Cvičení 7

Úkol: generování dat dle rozdělení, vykreslení rozdělení psti, odhad rozdělení dle dat, bodový odhad parametrů, centrální limitní věta, balíček Distfun, normalizace

Docházka a testík - 15 min.

Distfun 10 min.

Zapnutí Scilabu, seznámení s balíčkem Distfun (Atoms, Data Analysis and Statistic).

Tvar funkcí:

```
distfun_xy(...),
```

kde x je označení rozdělení: bino, poiss, norm, exp, ...

y je koncovka, co chci dělat:

- rnd - generování dat
- pdf - rozdělení psti
- cdf - distribuční fce
- inv - kvantily

Help k funkci najdu tak, že na ni kliknu pravým tlačítkem a najdu help. Tím zjistím, jak psát parametry.

Úloha 1

Generování dat

Generujte 10000 hodnot rovnoměrného rozdělení mezi 0 a 1. Toto opakujte 20-krát. Příslušné hodnoty oněch 20-ti rozdělení sečtěte. Součet pojmenujte Y. Data Y uložte do souboru.

```
mode(0);
y=zeros(1,10000);
for n=1:20;
  x=rand(1,10000,'uniform');
  y=y+x;
end
save Data y
```

Analýza dat

Nový soubor: Nahrajte data Y a vykreslete histogram. Jaké rozdělení nejlépe odpovídá datům Y? Proč? (Normální, centrální limitní věta)

```
mode(0);
load Data y;
histplot(40,y);
```

Jaká konkrétní gaussovka nejlépe odpovídá datům? Proveďte bodový odhad parametrů metodou momentů. Vypište výběrovou střední hodnotu, rozptyl a směrodatnou odchylku. Gaussovku s nalezenými parametry vykreslete k histogramu červeně.

```
stred_h=mean(y)
rozptyl=variance(y)
sm_odch=sqrt(rozptyl)
X=5:0.1:15;
Y=distfun_normpdf(X,stred_h,sm_odch);
plot(X,Y,'red');
```

Analýza rozdělení

Proč vyšla výběrová střední hodnota blízká číslu 10? (Při sčítání n. veličin se sčítají stř. hodnoty. Stř. h. rovnoměrného rozdělení od 0 do 1 je: $\frac{a+b}{2} = \frac{1}{2}$.)

Proč vyšel výběrový rozptyl blízký číslu $\frac{20}{12}$? (Při sčítání n. veličin se sčítají i rozptyly. Rozptyl rovnoměrného rozdělení od 0 do 1 je: $\frac{(b-a)^2}{12} = \frac{1}{12}$.)

Nyní připomeneme, že ve statistice můžeme mít čtyři různé situace:

známe data	kompletní = základní soubor
	nekompletní = výběr
známe rozdělení	diskrétní
	spojité

Nyní se můžeme od situace, kdy známe data, přenést do situace, kdy známe rozdělení. Tedy se přesuneme z druhého na čtvrtý řádek.

Úplně správná gaussovka má stř. h. 10 a rozptyl $\frac{20}{12}$. Vykreslete ji do stejného obrázku modře.

```
Z=distfun_normpdf(X,10,sqrt(20/12));
plot(X,Z,'blue');
```

Vidíme, že jsme správnou křivku odhadli velice přesně. Však máme 10000 dat.

Intervalový odhad dat

Najděte interval, kam by mělo padnout 90% nejpravděpodobnějších dat y. Jak to udělat? Uřízneme z každé strany 5%. K tomu potřebujeme 5%-ní a 95%-ní kvantil. Vrátíme se do situace, kdy známe jen data a budeme pracovat s červenou křivkou.

```
KV05=distfun_norminv(0.05,stred_h,sm_odch)
KV95=distfun_norminv(0.95,stred_h,sm_odch)
```

Spočetli jsme kvantily. Je pst, že hodnota y padne mezi ně, opravdu 90%? Ověřte pomocí distribuční funkce.

```
F1=distfun_normcdf(KV05,stred_h,sm_odch)
F2=distfun_normcdf(KV95,stred_h,sm_odch)
P=F2-F1
```

Statistický balíček

Nainstalujeme. Seznámíme je se statistickým balíčkem. Na Pavliných stránkách otevřeme návod k balíčku.

Pracovní adresář musíme otevřít v adresáři, kde je balíček nainstalován!!! Tedy musíme vidět podadresáře help, kombinatorika, pravdepodobnost, statistika a soubor „spust_pri_startu“.

Nejprve musíme spustit soubor „spust_pri_startu“.

Intervalový odhad střední hodnoty

Najděte 90%-ní interval spolehlivosti pro skutečnou střední hodnotu μ. Rozptyl neznáme, jen odhadujeme z dat. Proto t-int.

```
[a,b]=t_int(stred_h,rozptyl,10000,'o',0.1)
```

My víme, že správná hodnota μ je přesně 10. Jaká je pst, že 10 v našem intervalu vůbec nebude? Stalo se to někomu? Kdyžtak to můžeme zkusit několikrát. (Pst je 10% - Máme 90%-ní interval.)

Intervalový odhad rozptylu

Samostatně udělejte 90%-ní interval spolehlivosti pro rozptyl.

```
[a,b]=var_int(rozptyl,10000,'o',0.1)
```

Trefili jste hodnotu $\frac{20}{12} = 1,667$? Jaká je pst, že správnou hodnotu mineme? (10%) Co proti tomu můžeme udělat? (Dát víceprocentní interval.)

Test střední hodnoty

Test a intervalový odhad je téměř totéž. Jen je otázka jinak formulována.

Proveďte na datech y test, že střední hodnota $\mu = 10$. Test proveďte na hladině významnosti $\alpha = 0,1$.

Nolová hypotéza: $\mu = 10$.

```
[p,T,kvantil]=t_test(10,stred_h,rozptyl,10000,'o',0.1)
```

Výsledkem testu je p-hodnota. Pokud je p-hodnota menší než α, pak nulovou hypotézu zamítáme. Musíme však uvést, že máme x-procentní pravděpodobnost, že se mýlíme. Ono x je právě p-hodnota.

Tedy např. $\alpha = 0,1$ a p-hodnota = 0,03. Tedy zamítáme hypotézu, že $\mu = 10$, máme však 3%, že se mýlíme.

Obvykle se to formuluje takto: „Nulovou hypotézu na hladině významnosti 10% zamítáme.“ Nebo: „Získali jsme p-hodnotu 3%, tedy nulovou hypotézu zamítáme.“

Pokud je p-hodnota větší než α, pak nulovou hypotézu nezamítáme. To znamená, že data nejsou v tak příkrém rozporu s nulovou hypotézou, abychom ji byli nuceni zamítnout. Nulová hypotéza může, ale nemusí platit. Nevíme.

Zajímavějším výsledkem je rozhodně zamítnutí nulové hypotézy.

Všimněte si, že nulovou hypotézu zamítám právě tehdy, když se mi správná hodnota 10 nevejde do odhadnutého intervalu. Obojí se děje v 10% případů. To je právě ta podobnost mezi 90%-ním intervalem a testem na 10%-ní hladině významnosti.

Test rozptylu

Samostatně proveďte test rozptylu na hladině významnosti $\alpha = 0,1$. Nulová hypotéza: $\sigma^2 = \frac{20}{12}$.

Pozor! V nápovědě je chyba! Správné pořadí parametrů je následující:

```
[p,T,kvantil]=var_test(20/12,rozptyl,10000,'o',0.1)
```

Normování

Nový soubor.

Nahrajte data y. Odečtěte od nich jejich střední hodnotu. Výsledek podělte jejich směrodatnou odchylkou. Jaké rozdělení a s jakými parametry dostanete? (N(0,1))

Vykreslete histogram i hustotu psti.

```
mode(0)
load Data y;
y=y-mean(y);
y=y/stdev(y);
histplot(40,y);
X=-3:0.01:3;
Y=distfun_normpdf(X,0,1);
plot(X,Y);
```

Úloha 2

Generování dat

Vygenerujte 1001 dat s rovnoměrným rozdělením od 0 do 100. Můžeme si představovat, že to jsou např. auta na 100 km silnice.

Zjistěte jejich rozestupy. Tj. data srovnejte a od následujícího odečtěte předchozí. Tyto rozestupy nazvěte R.

Analýza dat

Vykreslete histogram. Jaké rozdělení nejlépe popisuje data? (Exponenciální - Poissonovský dopravní proud má exponenciální rozdělení rozestupů.)

Bodově odhadněte parametr a příslušnou hustotu psti vykteslete do histogramu.

```
mode(0);
x=distfun_unifrnd(0,100,1,1001);
x=gsort(x);
for n=1:1000
  R(n)=x(n)-x(n+1);
end
histplot(30,R);
D=mean(R)
X=0:0.01:0.7;
Y=distfun_exppdf(X,D);
plot(X,Y);
```