Introduction to Applied Statistical Methods in Clinical Medicine

Probability Distributions

Leif E. Peterson, Ph.D.
Dept. of Medicine
Baylor College of Medicine

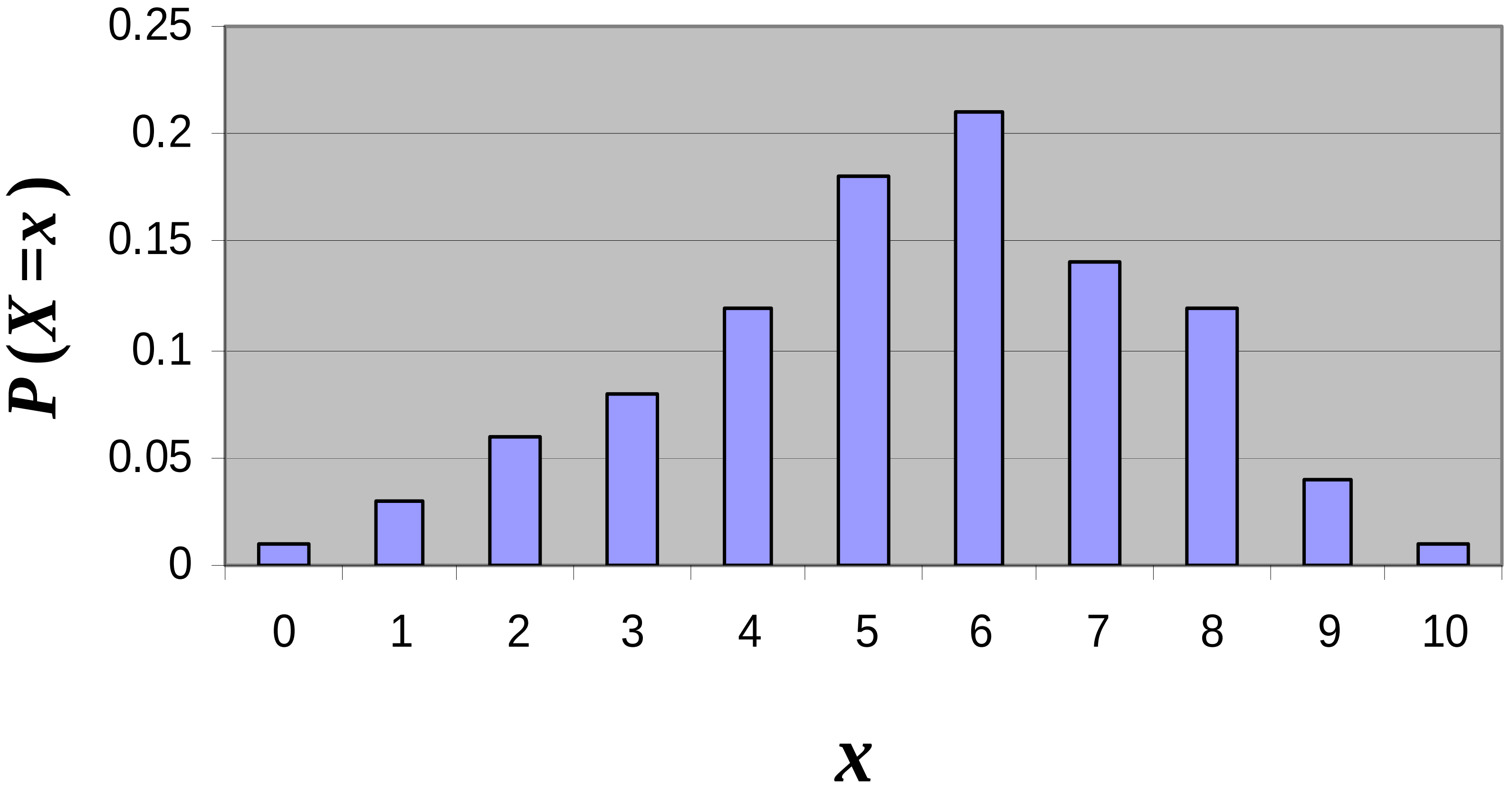
Discrete Probability Distribution

Definition: The probability distribution of a discrete random variable X is a table with all possible values that X assumes, i.e., $X=x$, with the frequency of occurrence of each value (probability)

Probability distribution of *X*, the number of seizures among 100 epileptic patients followed-up since initial diagnosis

x	Frequency of occurrence	$P(X=x)$	Prob
0	1	1/100	0.01
1	3	3/100	0.03
2	6	6/100	0.06
3	8	8/100	0.08
4	12	12/100	0.12
5	18	18/100	0.18
6	21	21/100	0.21
7	14	14/100	0.14
8	12	12/100	0.12
9	4	4/100	0.04
10	1	1/100	0.01
	100	100/100	1

Probability distribution of *X*, number of seizures among 100 followed-up epileptic patients



Properties of a probability distribution of a discrete random variable, *X*

Properties:

If x takes on values of $x_1, x_2, \ldots,$ then

Frequency function is $p(x_i) = P(X = x_i)$

$$0 \le p(x_i) \le 1$$

$$\sum_{i=1}^{n} p(x_i) = 1$$

Probability questions

What’s P(X=2)?=0.06

What’s P(X=6)?=0.21

What’s P(X=6) or P(X=7)? = 0.21 + 0.14= 0.35

Note: probabilities for mutually exclusive events can be added together according to Kolmogorov’s axiom:

When Events E1 and E2 are mutually exclusive,
$P(E1$ or $E2) = P(E1) + P(E2)$

Cumulative Distributions

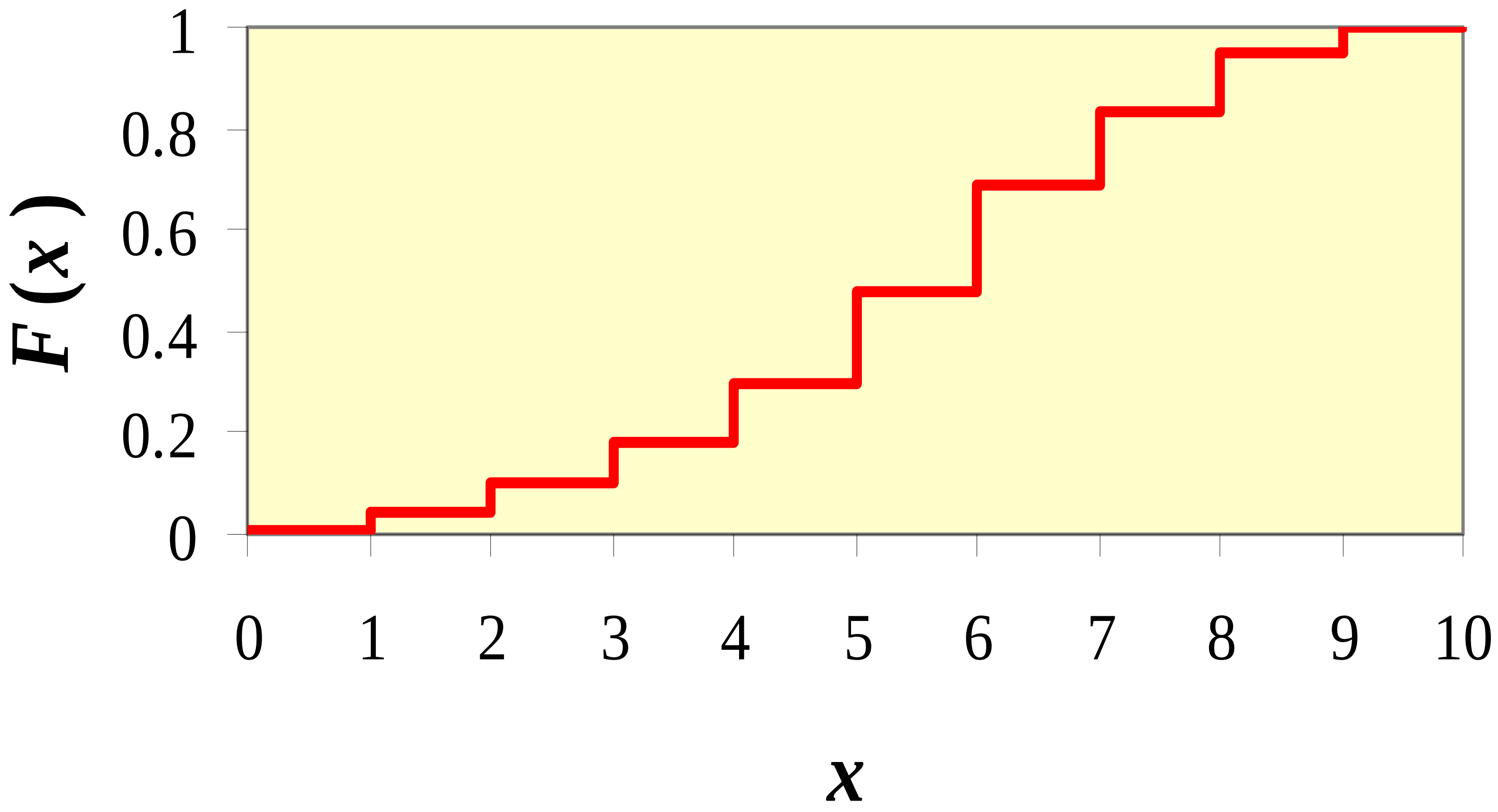
- Cumulative distributions can be more convenient to work with in some occasions
- Cumulative distributions are based on the sum of probabilities

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty$$

Cumulative Distribution Function, *F*(*x*) or "cdf"

x	Frequency of occurrence	P(*X* =*x*)	*p* (*x*)	*F* (*x*)
0	1	1/100	0.01	0.01
1	3	3/100	0.03	0.04
2	6	6/100	0.06	0.1
3	8	8/100	0.08	0.18
4	12	12/100	0.12	0.3
5	18	18/100	0.18	0.48
6	21	21/100	0.21	0.69
7	14	14/100	0.14	0.83
8	12	12/100	0.12	0.95
9	4	4/100	0.04	0.99
10	1	1/100	0.01	1
	100	100/100	1	

Cumulative Distribution Function for number of seizures among 100 epileptic patients



Note: the jump at x_i is $p(x_i)$

Properties of Cumulative Distribution Function

Properties:

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

Questions drawing on cdf

What’s the probability that a patient selected at random will have less than 5 seizures?

$$\begin{aligned} P(X<5) &= P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4) \\ &= 0.01+0.03+0.06+0.08+0.12 \\ &= 0.3 \end{aligned}$$

Questions drawing on cdf

What's the probability that a patient picked at random will have 5 or less (at most 5) seizures?

$$
\begin{aligned}
P(X \leq 5) &= P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5) \\
&= 0.01+0.03+0.06+0.08+0.12+0.18 \\
&= 0.48
\end{aligned}
$$

Questions drawing on cdf

What's the probability that a patient picked at random will have 5 or more (at least 5) seizures?

$$P(X \geq 5) = 1 - P(X < 5)$$

$$= 1 - 0.48$$

$$= 0.52$$

Binomial Distribution

- **Derived from Bernoulli trial:**

  Experiment or trial with only one of two mutually exclusive outcomes (failure or success)

- **Bernoulli Process:**
  - Each trial has one of two mutually exclusive outcomes
  - Probability of success remains constant from trial to trial
  - Trials are independent: outcome of any particular trial does not influence outcome of any other trial

Bernoulli Process

Trials: flip coin 3 times
Outcome:HTH
Coding: 101

$$\begin{aligned} P(1,0,1) &= pqp \\ &= p^2 q \end{aligned}$$

$$\begin{aligned} P(0,1,1) &= qpp \\ &= p^2 q \end{aligned}$$

$$\begin{aligned} P(1,1,0) &= ppq \\ &= p^2 q \end{aligned}$$

Binomial Probability

- Interest is the probability of x successes and n-x failures in n trials
- p is probability of success, 1-p is probability of failure (q=1-p)
- A sequence of x successes within n trials (n-x failures) occurs with probability

$$p^x(1-p)^{n-x}$$

- Number of ways to assign x success to n trials is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Binomial Probability

Fair coin is flipped three times (n=3). How many sequences are there for 2 heads (x=2) among the three flips?

	Possible outcome of single experiment with three trials							
Trial	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	**H**	**H**	**H**	**H**	**T**	**T**	**T**	**T**
2	**H**	**H**	**T**	**T**	**T**	**T**	**H**	**H**
3	**H**	**T**	**H**	**T**	**T**	**H**	**T**	**H**
		↑	↑					↑

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$$\binom{3}{2} = \frac{3!}{2!(3-2)!}$$

$$= \frac{3\times 2}{2}$$

$$= 3$$

Permutations of like objects

Definition: The total number of permutations of all n distinct objects for which r_1 are alike, r_2 are alike,…, r_k are alike is

$$N = \frac{n!}{r_1!r_2!\times\cdots\times r_k!}$$

r_1=2 *heads*
r_2=1 *tail*
HHT
HTH
THH

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$$

Binomial Probability

Probability of *x*=2 successes (heads) in *n*=3 independent trials (flips) is

$$
\begin{aligned}
P(x) &= \binom{n}{x} p^{x}(1-p)^{n-x} \\
&= \binom{3}{2} 0.5^{2}(0.5)^{3-2} \\
&= 3(0.25)(0.5) \\
&= 3(0.125) \\
&= 0.375 = \frac{3}{8}
\end{aligned}
$$

Binomial Probability

	Possible outcome of single experiment with three trials							
Trial	1	2	3	4	5	6	7	8
1	H	H	H	H	T	T	T	T
2	H	H	T	T	T	T	H	H
3	H	T	H	T	T	H	T	H
Prob	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Table of Binomial Coefficients

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

Binomial Frequency Function, B(*n*,*p*)

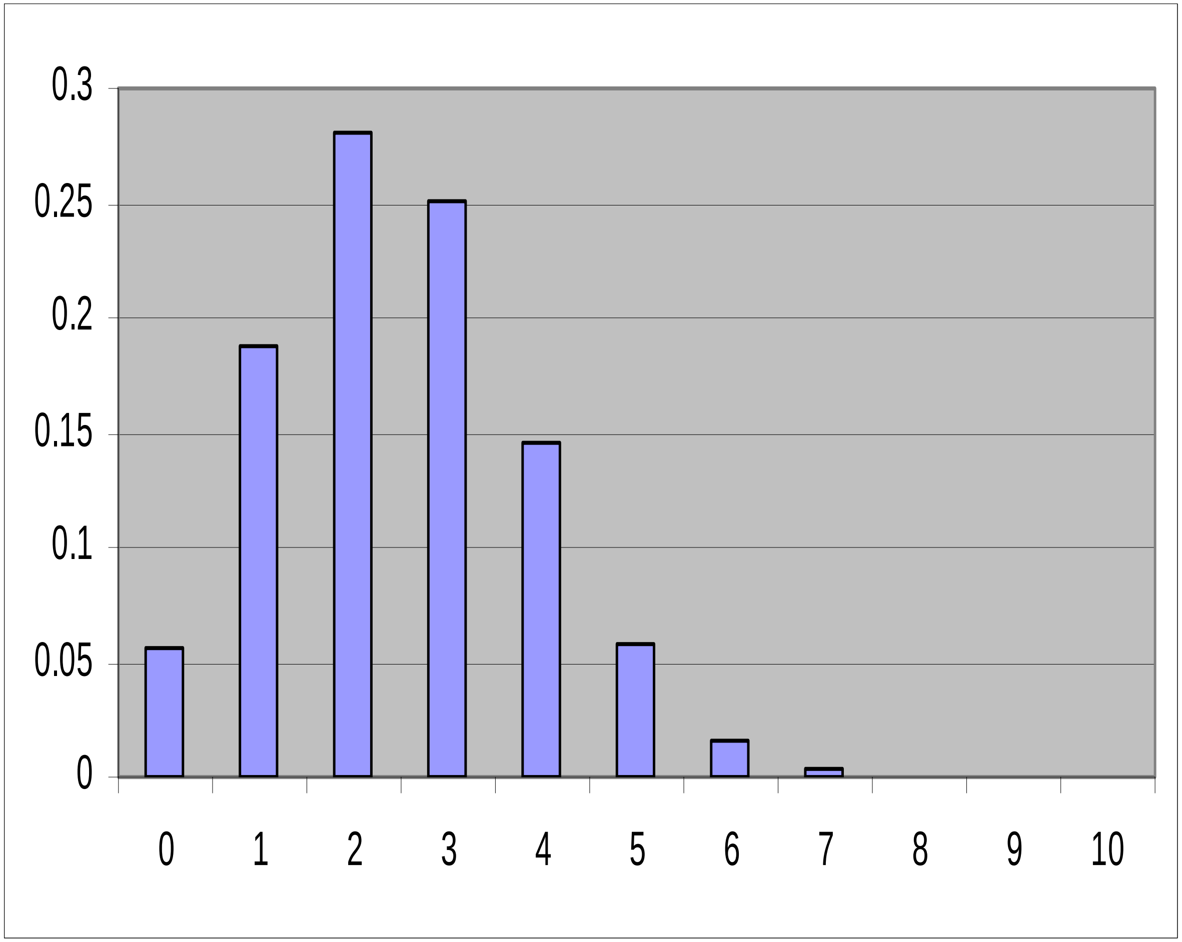
- If the probability that an offspring is affected with a trait is 0.25, then what is the frequency function for the probability that *X* offspring are affected out of 10?

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \binom{10}{x} 0.25^x (0.75)^{10-x} \qquad x = 0,1,2,\ldots,10$$

B(10, 0.25)

x	$P(X=x)$
0	0.056314
1	0.187712
2	0.281568
3	0.250282
4	0.145998
5	0.058399
6	0.016222
7	0.003090
8	0.000386
9	0.000029
10	0.000001



Poisson Frequency Function

For large *n* and small *p* ($npq \le 5$), the binomial is approximated by the Poisson distribution

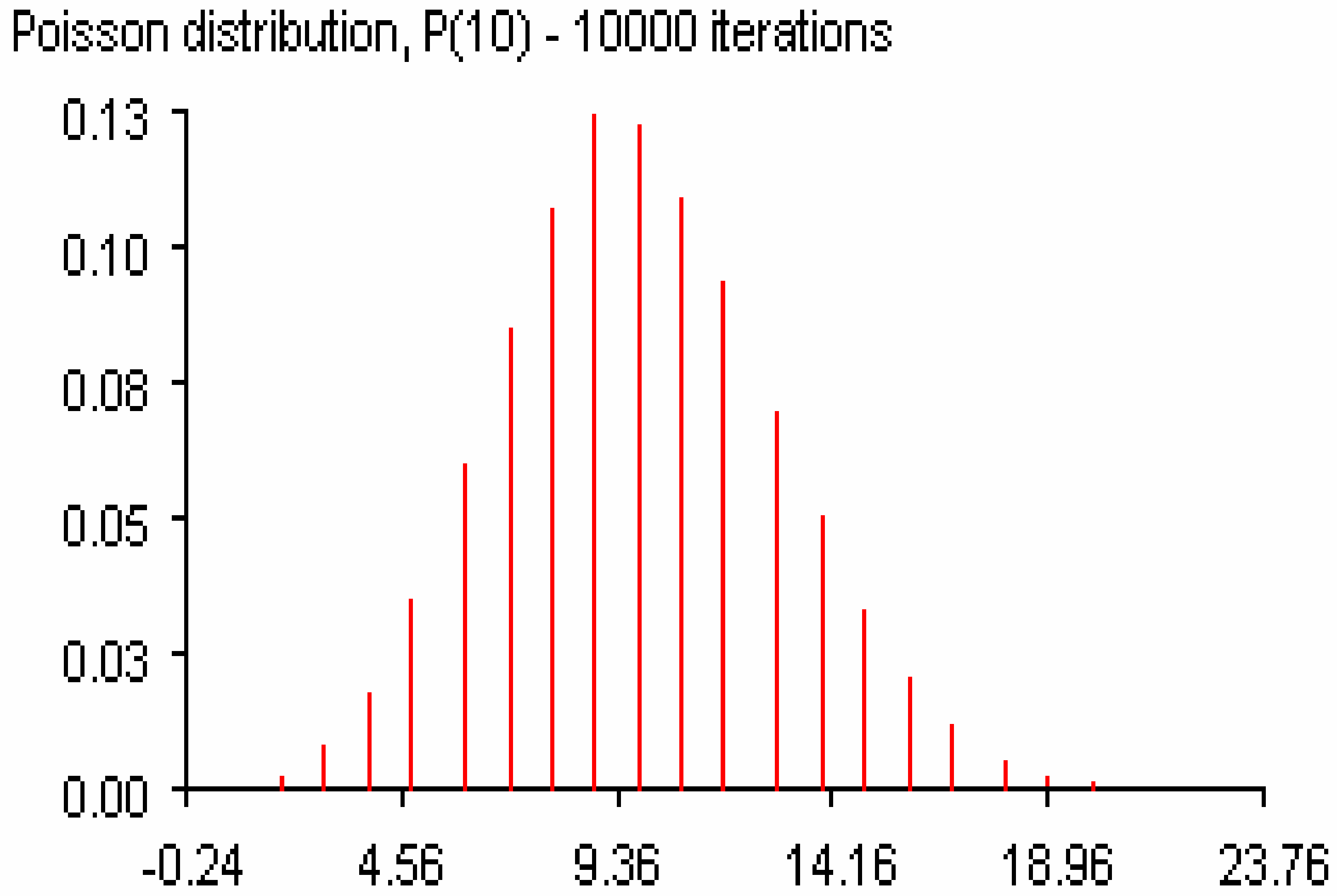
$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, \ldots$$

where $\lambda = np$ is the rate parameter.

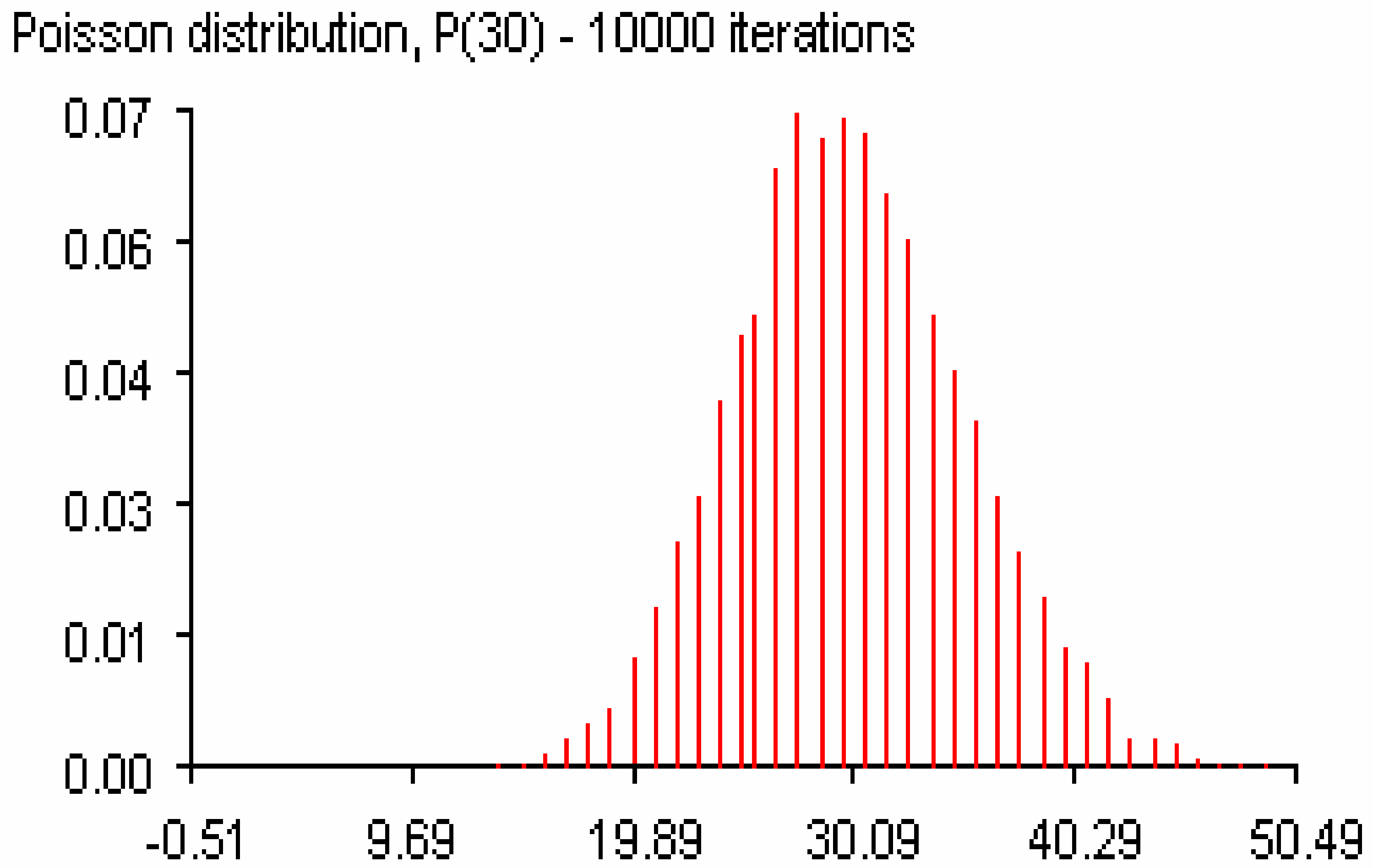
Example: If a telephone rings 6 times a day, what is the probability that it will ring 4 times on a given day?

$$P(X = 4) = \frac{6^4}{4!} e^{-6} = \frac{(1296)(0.0025)}{24} = 0.135$$

Poisson Distribution (λ=10), “P(10)”



P(30)


Poisson distribution, P(30) - 10000 iterations
0.07
0.06
0.04
0.03
0.01
0.00
-0.51
9.69
19.89
30.09
40.29
50.49

Multinomial Distribution

Definition: When there are more than *k* possible outcomes per trial, the probabilities for each outcome remaining the same for each trial, and the trials are independent, the probability of x_1 outcomes of the first, x_2 of the second kind,…, and x_k of the *k*th kind in *n* trials is

$$P(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\,x_2!\times\cdots\times x_k!} p_1^{x_1} p_2^{x_2} \times\cdots\times p_k^{x_k}$$

Multinomial Probabilities

Assume a single locus two-allele model (no HWE). Genotype frequencies for *AA*, *Aa*, and *aa* are P_{AA}=0.23, P_{Aa}=0.47, P_{aa}=0.30. What is the probability of observing 1 *AA*, 3 *Aa*, and 6 *aa* genotypes among 10 genotyped subjects?

$$P(1AA, 3Aa, 6aa) = \frac{10!}{1!3!6!}(0.23)^1(0.47)^3(0.30)^6$$

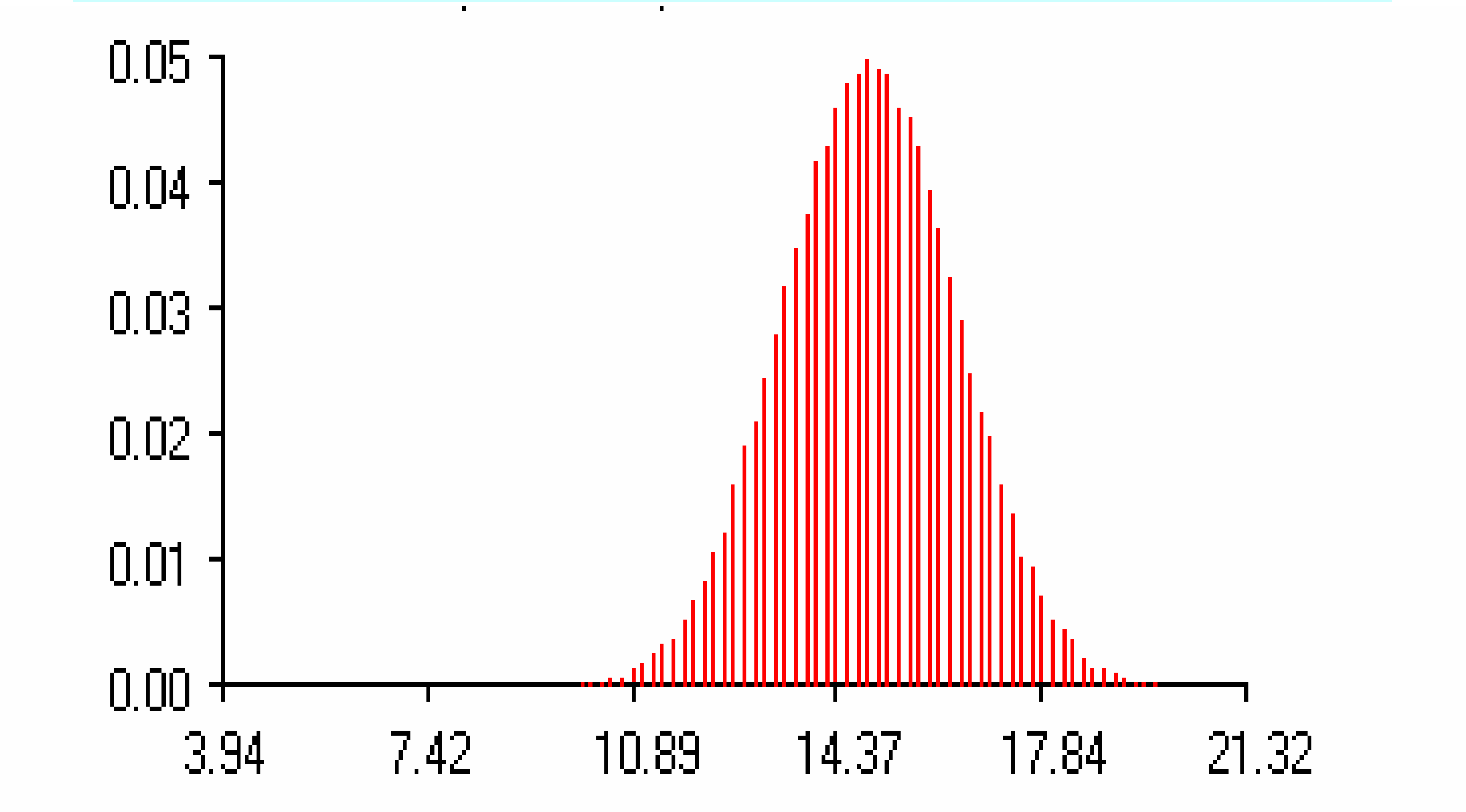
$$= 0.0146$$

Continuous Distribution
Normal Distribution, N(15,2.25)

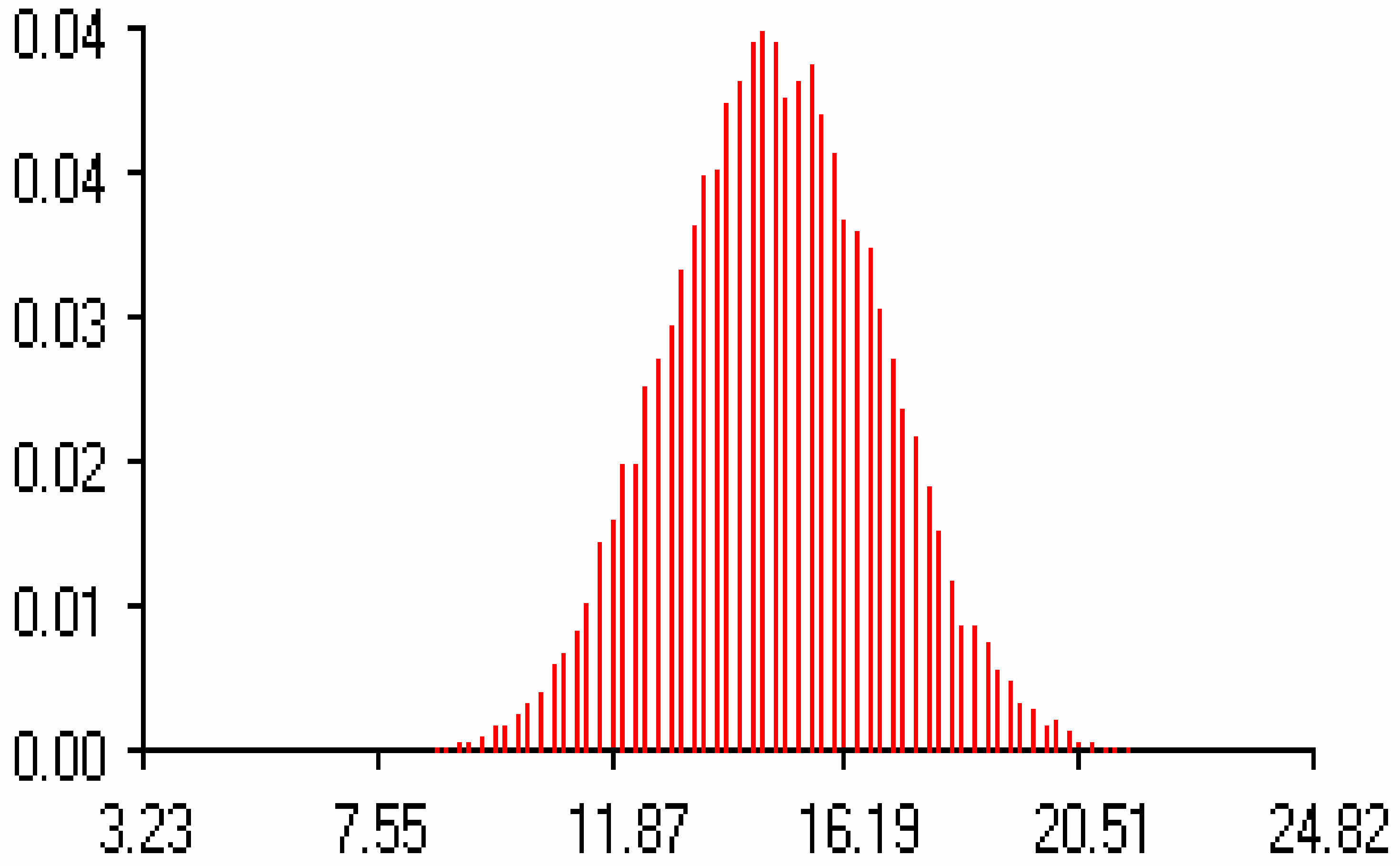
- Normal distribution has two parameters, mean, μ, and variance, σ^2.
- Probability density function (pdf) is:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2} \qquad -\infty < x < \infty$$

N(15,2.25)



N(15,4)



Standard Normal Distribution

$$\text{Consider } Z = \frac{x-\mu}{\sigma}$$

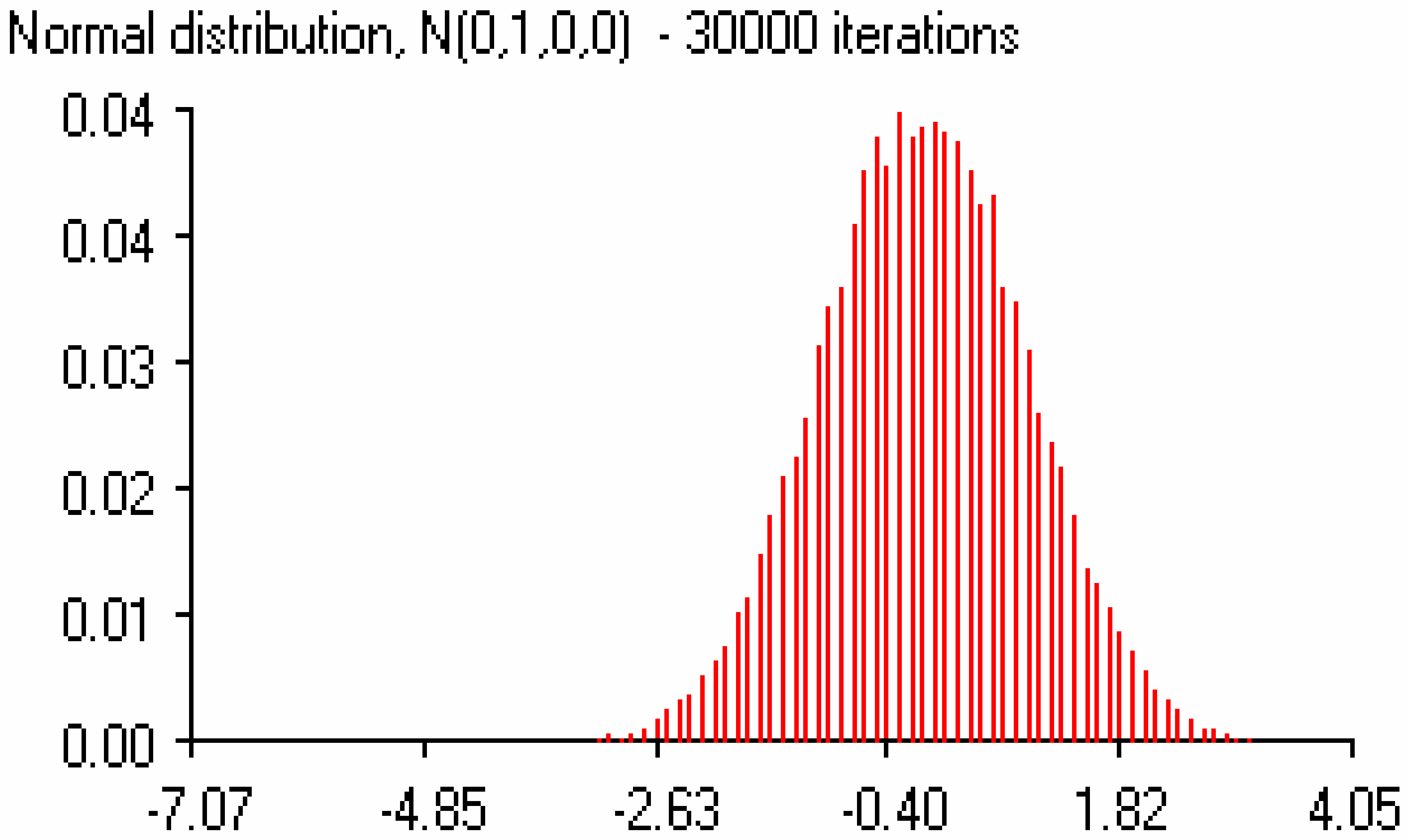
We know that

$$
\begin{aligned}
F(x) &= P(X \le x) \\
&= \mathrm{P}\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) \\
&= P\left(Z \le \frac{x-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right)
\end{aligned}
$$

thus

$$P(x_0 < X < x_1) = F(x_1) - F(x_0) = \Phi(x_1) - \Phi(x_0)$$

Standard Normal Distribution, N(0,1)



Standard Normal Distribution

Average patient weight, μ=150 LBS, σ=15.

What is the probability that a randomly selected patient weighs between 120 to 130 LBS?

$$
\begin{aligned}
P(120 < X < 130) &= \Phi\left(\frac{130-150}{15}\right) - \Phi\left(\frac{120-150}{15}\right) \\
&= \Phi(-1.33) - \Phi(-2) \\
&= (1-0.9082) - (1-0.9972) \\
&= 0.0918 - 0.0228 \\
&= 0.069
\end{aligned}
$$

Standard Normal Distribution

$$P(X \geq 170) = P\left(Z \geq \frac{170 - 150}{15}\right)$$

$$= \Phi(1.33)$$

$$= 1 - 0.9082$$

$$= 0.0918$$

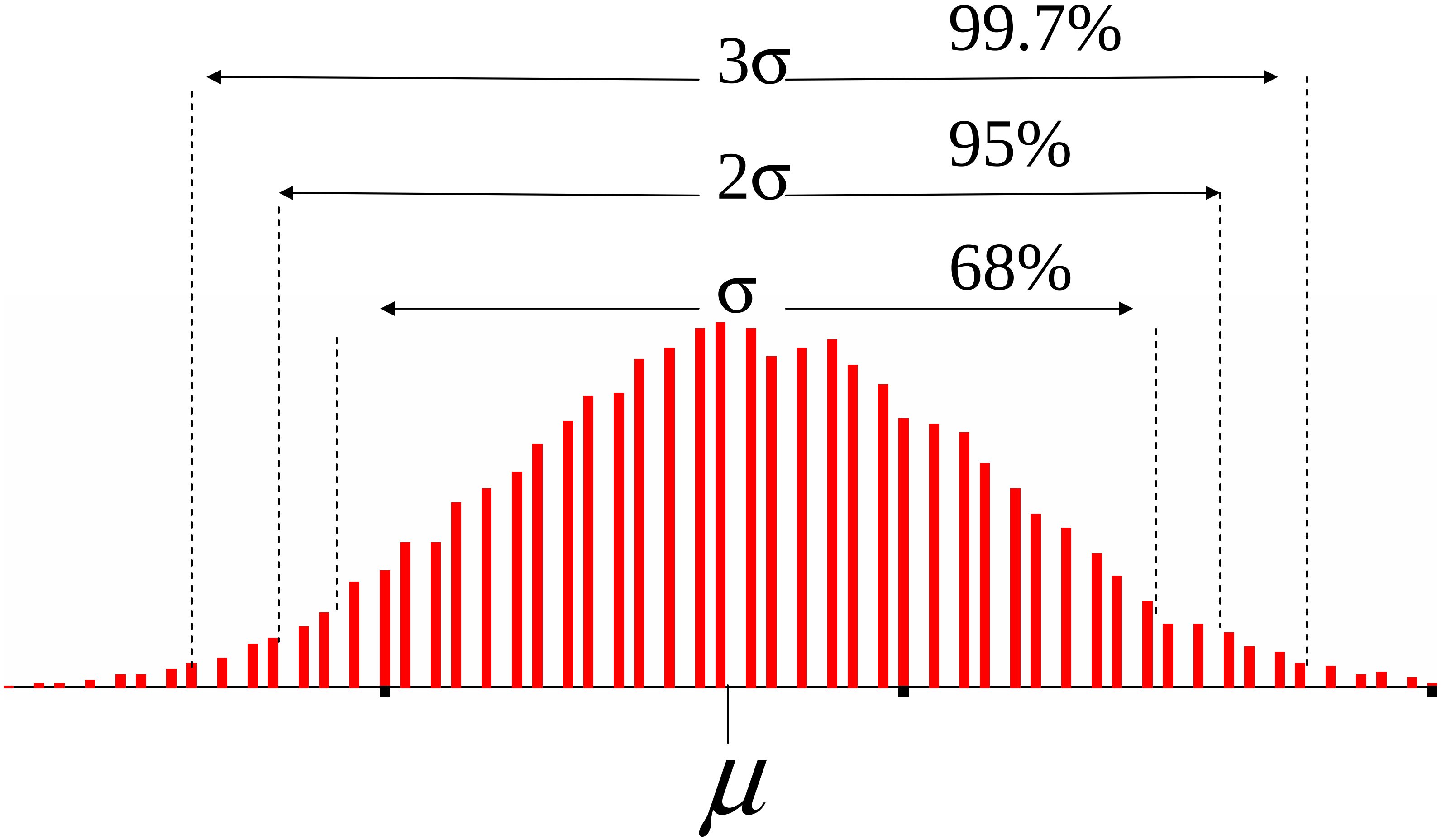
Standard Normal Distribution

$$\begin{aligned} P(-1 < X < 1) &= \Phi(1) - \Phi(-1) \\ &= 0.8413 - (1 - 0.8413) \\ &= 0.68 \end{aligned}$$

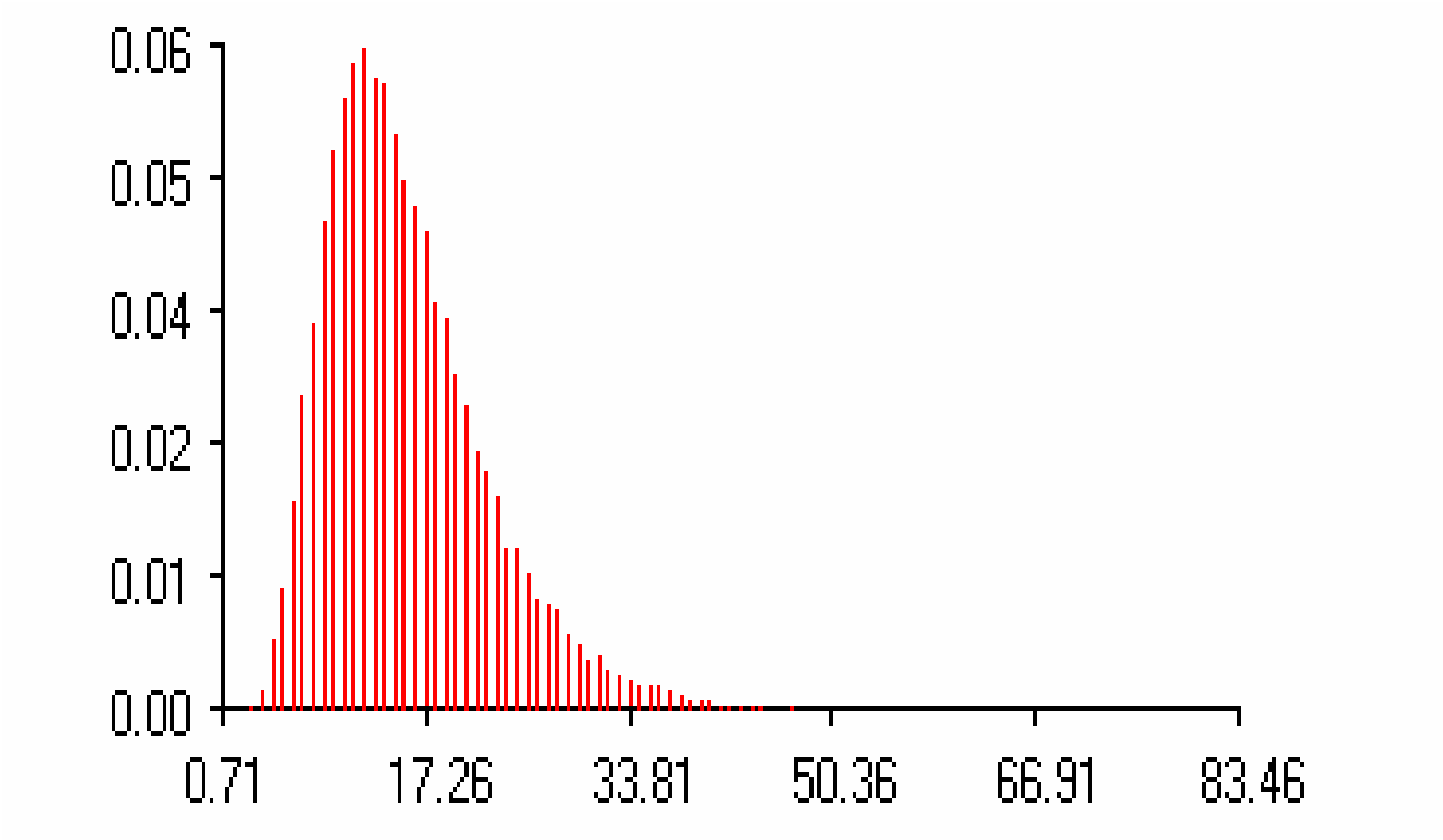
$$\begin{aligned} P(-2 < X < 2) &= \Phi(2) - \Phi(-2) \\ &= 0.9772 - (1 - 0.9772) \\ &= 0.95 \end{aligned}$$

$$\begin{aligned} P(-3 < X < 3) &= \Phi(3) - \Phi(-3) \\ &= 0.9987 - (1 - 0.9987) \\ &= 0.9974 \end{aligned}$$

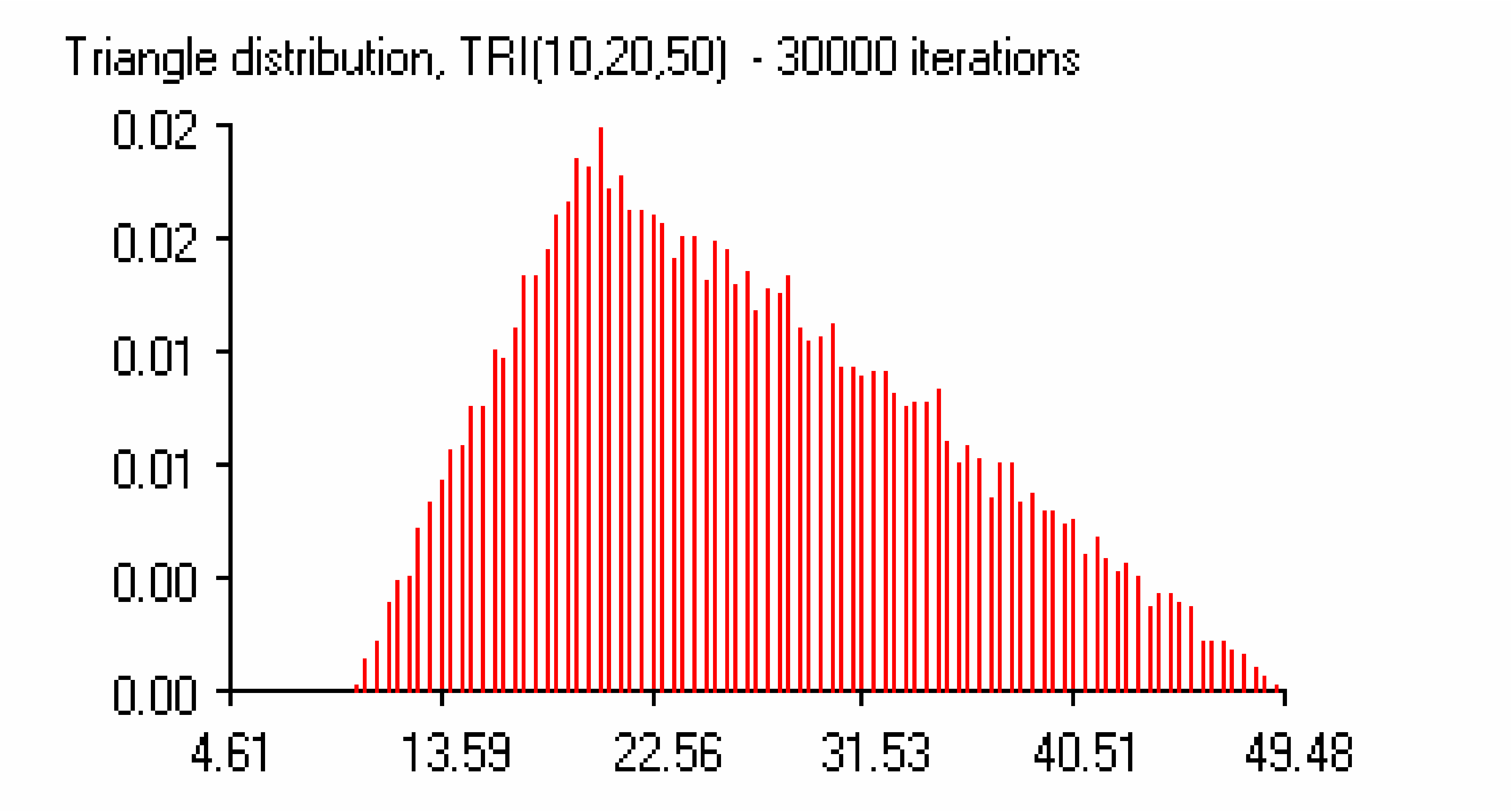
Standard deviations



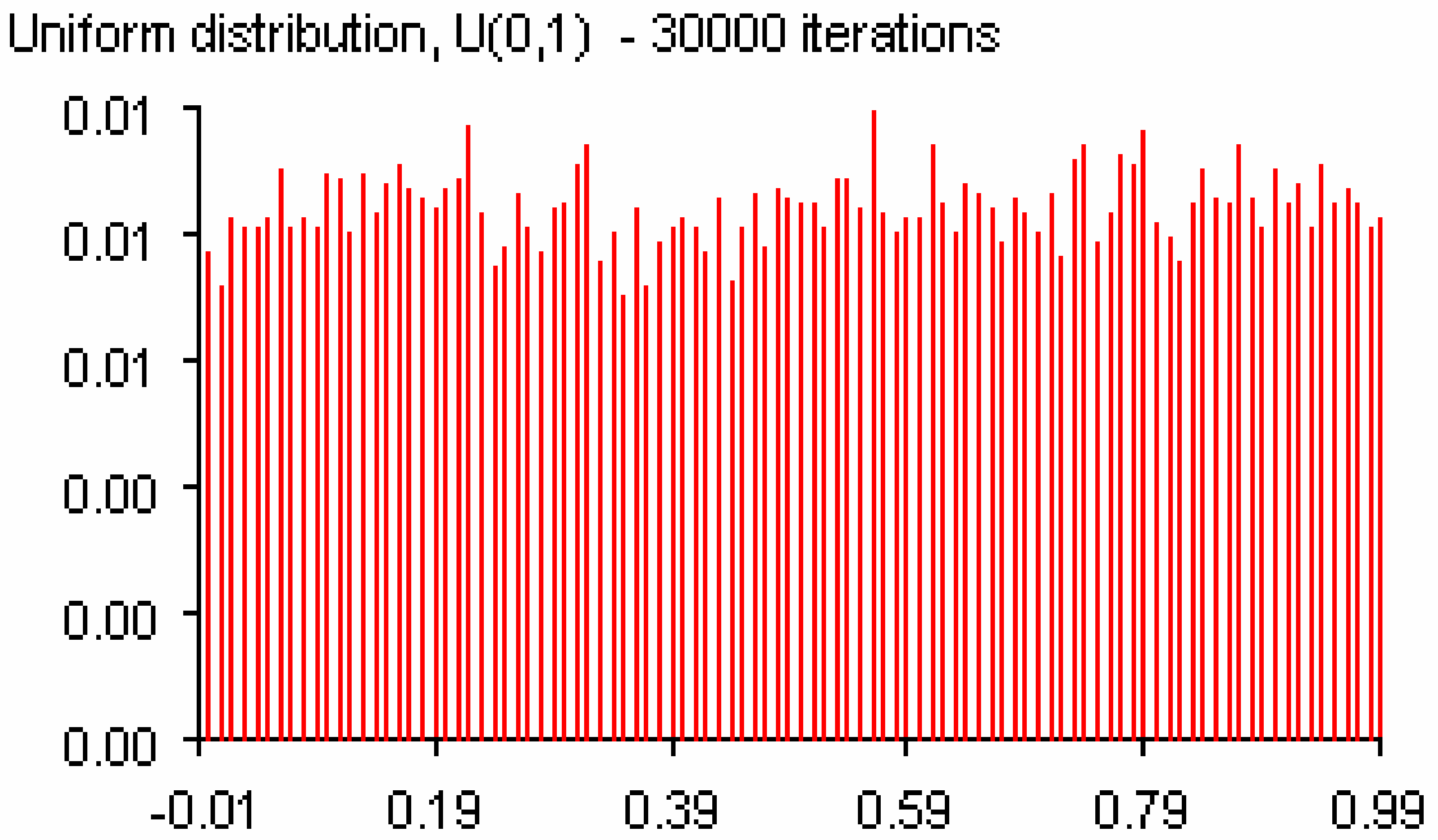
Log-normal Distribution, ln(Y) is distributed N(μ,σ^2), since Y=e^Z



Triangle Distribution
(Lack-of-knowledge Distribution)



Standard Uniform Distribution, U(0,1)



References

Daniel,W.W. Biostatistics: A Foundation for Analysis in the Health Sciences. New York(NY), Wiley, 1974.

Rice, J.A. Mathematical Statistics and Data Analysis. 2nd Ed. Belmont(CA), Wadsworth-Duxbury, 1995.

Freund, J.E., Williams, F.J. Elementary Business Statistics: The Modern Approach. Englewood Cliffs(NJ), Prentice-Hall, 1982.