Statistical toolbox

```
pdf ('jméno',parametry, rozměry)
cdf ('jméno', parametry, rozměry)
random('jméno', parametry, rozměry)
```

- Diskrétní rozdělení pravděpodobnosti
  - Alternativní
  - Binomické 'Binomial' n,p
  - Geometrické 'Geometric' p
  - Poissonovo 'Poisson' λ
- Spojité rozdělení pravděpodobnosti
  - Rovnoměrné 'Uniform' a, b
  - Normální 'Normal' μ, σ
  - Exponenciální 'Exponential' λ

Statistical toolbox

```
random('jméno', parametry, rozměry)
```

- **Diskrétní rozdělení pravděpodobnosti**

  'jméno' parametry

- 'Binomial' n,p

```
y=random('Binomial',n,p,500,1);
y=binornd(n,p,500,1)
```

  - 'Geometric' p

```
y=random('Geometric',p,500,1);
y=geornd(p,500,1);
```

  - 'Poisson' λ

```
y=random('Poisson',lambda,500,1);
y=poissrnd(lambda,500,1);
```

Statistical toolbox

```
random('jméno', parametry, rozměry)
```

- **Spojité rozdělení pravděpodobnosti**
  'jméno' **parametry**
- 'Uniform' a, b

```
y = random('Unif',a,b,500,1)
y = unifrnd(a,b,500,1)
```

  - 'Normal' μ, σ

```
y = random('Normal',nu,sigma,500,1);
y = normrnd(nu,sigma,500,1);
```

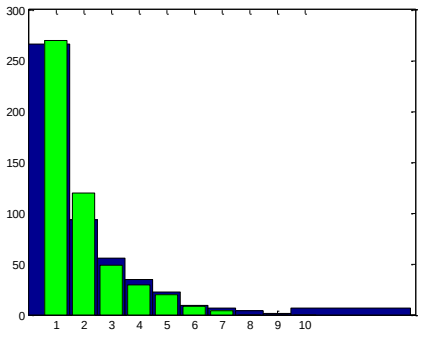
  - 'Exponential' λ

```
y=random('Exponential',1/lambda,500,1);
y=exprnd(1/lambda,500,1);
```

Geometrické x Exponenciální

- `yexp=random('Exponential',1/p,500,1);`
- `ygeo=random('Geometric',p,500,1)+1;`

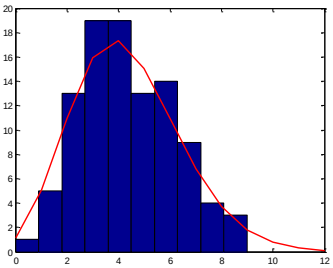
```
nbins=10;
hist(yexp,1:nbins)
hold on
freq = hist(ygeo,1:nbins);
bar(1:nbins,freq,'g')
hold off
```



Fitting Distribution

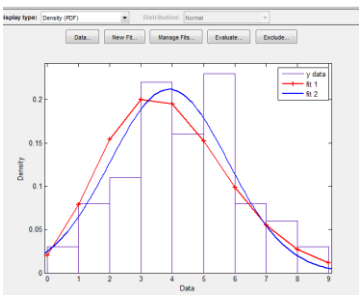
```
[lambda_est,lambda_ci]=poissfit(y) % E(y)=lambda
histfit(y,'','poisson')

[par_est,par_ci]=expfit(y);
[mu,sigma]=normfit(y);
[a,b]=unifit(y);
[p_est,p_ci]= binofit(y,n)
```



fitdist(y,'Exponential') stejný výstup jako dfittool

dfittool(y)



Testování hypotéz

Biolog, Statistik, Matematik a Informatik na safari. Zastaví džíp a pozorují dalekohledem.
Biolog "Podívejte se! Stádo zeber! A mezi nimi bílá zebra! To je fantastické! " "Existují bílé zebry! Budeme slavní!"
Statistik: "To není významné. Platí pouze, že hypotézu, že bílé zebry neexistují nemůžeme zamítnout!"
Matematik: "Ve skutečnosti víme, že existuje zebra, která je na jedné straně bílá."
Informatik: "Ale kdepak! To je výjimka!"

Typické zdroje hypotéz

- Požadavek na potřebnou kvalitu produktu
- Hypotéza je založena na předchozí zkušenosti
- Hypotéza vychází z teorie, kterou je třeba doložit
- Hypotéza je pouhým dohadem, založeným na náhodném pozorování

Srovnáváme dvě tvrzení:

H_0 – nulová hypotéza – většinou obhajuje stávající stav věcí

H_A – alternativní hypotéza ji odporuje

Chyby při testování hypotéz

Chyba 1. druhu: nulová hypotéza sice platí, ale my ji zamítáme. Rozhodnutí o zamítnutí H_0 je dáno hladinou významnosti testu, což je maximální přípustná pravděpodobnost chyby 1. druhu. Hladina testu se zpravidla značí symbolem α. Většinou volíme hladinu významnosti $\alpha=0,05$ nebo $\alpha=0,01$.

K **chybě 2. druhu** dochází, když nulová hypotéza neplatí, ale my ji nezamítneme (nepoznáme, že neplatí). Doplněk pravděpodobnosti chyby 2. druhu do jedničky $(1 - \beta)$ se nazývá **síla testu**. Je to pravděpodobnost, že nulovou hypotézu zamítneme, když tato hypotéza neplatí, tedy **pravděpodobnost, s jakou neplatnost hypotézy objevíme**.

Chyby v testování hypotéz

H_0 je ve skutečnosti: Já se rozhodnu takto⬇	**správná**	**nesprávná**
Zamítám H_0	**Chyba 1. druhu**	Správné rozhodnutí
Nezamítám H_0	Správné rozhodnutí	**Chyba 2. druhu**

- Pravděpodobnost chyby 2. druhu (β) obvykle neznáme. *1- β* je **síla testu**
- Čím větší nároky kladu na α (0.05 ⇨ 0.01⇨ 0.001), tím vyšší bude β
- β klesá i s rostoucím počtem pozorování

Zjištění statistické významnosti nikdy nemůže nahradit rozhodnutí o vědeckém (věcném) významu výsledků!

Chí kvadrát test dobré shody

H_0 – náhodný výběr byl proveden z rozdělení stanoveného typu

Provést: intervalové rozdělení četností

Podmínky: - žádný interval s nulovou četností; maximálně 20% intervalů s četností menší než 5

Testovací kritérium:

$$\chi = \sum_{i=1}^{n} \frac{(A_i - E_i)^2}{E_i}$$

kde: A_i je pozorovaná četnost, E_i je očekávaná četnost a n je počet intervalů.

Kritický obor: $\left(\chi^2_{\alpha,n-1}; \infty\right)$

Pokud je hodnota testovacího kritéria vyšší, než příslušná kritická hodnota rozdělení chí-kvadrát pro (n - 1) stupňů volnosti (kde n = počet intervalů), hypotézu o shodě dvou rozdělení **zamítáme** (na příslušné hladině významnosti)

Kolgomorov-Smirnovův test shody pro jeden výběr

H_0: **náhodný výběr byl proveden z rozdělení stanoveného typu**

Používá se v případech, kdy se nedoporučuje χ^2 test (při počtu tříd >2 nemá být více než 20% četností menších než 5 a žádná menší než 1, při k=2 nemá být žádná menší než 5).

Testovací kritérium:

$$D_1 = \frac{1}{n} \cdot \max \left| N_{a,j} - N_{e,j} \right|$$

kde

$N_{a,j}$ = aktuální kumulativní četnost v j-tém řádku

$N_{e,j}$ = očekávaná kumulativní četnost v j-tém řádku

`ks_test.m`

Testování hypotéz v MATLABu

H0: Data pocházejí z daného rozdělení,

h=1 pokud H0 zamítáme (na 5% hladině významnosti)
h=0 nemůžeme H zamítnout
p-hodnota – pst, že taková, (či ještě horší) data vybereme z předepsaného rozdělení. Čím je p menší, tím průkazněji zamítáme H0 – hladina významnosti.

Chí kvadrát dobré shody

```
[h,p]=chi2gof(y,'cdf',{@expcdf,par_est});
```

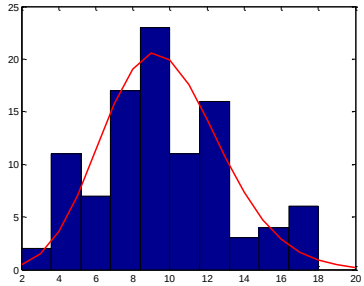
Kolmogorův-Smirnovův test

```
[h_ks,p_ks]=kstest(y,[y,expcdf(y,par_est)]);
```

Určete, z jakého rozdělení pocházejí data

```
load('data00.mat')
Nebo z menu File -> Open
```

```
[h1,p1] = chi2gof(y1,'cdf',{@geocdf,1/mean(y1)});
[h_ks,p_ks]=kstest(y1,[y1,geocdf(y1,1/mean(y1))]);
```

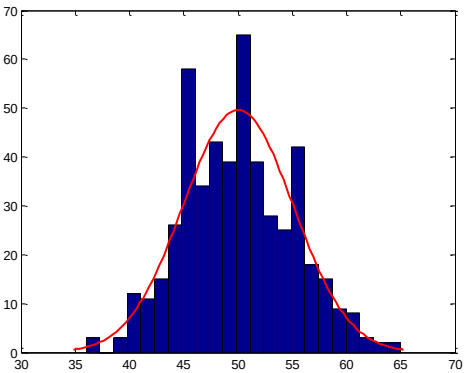


Kolmogorov-Smirnovův test

```
[h2,p2]=kstest2(x,y);
```

Generujte data z binomického rozdělení N=100, p=0.5 a nahraďte je normálním rozdělením. Testujte, zda data pocházejí ze stejného rozdělení

```
y= binornd(100,0.5,500,1);
[nu,sigma]=normfit(y);
y2=normrnd(nu,sigma,500,1);
[h2,p2]=kstest2(y,y2);
histfit(y)
```



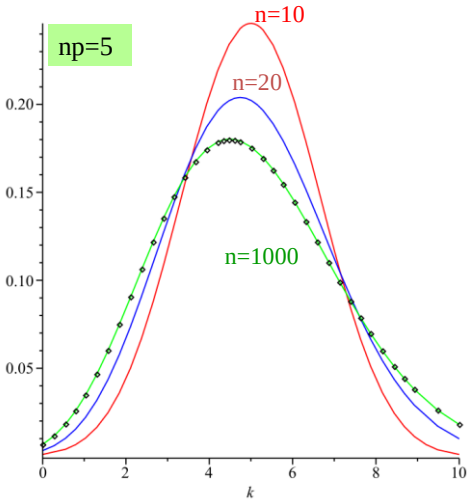
Poissonovo x normální

- Pokud n→∞, pak náhodná veličina s binomickým rozdělením konverguje k Poissonovu rozdělení, *np*=λ

$$Binom(n, p) \to Poisson(n \cdot p)$$

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E[X] = \lambda$$



Náhodná procházka = Markovský diskrétní stochastický proces

Např.

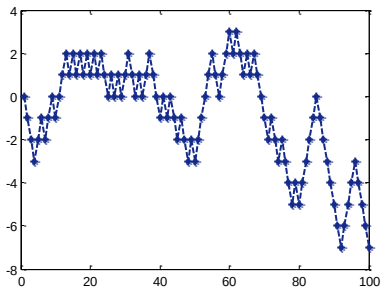
Spravedlivá hra, kde je pravděpodobnost úspěchu při každé sázce p = 0.5.

Stav procesu bude ohodnocen stavem peněženky jednoho z hráčů

{…, -2, -1, 0, 1, 2, …}, Začínáme na 0. Určete rozdělení pravděpodobnosti po 100 sázkách.

Model hry jednoho hráče

```
steps=100;
S=zeros(steps,1)
for i =1:steps-1
    if rand>0.5
    S(i+1)=S(i)+1
    else
       S(i+1)=S(i)-1
    end
end
plot(1:steps,S,'*m--')
```

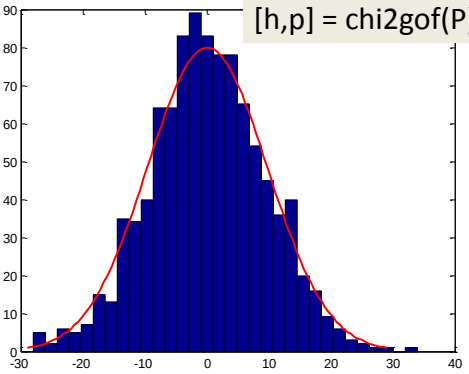


Náhodná procházka = Markovský diskrétní stochastický proces

Experiment opakujeme pro 1000 hráčů, určíme rozdělení pravděpodobnosti

```
function[vyhra]=hra(steps,p)
% [S(i)]=hra(steps,p)%No steps, probability of
success
S=zeros(steps,1);
for i =1:steps-1
    if rand<p
    S(i+1)=S(i)+1;
    else
       S(i+1)=S(i)-1;
    end
end
vyhra=S(i);
end
```

[h,p] = chi2gof(P)



Náhodná procházka = Markovský diskrétní stochastický proces

Rozdělení pravděpodobnosti stavu peněženky se krok od kroku liší. Zajímá nás, jestli pro počet sázek jdoucí k nekonečnu konverguje k tzv. ustálenému rozdělení psti.

[h,p] = chi2gof(P)