

Vedecko-technické výpočty

Základné štatistické
rozdelenia



Prehľad



- Štatistické rozdelenie
- Centrálna limitná veta
- Generovanie náhodných čísel
 - Trojuholníkové rozdelenie
 - Poissonove rozdelenie
 - Gaussove rozdelenie

Už vieme...



- Ak sa x mení spojitě, potom pravdepodobnosť, že x bude namerané v intervale $\langle x, x+dx \rangle$ je daná funkciou výsledkom $f(x)dx$, ktorá sa nazýva hustotou pravdepodobnosti (viac o rôznych distribúciách náhodných javov preberieme v ďalšej prednáške).
- Pravdepodobnosť že x nájdeme v intervale $\langle a, b \rangle$ je definovaná ako:

$$P(x \in \langle a, b \rangle) = \int_a^b f(x)dx$$

Otázky sú definované nasledovne:

- a) Aké rôzne rozdelenia pravdepodobností môžeme definovať?
- b) Ako môžeme generovať náhodné čísla s daným rozdelením.

Hlavné parametre rozdelenia



Pri štatistických rozdeleniach nás zaujíma najmä:

- Funkcia hustoty pravdepodobnosti
- Stredná hodnota rozdelenia
- Disperzia

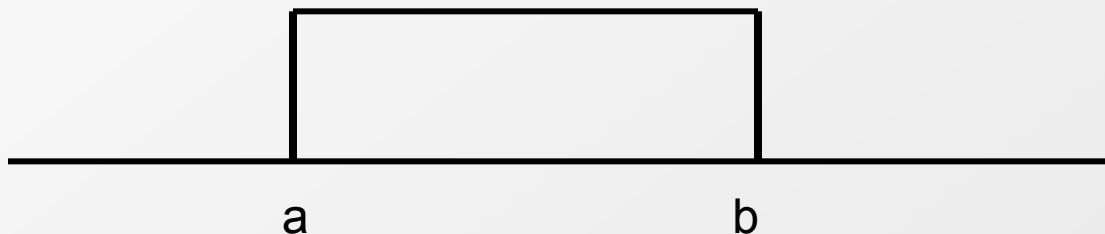
Rovnomerné rozdelenie



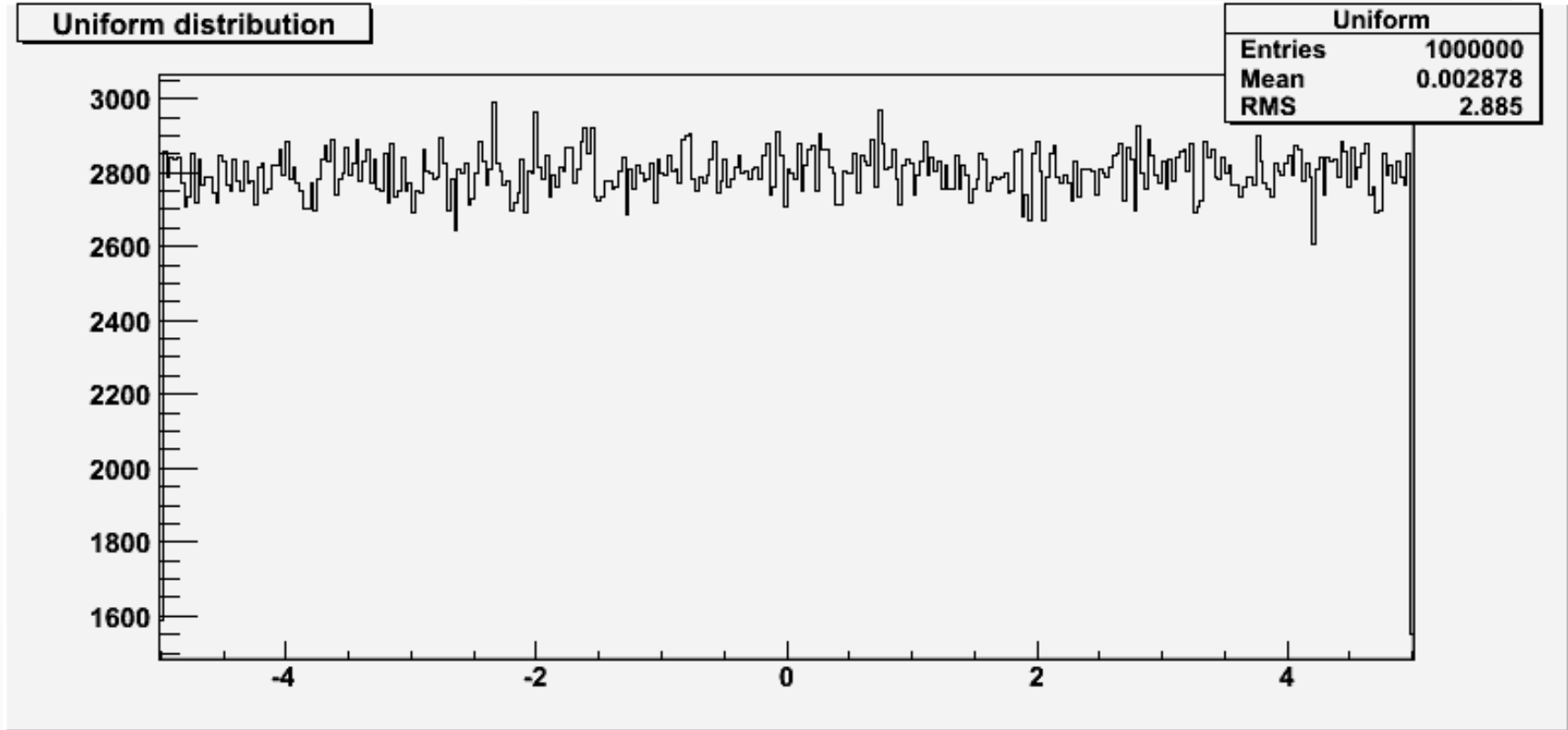
Najjednoduchšie rozdelenie. Opisuje ľubovoľný jav pri ktorom máme konštantnú pravdepodobnosť, že získame veličinu náhodnú veličinu s určitou hodnotou v rámci daného intervalu.

$$f(x; a, b) = \frac{1}{b-a} \text{ pre } x \in \langle a, b \rangle$$

$$f(x; a, b) = 0 \text{ pre } x \in (-\infty, a) \text{ a } x \in (b, +\infty)$$



Rovnomerné rozdelenie



Stredná hodnota

$$\bar{x} = \frac{a+b}{2}$$

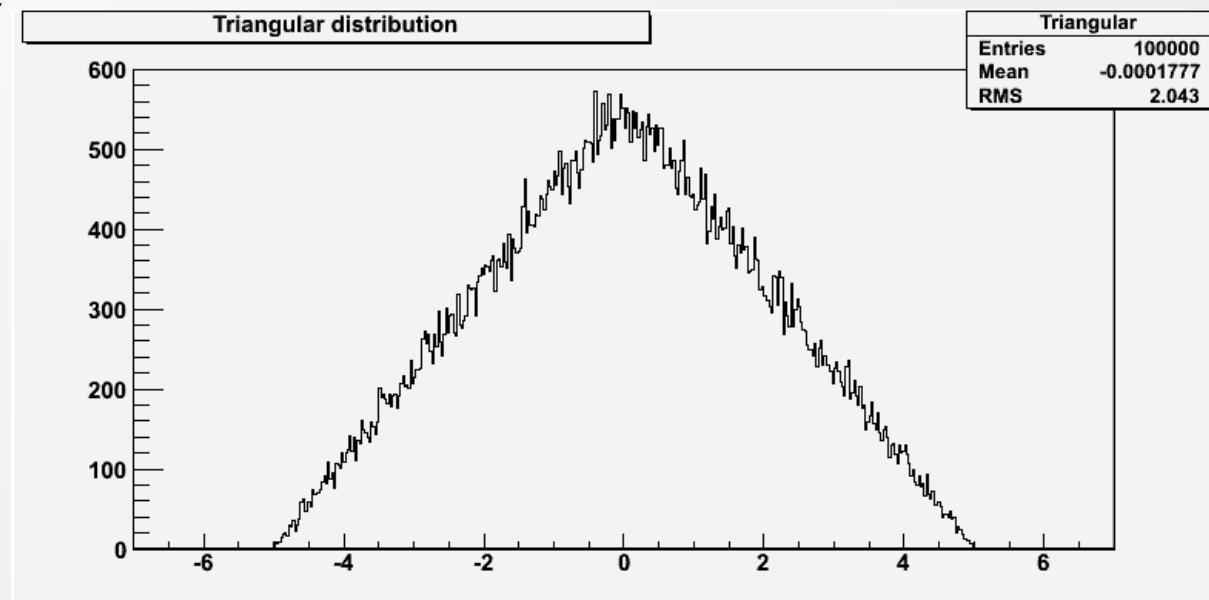
Disperzia

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Rozdiel dvoch rovnomerných čísel

Máme priemer dvoch náhodných čísel s rovnomernou distribúciou. Aké ma rozdelenie?

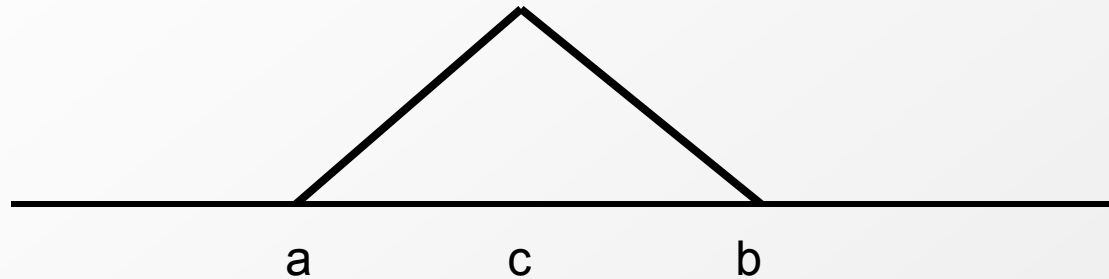
```
Void TriangleRandomNumber() {  
    double U1= (rand() / (double) RAND_MAX) ;  
    double U2= (rand() / (double) RAND_MAX) ;  
    double w= (U1+U2) / 2 ;  
    return w ;  
}
```



Trojuholníkové rozdelenie



Pravdepodobnostná funkcia je vyjadrená trojuholníkom, či už rovnoramenným, alebo všeobecným.



Využíva sa napr. v ekonómii alebo pri spracovaní zvuku. Opisuje taktiež čas medzi dvomi rôznymi náhodne prichádzajúcimi signálmi.

Trojuholníkové rozdelenie



$$f(x; a, b, c) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} \text{ pre } a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} \text{ pre } c \leq x \leq b \\ 0 \text{ všade inde} \end{cases}$$

Stredná hodnota

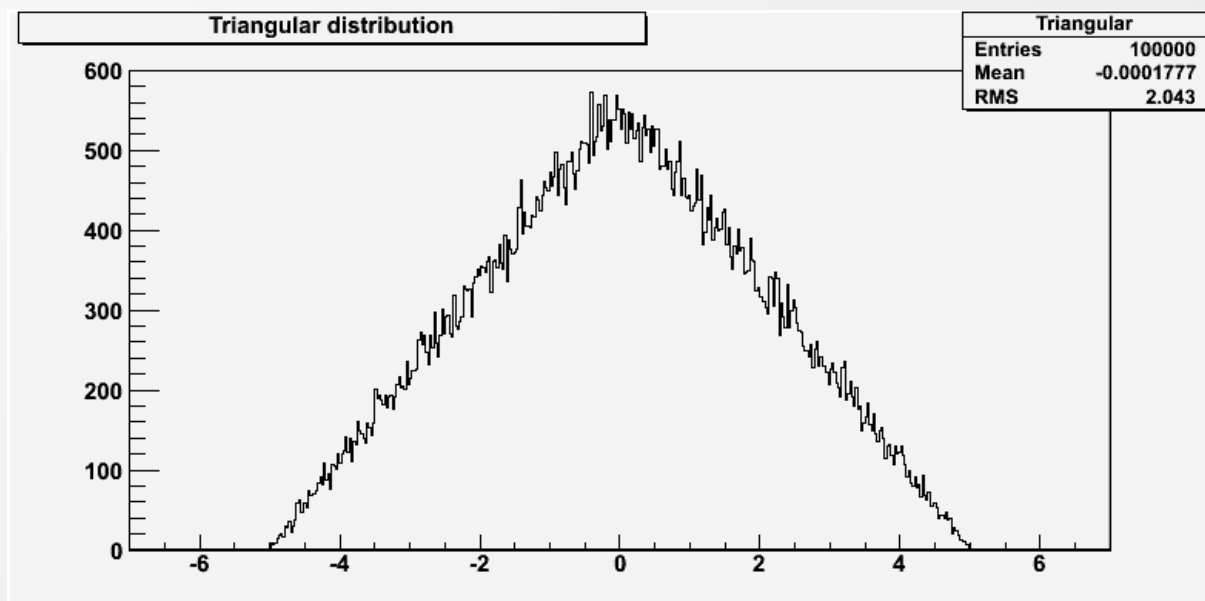
$$\bar{x} = \frac{a+b+c}{3}$$

Generovanie rovnomerného trojuholníkového rozdelenia



Rovnomerné rozdelenie získame napr. Ako priemer dvoch náhodných čísel s rovnomernou distribúciou.

```
Void TriangleRandomNumber() {  
    double U1= (rand() / (double)RAND_MAX) ;  
    double U2=(rand() / (double)RAND_MAX) ;  
    double w=(U1+U2) / 2 ;  
    return w ;  
}
```



Binomické rozdelenie



- Vhodné ak hodnota môže nadobúdať iba dve hodnoty (napr. výsledok hodu mincou)
- Ak je pravdepodobnosť získať výsledok p potom pravdepodobnosť získania r výsledkov z n pokusov bez ohľadu na poradie je daná binomickým rozdelením.

$$f(r; n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \quad r = 0, 1, 2, 3, \dots, n \text{ pre } p \in \langle 0, 1 \rangle$$

Stredná hodnota

$$\bar{r} = n \cdot p$$

Disperzia

$$\sigma^2 = n \cdot p \cdot (1-p)$$

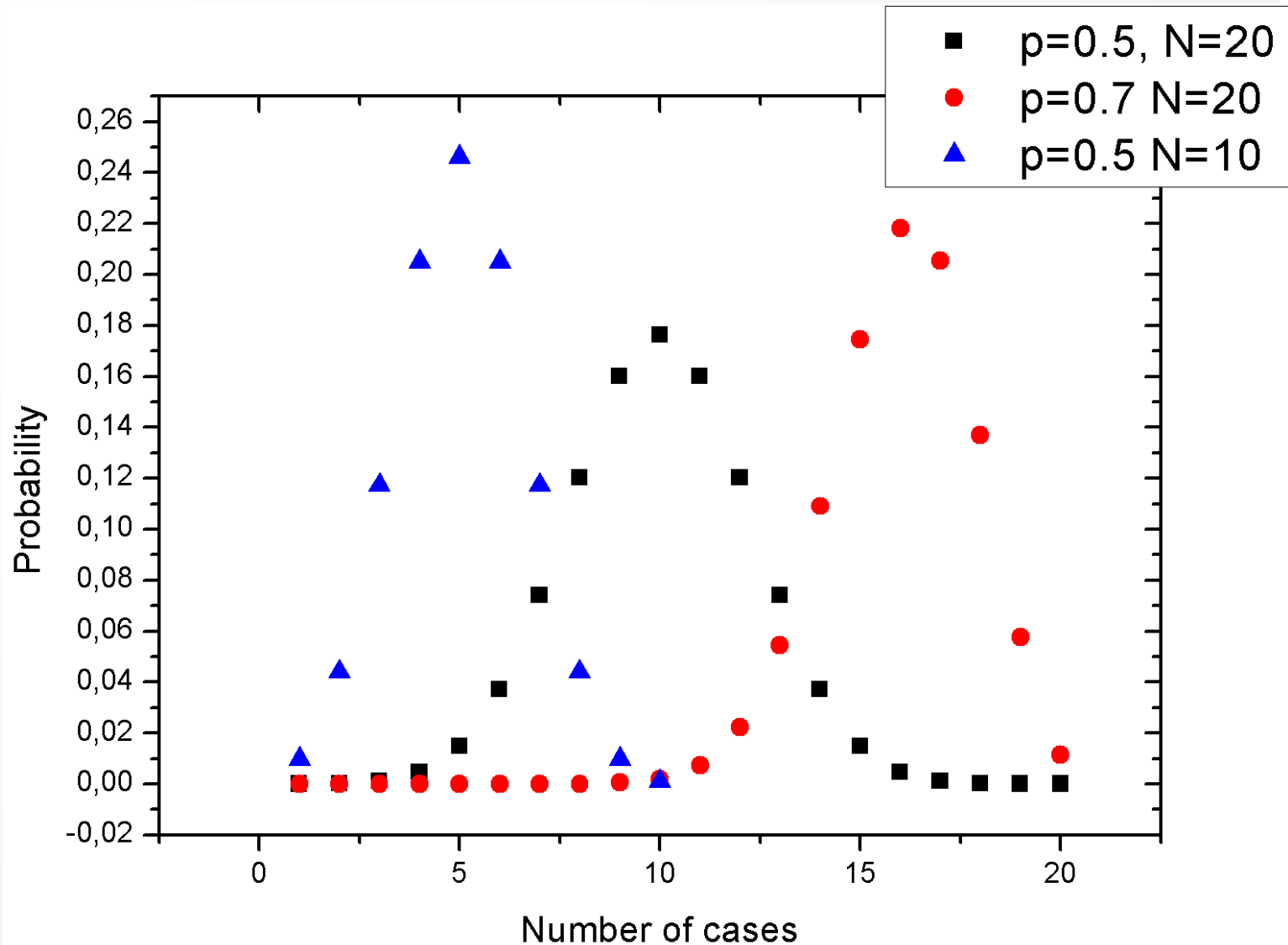
Význam binomického rozdelenia



- Opisuje napríklad keď z veľkeho počtu prípadov, vyberáme podmnožinu a sledujeme, či prvky podmnožiny spĺňajú určitú vlastnosť.
- Napríklad – človek môže mať tlak krvi v normále alebo nie. Ak si vieme štatistickú pravdepodobnosť p , že tlak krvi je v normále, a vyberieme si vzorku n ľudí, binomické rozdelenie nám povie aká je pravdepodobnosť, že práve r ľudí bude mať tlak krvi v normále.
- Príklad: Máme 5 hodov kockou. Aká je pravdepodobnosť, že padnu 2 šestky?

$$f(r; n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$
$$= \frac{5!}{2!(5-2)!} (1/6)^2 (1-(1/6))^{5-2} = 0.16$$

Príklad



Poissonovo rozdelenie



- Poissonovo rozdelenie udáva pravdepodobnosť nájdenia práve k javov v danom intervale (napr. priestore a čase) nezávisle od seba.
- Aproximácia binomického rozdelenia, pre veľký počet pokusov (zvyčajne ak $n > 30$ a $p < 0.1$).
- Toto rozdelenie sa používa na popis experimentálnych hodnôt, ktoré reprezentujú počet udalostí za jednotku času (napr. počet rádioaktívnych rozpadov, počet zákazníkov v obchode, počet áut prechádzajúcich cez úsek diaľnice a pod.)

Poissonovo - vyjadrenie



Hustota pravdepodobnosti tohto rozdelenia je

$$f(k; \mu) = \frac{\mu^k e^{-\mu}}{k!} \quad k = 0, 1, 2, 3, \dots \quad \mu > 0$$

Ak máme n pokusov s pravdepodobnosťou p tak stredný počet realizovaných prípadov je $\mu = np$.

Funkcia pravdepodobnosti nám definuje pravdepodobnosť, že sa pri danom μ realizuje práve hodnota k

- Stredná hodnota $\bar{k} = \mu$
- Disperzia $\sigma^2 = \mu$

Príklad jazdeckej jednotky pruskej armády



Prvá reálna aplikácia Poissonovho rozdelenia v reálnej situácii (Ladislaus Bortkiewicz, 1898)
Za 20 rokov bolo v pruskej armade v jednotke s 10 jazdcami 122 ľudí ukopaných koňmi k smrti. V niektoré roky nenastali žiadne úmrtia, ale vyskytol sa aj rok, keď zahynuli štyria. Otázka je, či je proces čisto náhodný a ako opísať pravdepodobnosť, že niekto zahynie.



To je ekvivalentné, akoby sme sledovali vzorku 200 členov zaradených do jednotky počas jedného roku (pričom mŕtvy člen sa vždy nahradí novým). Teda 200 vojako-rokov. Na koľkých pozíciách príde ku koľkým úmrtiam?

Príklad jazdeckej jednotky pruskej armády



$$f(k; \mu) = \frac{\mu^k e^{-\mu}}{k!} \quad k = 0, 1, 2, 3, \dots \quad \mu > 0$$



Priemerná hodnota (úmrtia na jednej pozícii)
 $\mu = 122/200 = 0,61$ úmrtia na jedného vojaka na rok.

Príklad jazdeckej jednotky pruskej armády



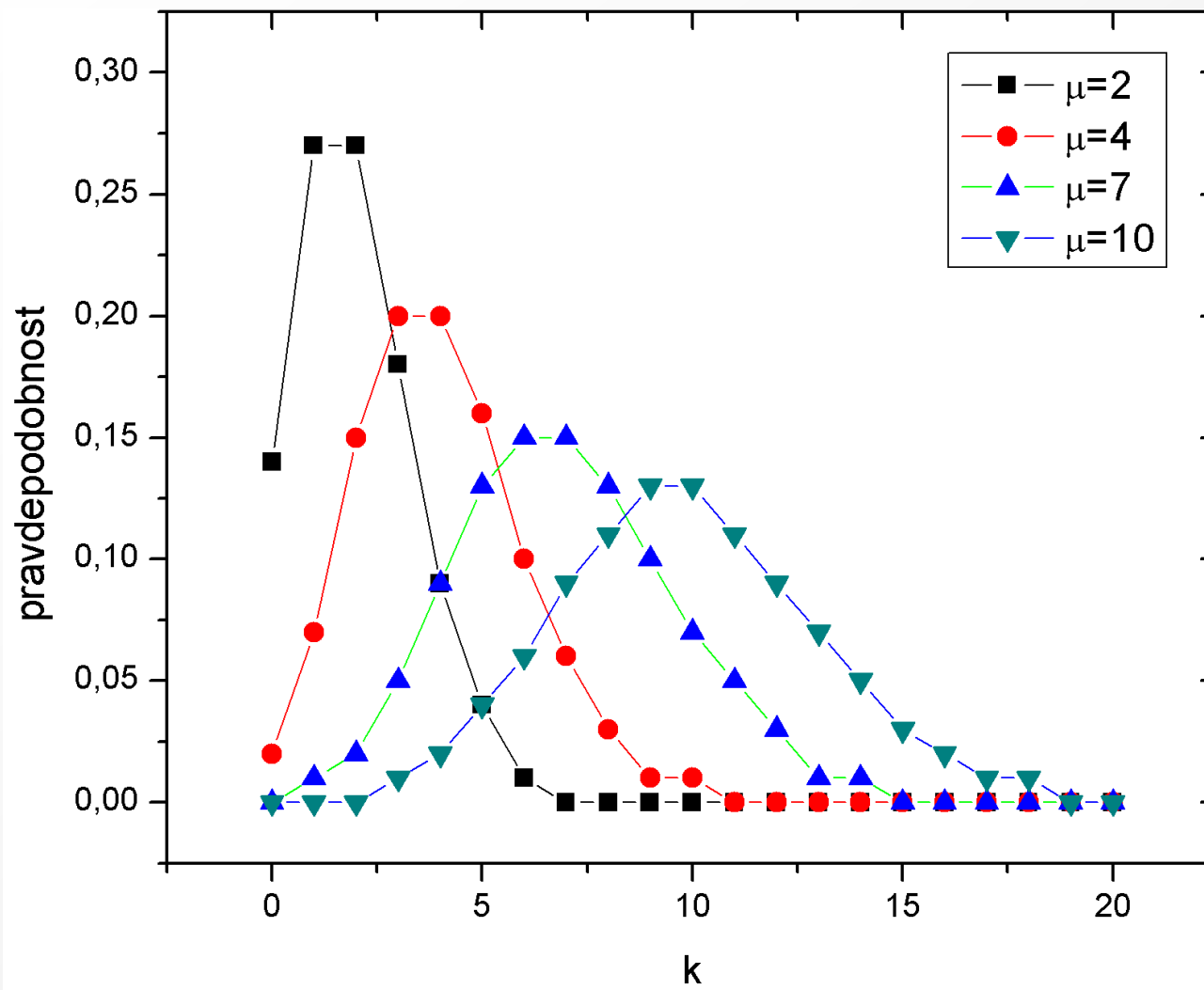
$$f(k; \mu) = \frac{\mu^k e^{-\mu}}{k!} \quad k = 0, 1, 2, 3, \dots \quad \mu > 0$$

Priemerná hodnota (úmrtia na jednej pozícii)
 $\mu = 122/200 = 0,61$ úmrtia na jedného vojaka na rok.



# úmrtí /vojaka /rok	Pravdepodobnosť	Skutočných úmrtí	Odhadnutých úmrtí
0	0,5433508691	109	108,6701738149
1	0,3314440301	65	66,2888060271
2	0,1010904292	22	20,2180858383
3	0,0205550539	3	4,1110107871
4	0,0031346457	1	0,626929145
5	0,0003824268	0	0,0764853557

Poisson pre jednotlivé prípady



Kumulatívna distribučná funkcia



Zoberieme si sumu jednotlivých členov – tzv. kumulatívnu distribučnú funkciu

$$D(k; \mu) = \sum_{k=0}^{+\infty} \frac{\mu^k e^{-\mu}}{k!} = e^{-\mu} \sum_{k=0}^{+\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{\mu} = 1$$

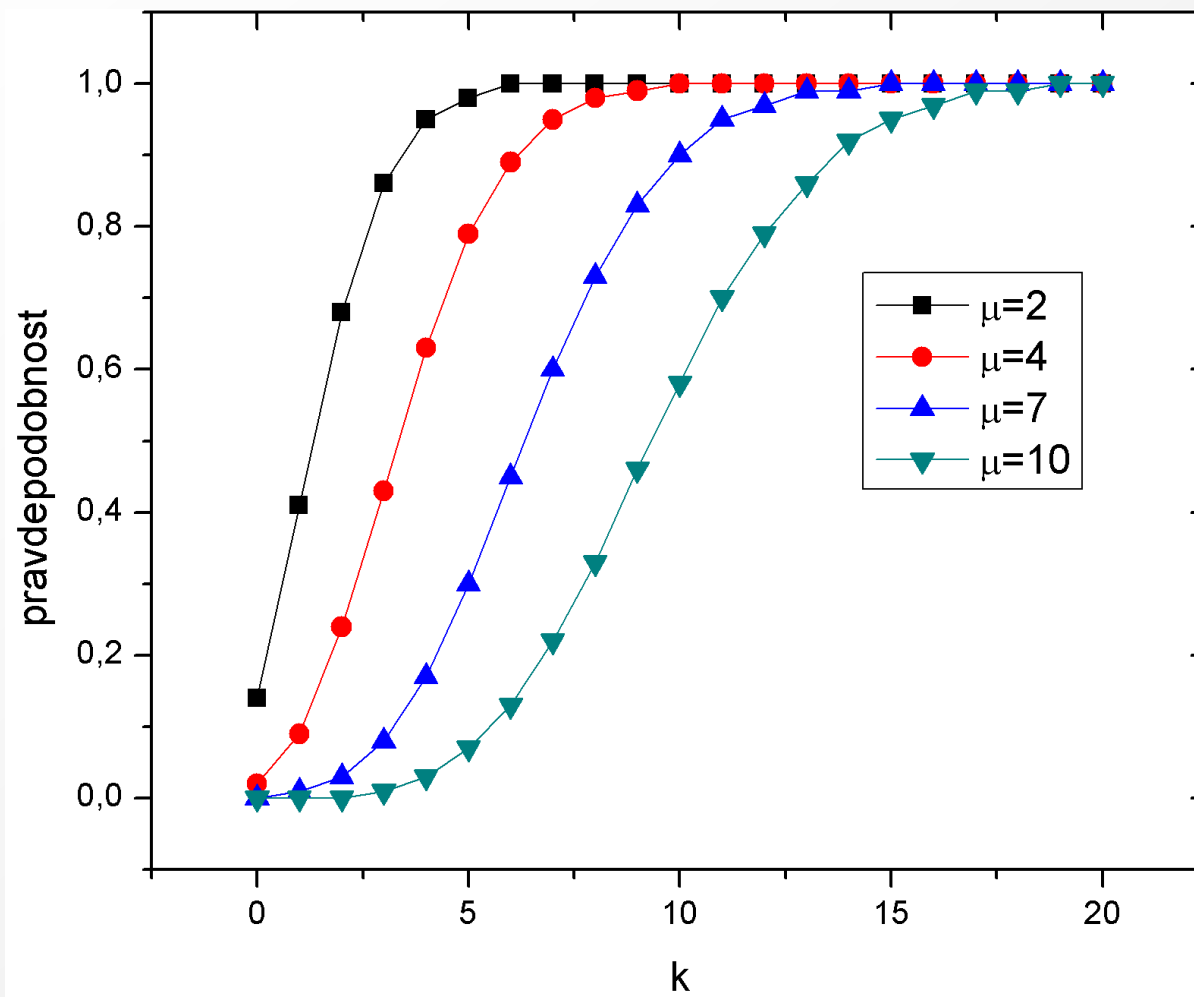
Este sa pozrieme na pomer dvoch susednych clenov

$$\frac{f(k+1; \mu)}{f(k; \mu)} = \frac{\frac{\mu^{k+1} e^{-\mu}}{(k+1)!}}{\frac{\mu^k e^{-\mu}}{k!}} = \frac{\mu}{(k+1)}$$

Poisson pre jednotlivé prípady



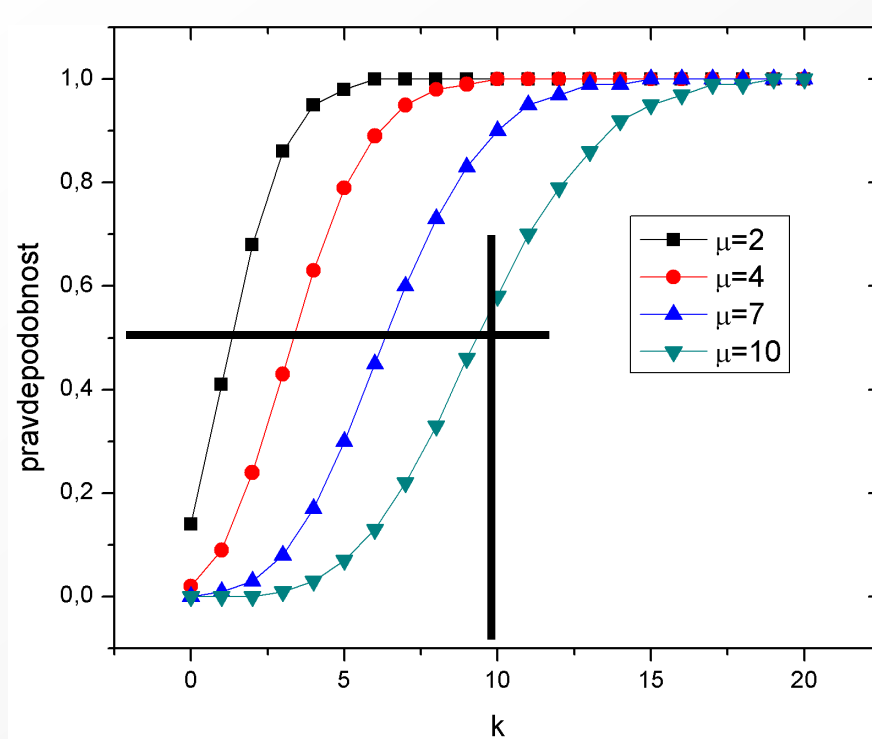
Kumulatívna distribučná funkcia



Generovanie náhodných čísel s Poisson distribúciou



Idea: Vygenerujem číslo s rovnomerným rozdelením od 0 po 1 a označím ho p . Pre distribúciu s daným μ vypočítame kumulatívne funkcie D pre všetky hodnoty k .



Ako náhodné číslo s Poissonovým rozdelením akceptujeme prvé číslo k , pre ktoré je hodnota $D > p$.

Táto tzv. kumulačná metóda je viac-menej využiteľná vždy, keď vieme jednoducho vypočítať kumulatívnu funkciu D .

Optimalizácia generátora



Opakovaný výpočet faktoriálov a exponenciálnych funkcií je časovo náročný. Preto si doratáme člen $f(k+1)$ pomocou člena $f(k)$ a pripočítame k predchádzajúcej sume, pričom $P(0) = e^{-\mu}$

= využijeme

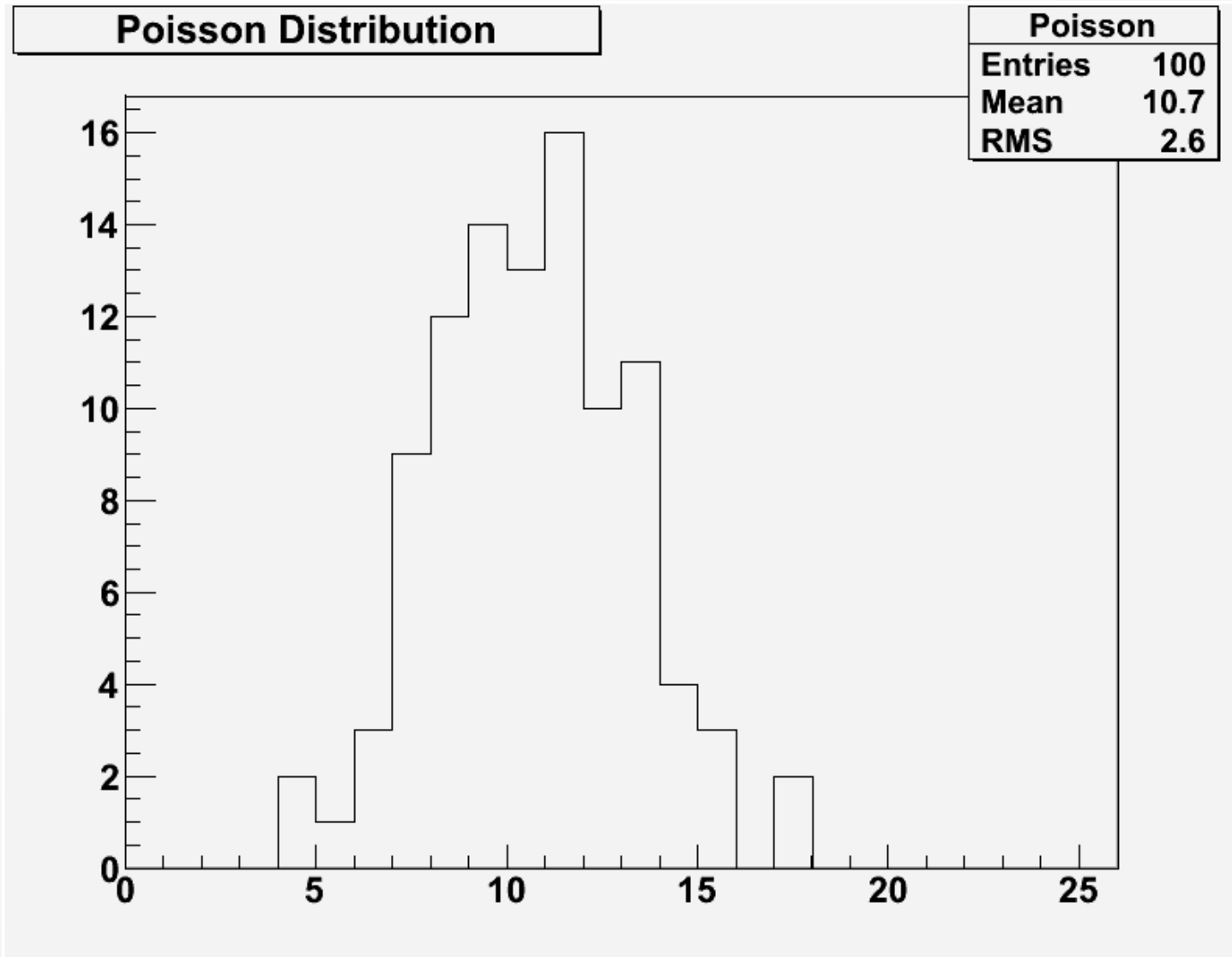
$$\frac{f(k; \mu)}{f(k-1; \mu)} = \frac{\frac{\mu^k e^{-\mu}}{(k)!}}{\frac{\mu^{k-1} e^{-\mu}}{(k-1)!}} = \frac{\mu}{k}$$

Generovanie náhodných čísel s Poisson distribúciou



```
const int PoissonRandomNumber(const double mi)
{
    int k=0;                //Počítadlo
    const int max_k = 1000; //horný limit pre k
    double p = rand() / (double)RAND_MAX; //rovnom.rozd.(0,1)
    double P = exp(-mi);    //P(0) - prvý člen kumulatívu
    double sum=P;           //kumulačná funkcia (ozn.f)
    if (sum>=p) return 0;   //už prvý člen splní podm.
    for (k=1; k<max_k; ++k) { //Prechádzame cez všetky k
        P*=mi / (double)k;  //kumulat. pre ďalšie k
        sum+=P;             //Navyšujeme kumulatív
        if (sum>=p) break;  //Stop prekročíme učený limit
    }
    return k;              //návrat vygenerovaného čísla
}
```


Príklad pre $\mu=10$



Poisson pre špeciálne prípady



- Aproximáciou binomického rozdelenia pre $n \rightarrow \infty$ s pravdepodobnosťou $p \rightarrow 0$.
- Pre vysoké μ je podobný gaussovmu rozdeleniu

Gauss rozdelenie (normálne)



- Normálne rozdelenie je tiež nazývané Gaussovým rozdelením.
- Je to najdôležitejšie rozdelenie z pohľadu spracovania experimentálnych dát.
- Popisuje náhodné pozorovania pre väčšinu experimentov a procesov v prírode.
- Jeho dôležitosť je do značnej miery daná tzv. centrálnou limitnou vetou, ktorú si spomenieme neskôr.

Gauss rozdelenie - vyjadrenie



Hustota pravdepodobnosti je definovaná ako:

$$f(x; \mu, \sigma) = \frac{1}{\sigma \cdot \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ podm. } \sigma > 0$$

Stredná hodnota

$$\bar{x} = \mu$$

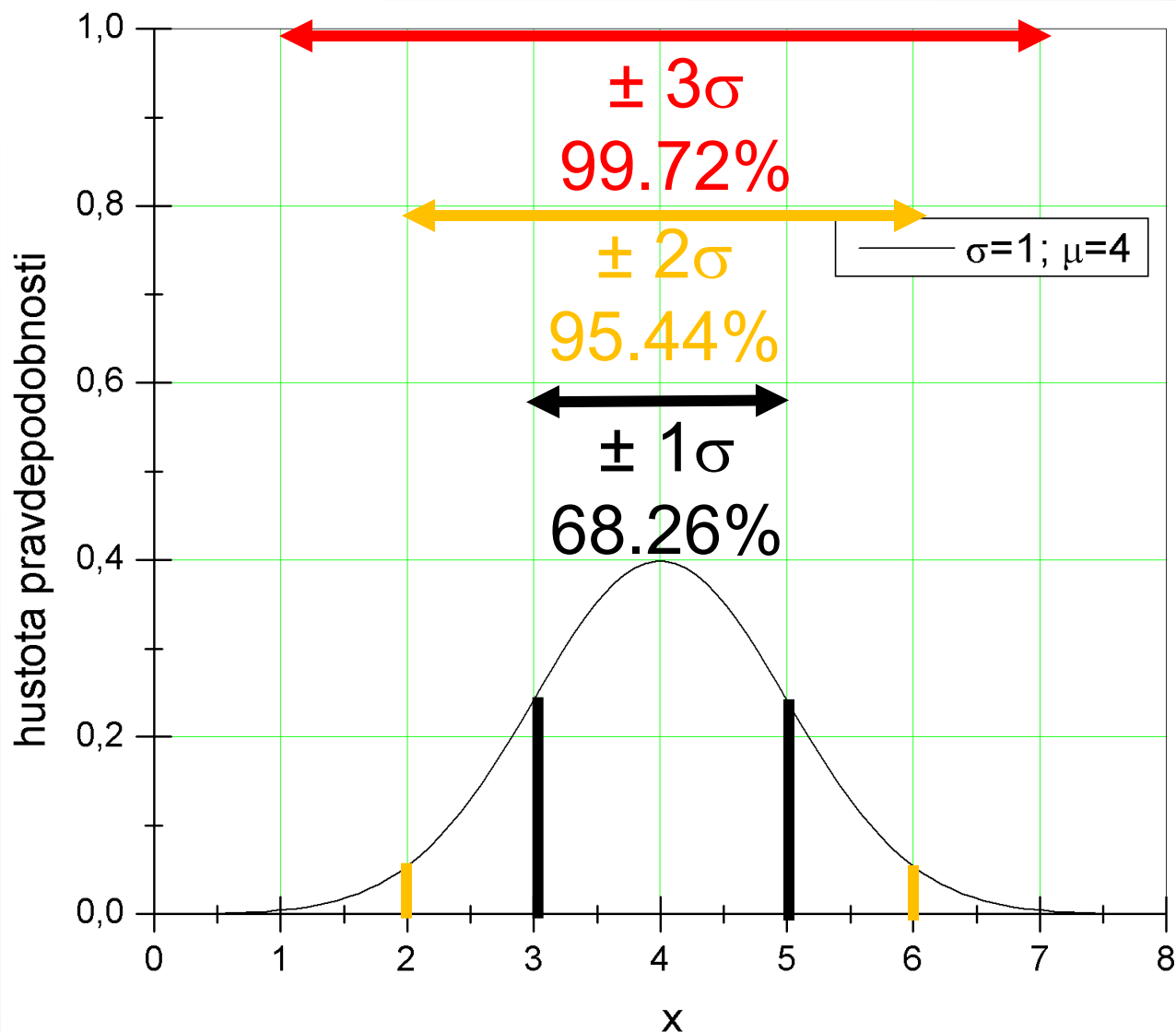
Disperzia

$$\sigma^2 = \sigma^2$$

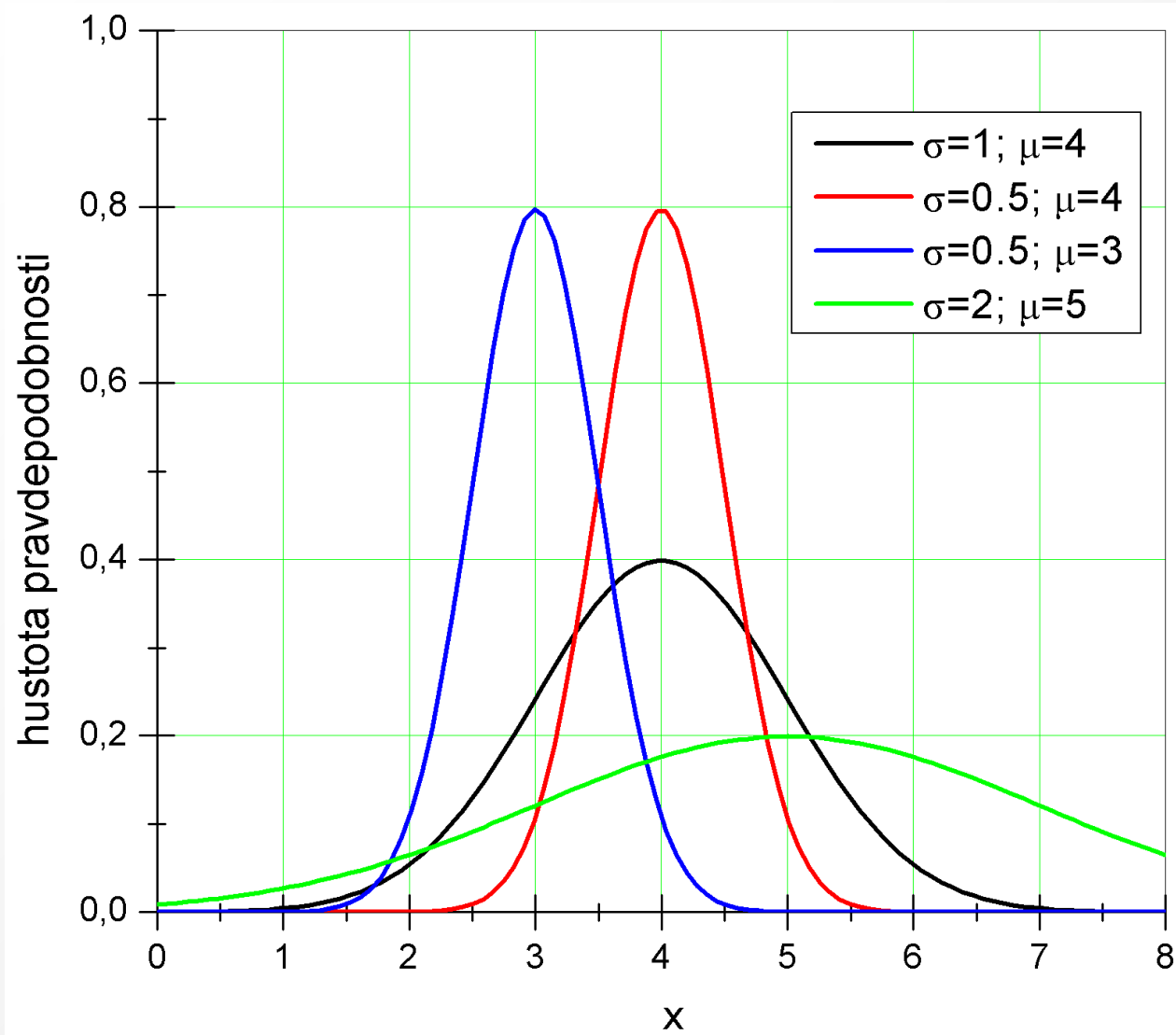
Kumulatívna distribučná funkcia:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x}{\sqrt{2}} \right) \right], x \in \mathbb{R}$$

Priebeh Gauss funkcie



Vplyv parametrov na Gauss



Kumulatívna distribučná funkcia (Gauss)



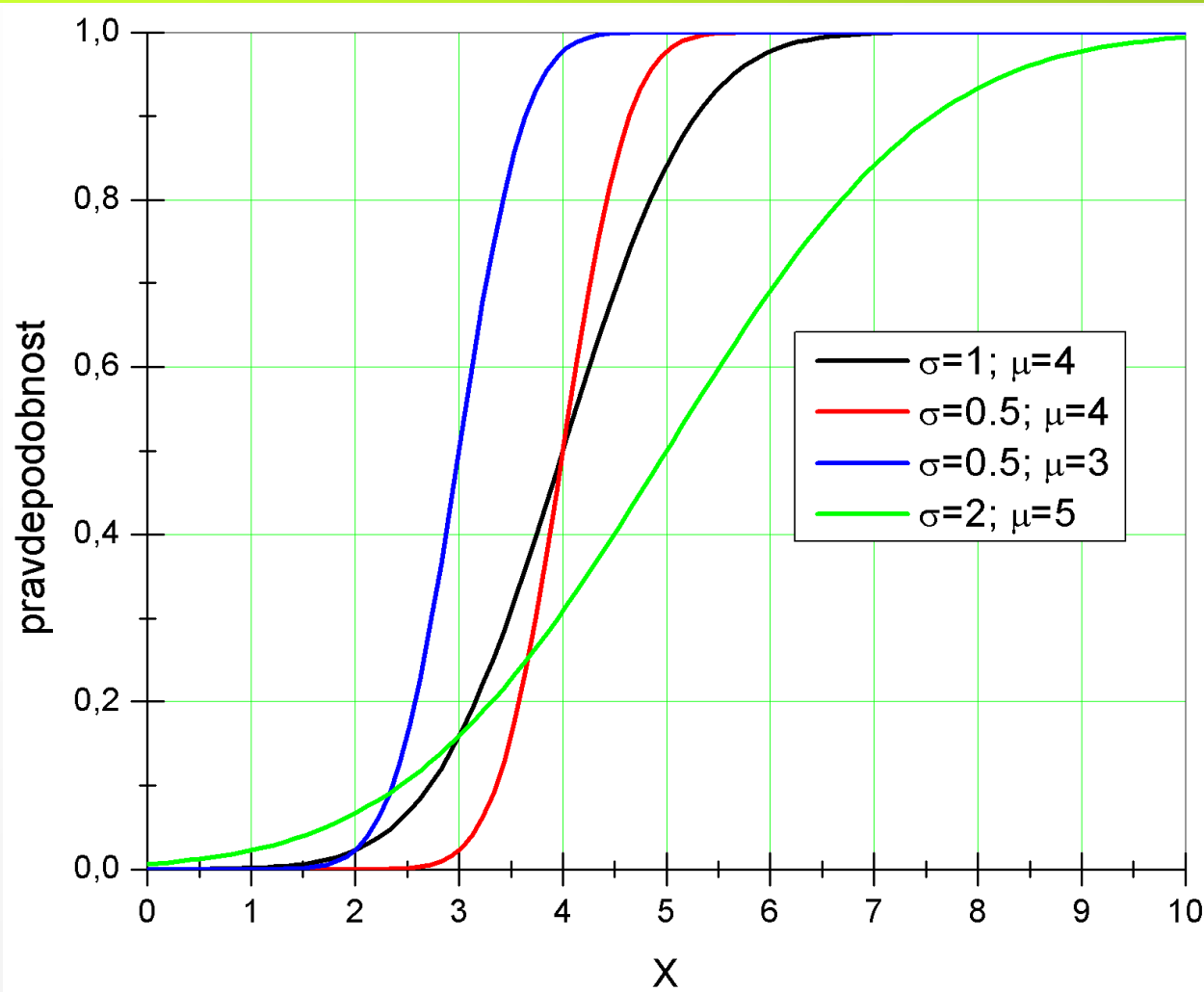
Kumulatívna distribučná funkcia využíva intergrál z funkcie s hustotou pravdepodobnosti

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right], x \in \mathbb{R}$$

Pričom vo všeobecnosti platí:

$$F(x; \mu, \sigma^2) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right) \right], x \in \mathbb{R}$$

Kumulatívna distribučná funkcia (Gauss)



Centrálna limitná veta

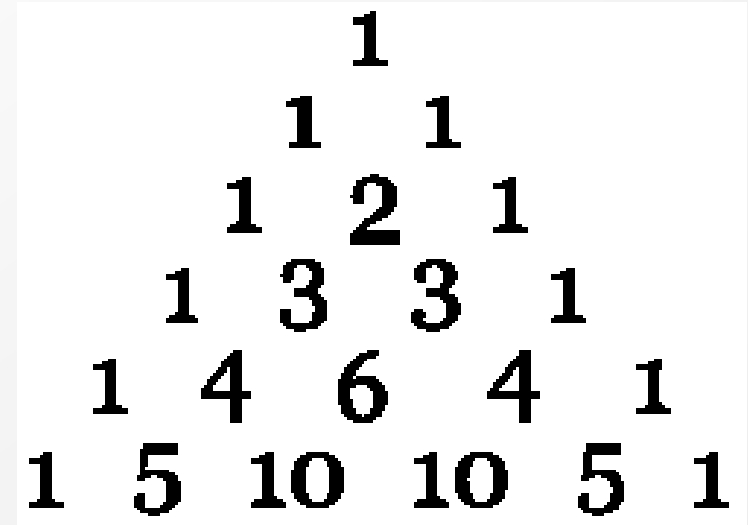


Ak máme postupnosť náhodných premenných X_n , ktoré sú rozdelené podľa rôznych rozdelení pravdepodobnosti, s konečnými strednými hodnotami a disperziami, potom pre dostatočne veľké n získavame približne Gaussovo rozdelenie pre ich sumu:

$$\sum_{i=1}^n X_i$$

Dôležitá poznámka: „rozdelené podľa rôznych rozdelení“ treba chápať, že z daných náhodných premenných ani jedna nemusela byť pôvodne rozdelená Gaussovsky.

Centrálna limitná veta dáva jednoduchú možnosť ako získať náhodné čísla s Gauss rozdelením približnou metódou.



Inak povedané, spustené guľičky sa prerozdeľujú podľa pascalovho trojuholníka určujúceho koeficienty binomického rozdelenia.

Generovanie náhodnej premennej s Gauss rozdelením



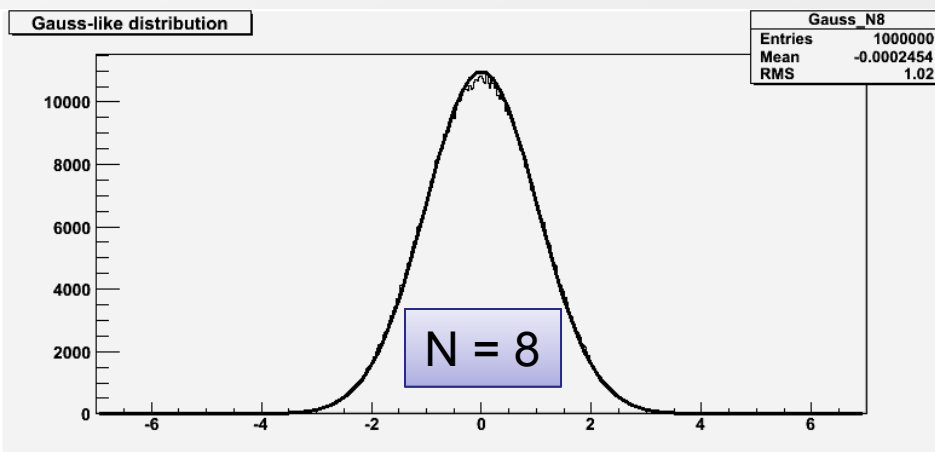
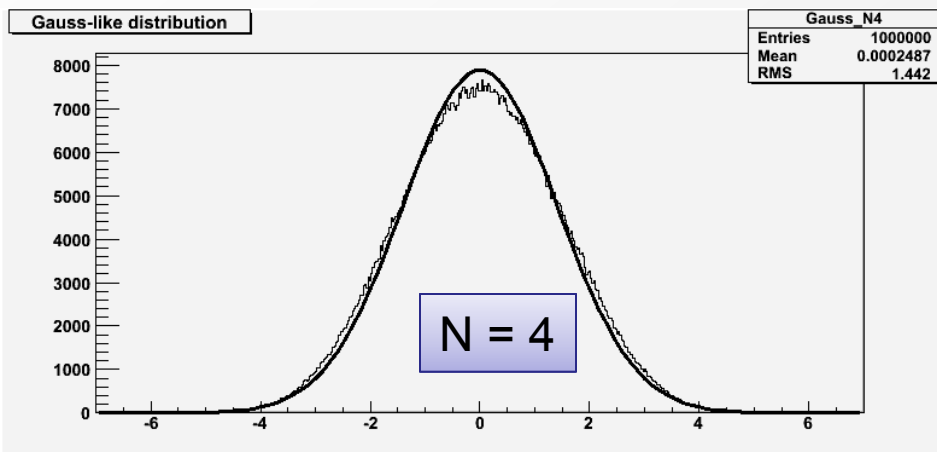
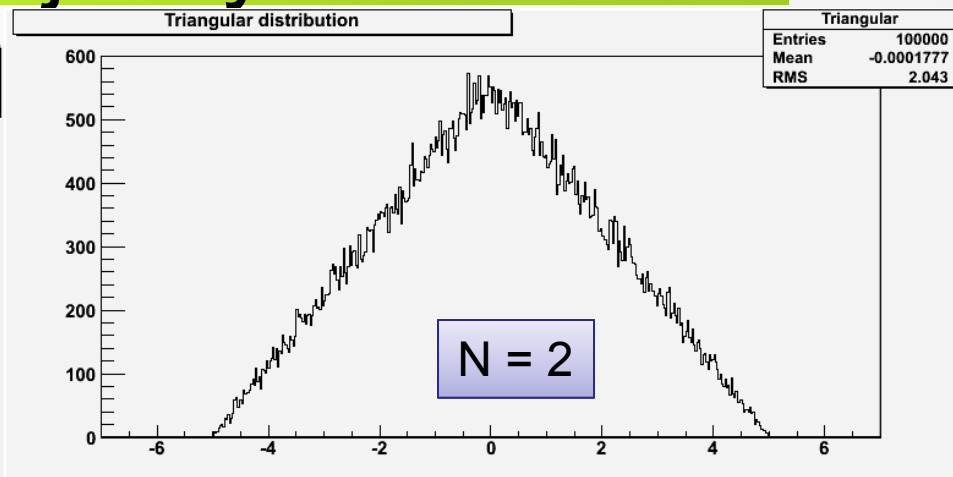
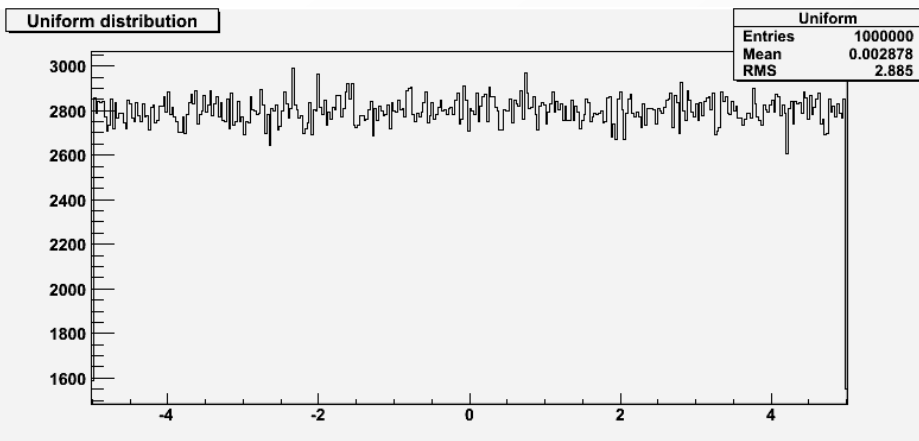
- Zvyčajne vychádzajú generátory z náhodného čísla s rovnomerným rozdelením.
- Možnosti máme približné a presné.
- Napríklad inverzná metóda s využitím kumulatívnej distribučnej funkcie Φ^{-1} je exaktná, ale chýba nám vhodné analytické vyjadrenie.
- Jednou z možností je napr. aproximácia pomocou polynómu. Problémom je však vyjadrenie hodnôt vzdialených o $4-5 \times \sigma$ od strednej hodnoty

Centrálna limitná veta



```
Void CentralLimit() {  
    double U1=0;  
    int N = 12; // počet iterácií  
    for(int i=0; i<N; i++)  
        U1=U1+10*rand() / (double) RAND_MAX - 5.0;  
    U1=U1/i;  
    return U1;  
}
```

Praktické overenie centrálnej limitnej vety

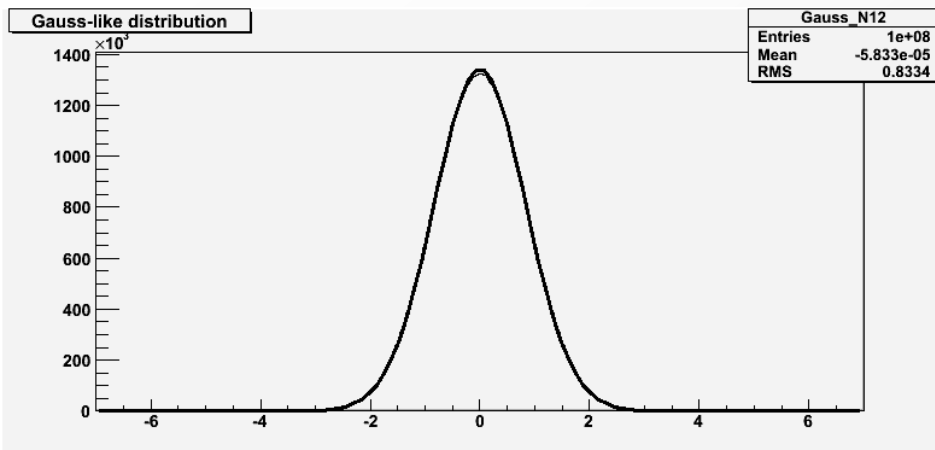


Centr. lim. veta Pre $N = 12$

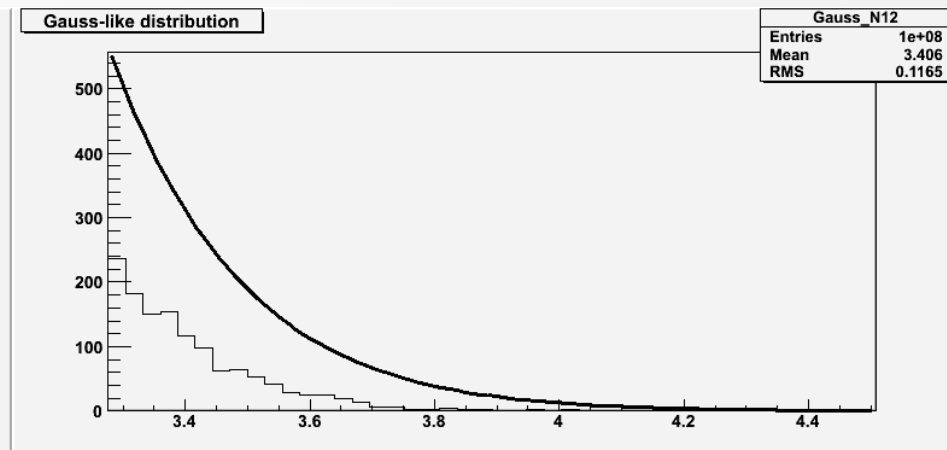


Poznatok 1: Výrazné spomalenie výpočtu

Poznatok 2: Aká je miera presnosti?



Vynikajúci súlad pre celkový tvar distribúcie...



... stále sú však násobné rozdiely pre oblasť $4 \times \sigma$, ktoré môžu byť problémom pri simuláciách a opisoch situácií s vysokou citlivosťou.

Upozornenie: Centrálna limitná veta je použiteľná len ak sa nepožaduje skutočná gauss distribúcia.

Box-Muller metóda (1958)



Dve premenné U_1 a U_2 s rovnomerným rozdelením z intervalu $(0,1)$ prevedieme matematickou transformáciou na dve premenné Y_1 a Y_2 s Gauss rozdelením.

$$Y_1 = \sqrt{-2 \ln(U_1)} \cos(2\pi U_2)$$

$$Y_2 = \sqrt{-2 \ln(U_1)} \sin(2\pi U_2)$$

Box-Muller implementácia



```
Void GaussRandomNumber(double &Gauss1, &Gauss2) {  
    double U1=rand() / (double) RAND_MAX;  
    double U2=rand() / (double) RAND_MAX;  
    double a=sqrt(-2*log(U1));  
    double b=(2*pi*U2);  
    Gauss1=a*cos(b);  
    Gauss2=a*sin(b);  
}
```

+ Veľmi jednoduchá implementácia

- Nevýhody: Numerická nestabilita pre $U1$ v okolí 0.
- Pomalá metóda – volá sa priveľa funkcií. (Čiastočné riešenie je vrátiť užívateľovi jednu hodnotu a druhú odložiť pre ďalšie volanie funkcie)

Nevýhody sú kritické pri masívných stochastických výpočtoch.

Box-Muller v polárnych súradniciach



Problémy rieši čiastočne polárna forma Box-Muller transformácie. Originálne už z roku 1968.

Akceptujú sa iba niektoré vygenerované čísla (tzv. rejection methods). V tomto prípade sa akceptuje cca $\pi/4$ čísel.

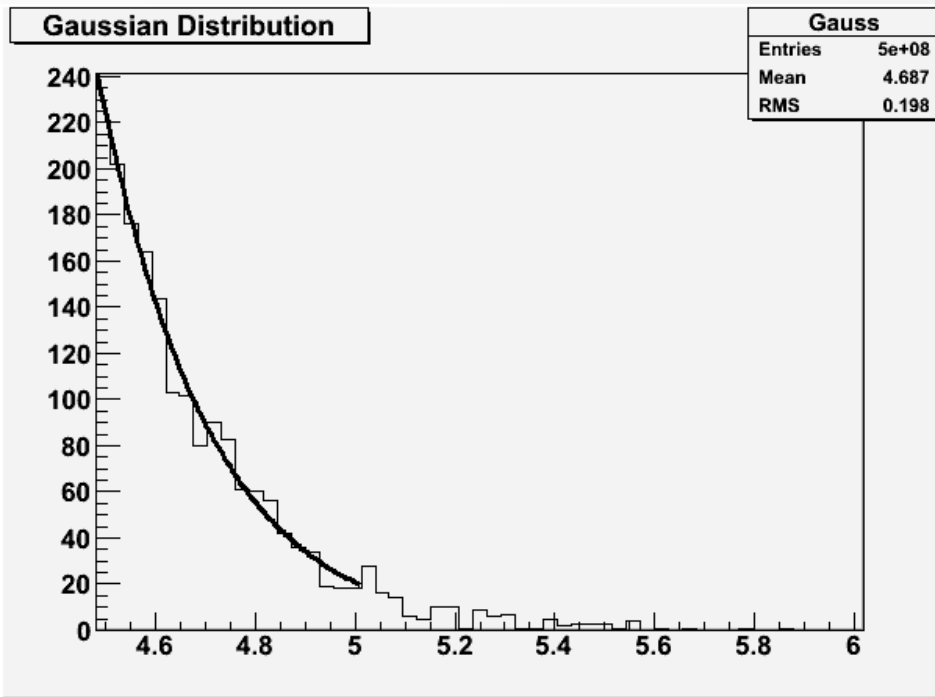
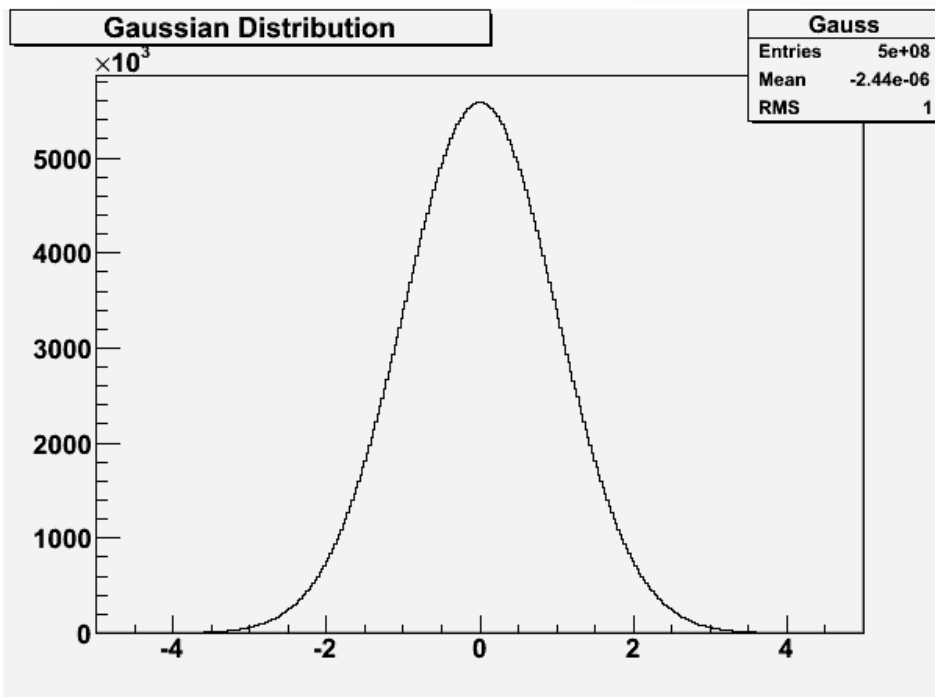
Polárna metóda – impl.



```
Void GaussRandomNumber(double &Gauss1, &Gauss2) {  
    do {  
        double U1=2.0*(rand()/(double)RAND_MAX)-1.0;  
        double U2=2.0*(rand()/(double)RAND_MAX)-1.0;  
        double w=U1*U1+U2*U2;  
    } while (w>=1.0); // akceptuje sa iba W<1  
    w = sqrt((-2.0*log(w))/w);  
    Gauss1=U1*w;  
    Gauss2=U2*w;  
}
```

Netreba volať sin a cos $\Rightarrow$ čiastočne lepší výkon

Finálny výsledok



Výborný súlad aj pri $5 \times \sigma$

Nad $4.5 \times \sigma$ je iba 3×10^{-6} hodnôt

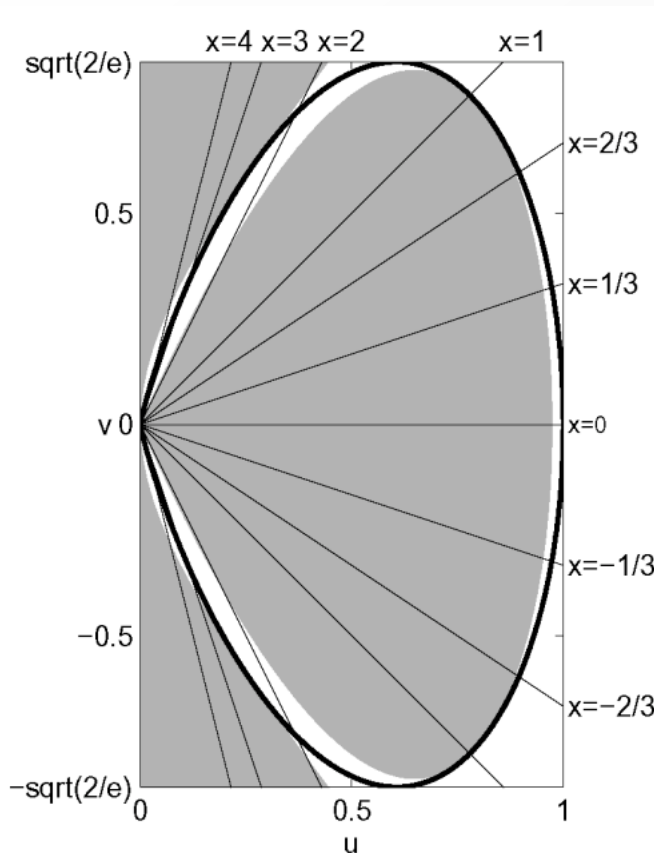
Test pri 6σ a viac je už dosť časovo náročný

Ratio-of-Uniforms



Ďalšia relatívne jednoduchá presná metóda.

Autori Kinderman a Monahan (1977), vylepšil J. L. Leva (1992)



Body vo vnútri krivky sú akceptované,
body mimo krivky sú zamietnuté.

Generujú sa čísla U_1 a V_1 s rovnomerným
rozdelením, pričom $U_1 \in (0,1)$ a $V_1 \in (-1,1)$.
 V_1 sa následne upraví ako

$$V_1 \leftarrow V_1 \cdot \sqrt{\frac{2}{e}} / U_1$$

Ratio of Uniforms (idea)



Potom sa získané hodnoty U_1 a V_1 podrobia sérii testov

1) Automaticky sa akceptujú body, kde $V_1^2 \leq (5 - 4e^{1/4}U_1)$

2) Ak bod nevyhovuje podmienke 1) urobí sa test a posudzujú sa ďalej iba body kde $V_1^2 < (4e^{-1.35}/U_1 + 1.4)$

3) Ak bod vyhovuje podmienke 2) akceptuje sa ak

$$V_1^2 \leq (-4 \log(U_1))$$

Výhoda je, že iba pre časť hodnôt sa počíta logaritmus.

Pre zvyšné sa využívajú iba jednoduché numerické operácie.

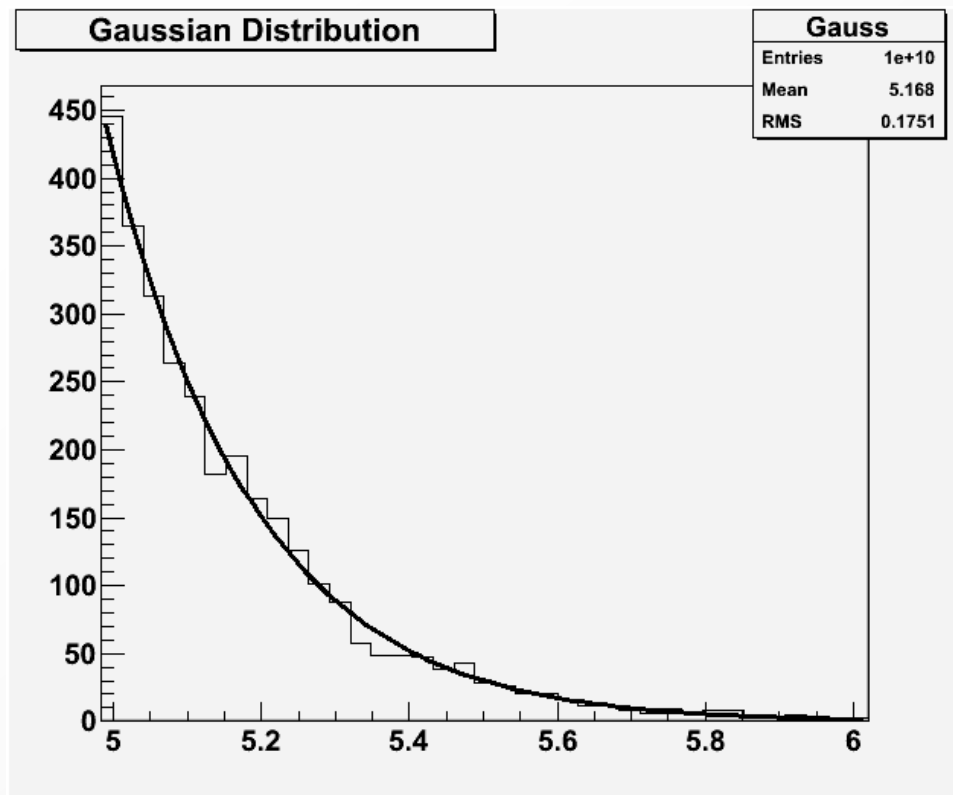
Nevýhoda: generujeme 2 rovnomerné čísla a získame iba jedno s gauss rozdelením.

Implementácia



```
Double U1,V1,Result = 0;
bool AcceptedResult = 0;
do {
    U1=(rand() / (double)RAND_MAX);
    V1=(2.0*(rand() / (double)RAND_MAX)-1.0)*sqrt(2/exp(1))/U1;
    if ((pow(v1,2))<=(5-4*pow(exp(1),0.25)*U1)) {
        AcceptedResult=1;
        Result = V1;
    } else
    if ((pow(V1,2))<(4*pow(exp(1),-1.35)/U1+1.4)) {
        if ((pow(V1,2))<=-4*(log(U1))) {
            AcceptedResult=1;
            Result = V1;
        }
    }
} while (AcceptedResult!=1);
Return Result
```

Finálny výsledok



Výborný súlad aj pri $6 \times \sigma$

Stredná hodnota fitu -8.7×10^{-6}

Stredná kvadratická odchýlka 1.00004

Výborne vystihuje gauss rozdelenie

Porovnanie



Čas potrebný pre vygenerovanie 1×10^9 čísel.

Centrálna limitná veta (N = 12)	Box-Muller	Polar	Ratio of Uniforms
6 min 6 sek	3 min 10 sek	3 min 3 sek	2 min 16 sek

Upozornenie: Výsledky silne závisia od optimalizácie kompilátora, realizácie kódu (napr. kontrol. výpisov) a samozrejme HW.

Poznámka: Box-Muller a Polar nám dávajú dva výsledky. Ak generujeme náhodné čísla s gauss-rozdelením často a dokážeme uschovať jedno číslo do ďalšieho volania generátora tak s tým získame faktor 2 vo výkone.



THE END