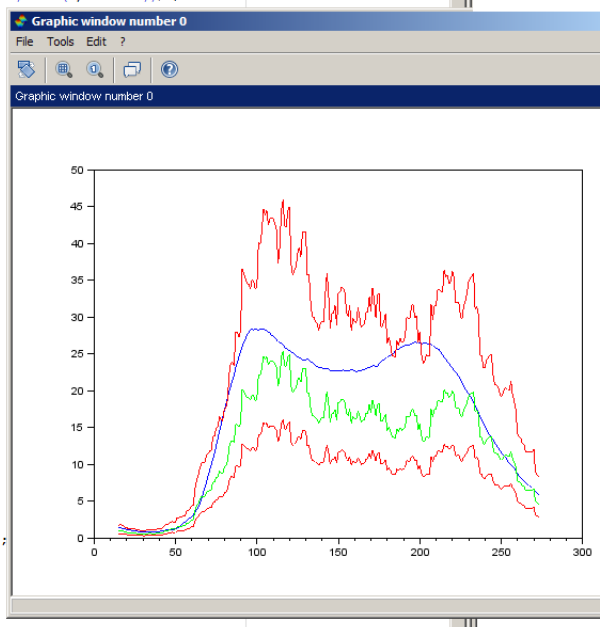


Vidíme, že zatímco pro nižší intenzity si grafy pěkně odpovídají, (pomalejší náběh rozptylu je způsoben tím, že se hodnoty počítají z 30 sousedních hodnot), pro vyšší hodnoty je již rozptyl poněkud nižší. Tento efekt může být způsoben buď tím, že už vznikají "buřtíky" nebo řízením, např. semafory.

Intervalový odhad rozptylu

Můžeme si ale také myslet, že nižší průběh rozptylu je pouze náhodný. Vždyť bodový odhad rozptylu je velmi nepřesný, jak zjistíme, pokud si vykreslíme intervalový odhad rozptylu. Vykreslujeme 95% oboustranný interval spolehlivosti (červeně). Abychom kompenzovali pomalejší náběh rozptylu (který je počítán z 30 sousedních hodnot), použijeme místo funkce f její průměr z 30 sousedních hodnot. Nakonec provedeme test podílu na podíl hodnot, které vybíhají z 95% intervalu spolehlivosti:

```
[I,O,R,H,D]=Nacti_data();

//Nacteni intenzit pro dny v týdnu
for n=1:7
    Int1(n,:)=I(D==n);
end;
Int=(Int1(:,1:288)+Int1(:,289:576)+Int1(:,577:864)+Int1(:,865:1152))/4;
I1=Int1(1,1:288);
f=Int(1,:);

e=I1-f;

//Bodovy odhad rozptylu
kolik=30;
for k=1:length(e)-(kolik-1)
    r(k)=variance(e(k:k+(kolik-1)));
end;
rx=fix(kolik/2):length(r)+fix(kolik/2)-1;

//Intervalovy odhad rozptylu
[a,b]=var_int(0.95,kolik,r,'o');

//Prumerne f
for k=1:length(e)-(kolik-1)
    ff(k)=mean(f(k:k+(kolik-1)));
end;

//Podil hodnot mimo 95_ interval
fmimo=ff(0.75*ff>b | 0.75*ff<a);
nplus=length(fmimo);
ncelkem=length(ff);
podil=nplus/ncelkem;
disp('Podil hodnot mimo 95_ IS');
disp(podil);

//Test podilu
[p,T,kv]=prop_test(0.99,0.05,podil,ncelkem,'o');
disp('Test podilu ----- p-hodnota');
disp(p);

//Vykresleni
plot(rx,0.75*ff,'blue');
plot(rx,r,'green');
plot(rx,a,'red');
plot(rx,b,'red');
```



Vidíme, že naše křivka zprůměrovaných $\frac{3}{4}f(t)$ je opravdu většinou v mezích 95% intervalu. Mimo se ocitla pouze 13 krát, což je 5,02% případů. U 95% intervalu spolehlivosti bychom přitom očekávali, že křivka bude mimo zhruba v 5% případů. To je výborná shoda, jak nám napovídá i p-hodnota 0,9886 v provedeném testu podílu. Nezamítáme tedy nulovou hypotézu, že mimo 95% interval spolehlivosti je 5% hodnot, tedy ani hypotézu, že rozptyl šumu je roven $\frac{3}{4}f(t)$. Tento vztah byl vyvozen z předpokladu, že rozdělení intenzit je Poissonovské. Nezamítáme tedy ani tento předpoklad.

Test podílu