BIM 105
Probability and Statistics for Biomedical Engineers

David M. Rocke

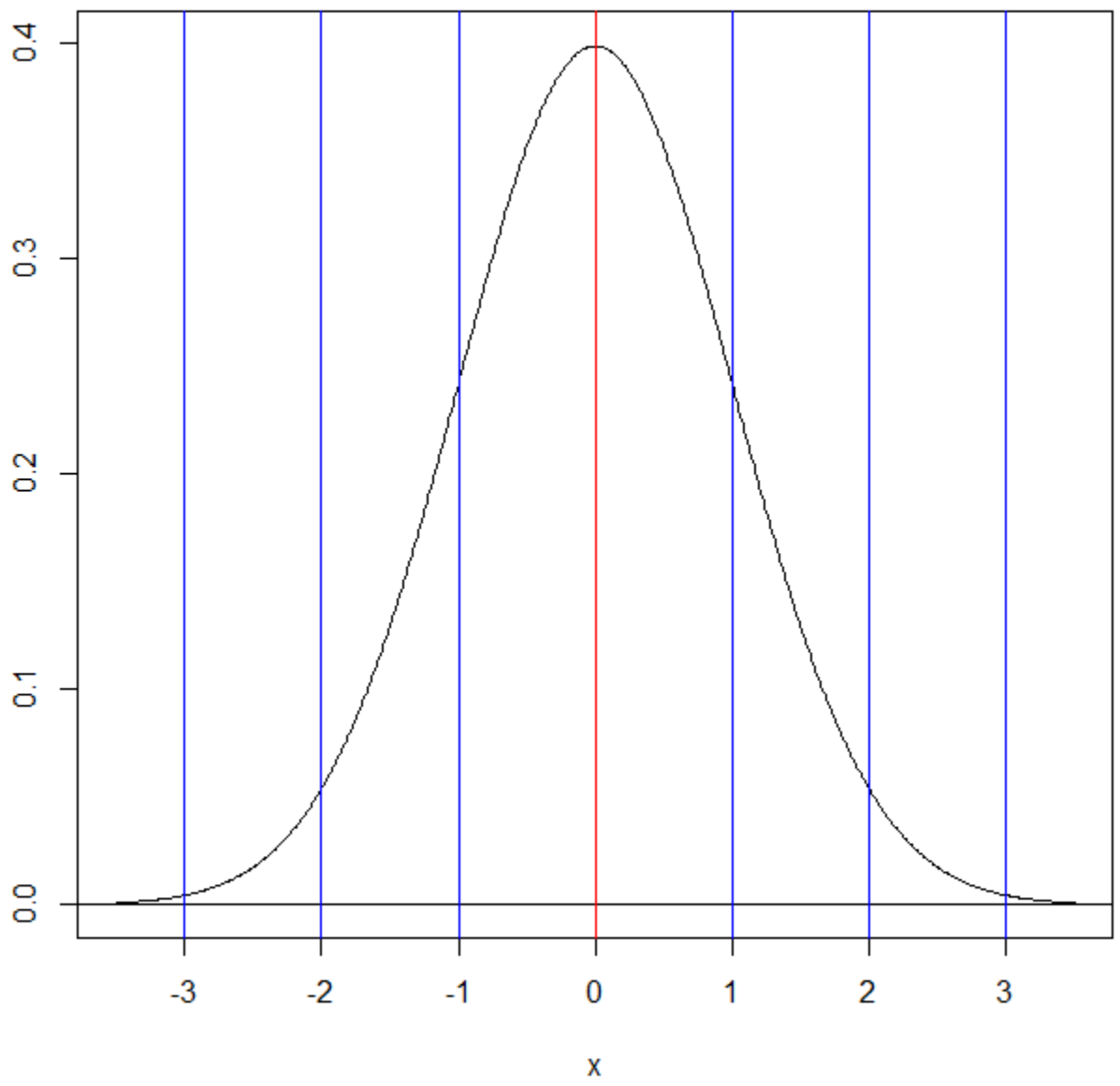
Department of Biomedical Engineering

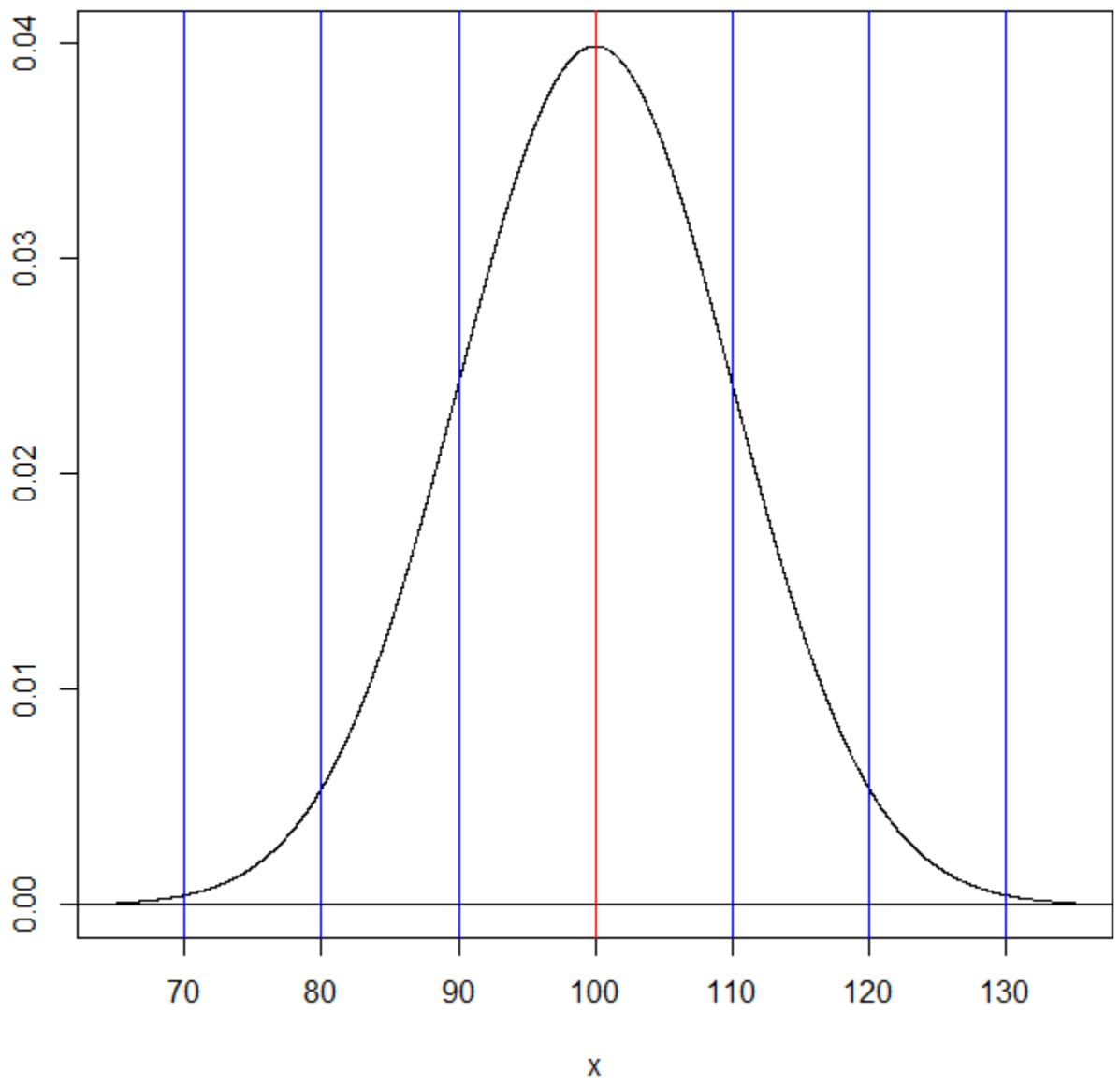
Types of Continuous Data and Distributions

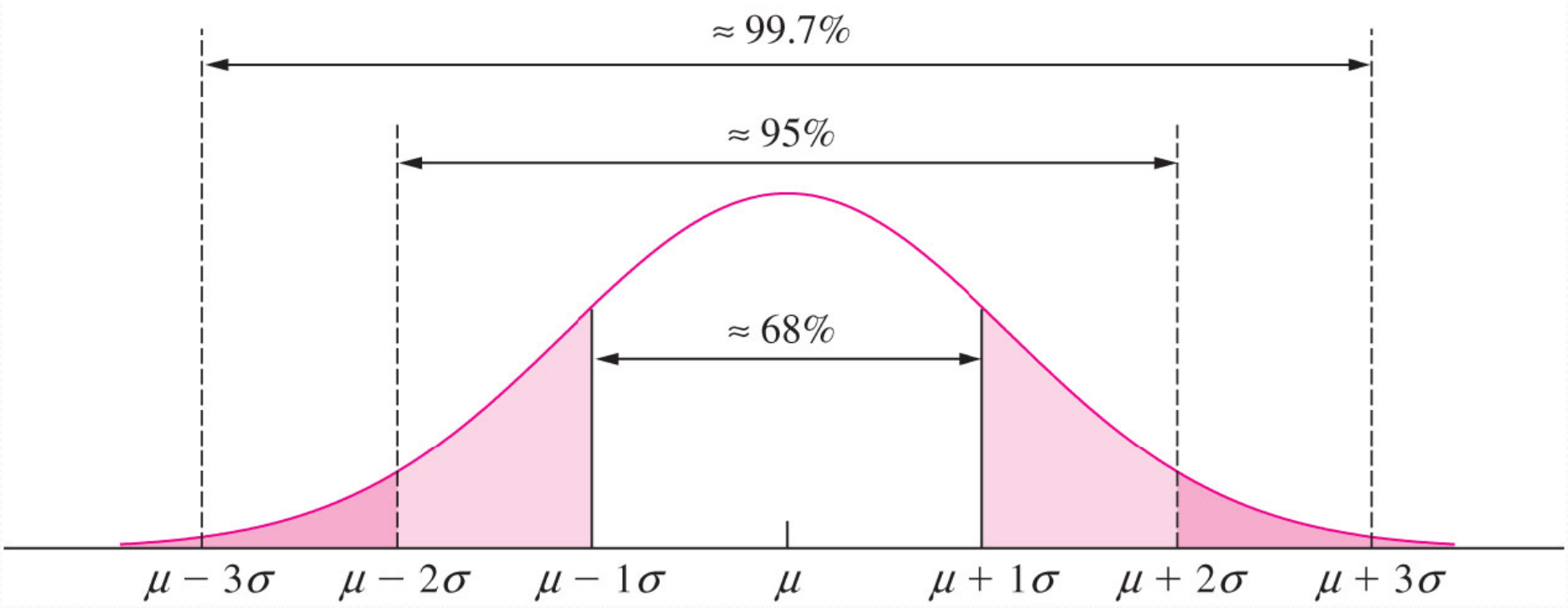
- Continuous data arise from measurements, logs of measurements, or differences of measurements.
- The range of the data is often either the whole real line $(-\infty, \infty)$, the non-negative reals $[0, \infty)$, or sometimes an interval.
- The amount of toluene in a sample (ng/L) lies in $[0, \infty)$.
- This may be based on the peak height above the baseline from a GC/MS, which can be negative, so we treat it as lying in $(-\infty, \infty)$.
- If we round the reading to a whole number of ng/L, then the rounding error lies in [-0.5, 0.5] ng/L.

The Gaussian (Normal) Distribution

- For data that have a "mound-shaped" distribution, we often use the Gaussian/Normal.
- Some people avoid the term "Normal" because it implies that this is ordinarily the distribution that data will come from and that it is abnormal for the distribution to be otherwise, both of which are contrary to fact and experience.
- Nonetheless, it is the most important distribution in statistics, for reasons that will emerge later.
- There is a Gaussian distribution for every choice of mean μ and standard deviation $\sigma > 0$.


Normal Distribution with Mean 0 and Standard Deviation 1
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x


Normal Distribution with Mean 100 and Standard Deviation 10
0.00
0.01
0.02
0.03
0.04
70
80
90
100
110
120
130
x



Normal distribution with mean μ and standard deviation σ

The PDF is

$$f(x) = \phi(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$\phi(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$X \sim N(\mu, \sigma^2)$$

The CDF is by definition

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

$$\Phi(x; \mu, \sigma^2) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/(2\sigma^2)} dt$$

There is no closed-form solution to this integration, so we need a table or computer program.

How to Use a Table of the Gaussian/Normal Distribution

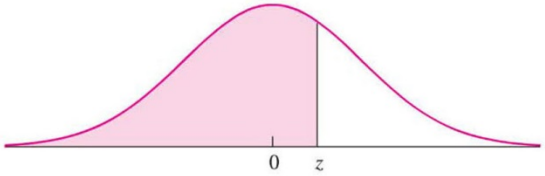
- The table A.2 is expressed in standard units, so that it gives the CDF for a standard normal, one with mean 0 and variance 1.
- So any question needs to be translated to standard units, sometimes called z-scores.
- For example, suppose the amount of calcium in a liter of cell-growth medium is supposed to average 40gm, and has a standard deviation of 2.5gm.
- 42 gm is (42-40)/2.5 = 0.8 in standard units.
- If the amount of calcium is normally distributed, then the chance that there is less than 42gm is the same as the chance that at standard normal random variable is less than 0.8, which from the table is 0.7881.

- For example, suppose the amount of calcium in a liter of cell-growth medium is supposed to average 40gm, and has a standard deviation of 2.5gm.
- To find the probability that the amount of calcium is between 37 and 42, we translate this into a statement about a standard normal random variable.

$$P(37 < X < 42) = P\left(\frac{37-40}{2.5} < Z < \frac{42-40}{2.5}\right)$$
$$= P(-1.20 < Z < 0.80)$$
$$= P(Z < 0.80) - P(Z < -1.20)$$
$$= 0.7881 - 0.1151$$
$$= 0.6730$$

$$P(X < 41.7) = P(Z < 0.68) = 0.7517$$
$$P(X \geq 38.5) = P(Z \geq -0.6)$$
$$= 1 - P(Z < -0.6)$$
$$= 1 - 0.2743$$
$$= 0.7257$$

TABLE A.2 Cumulative normal distribution (continued)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Normal CDF in MATLAB

```
>> normcdf(0.8) % cdf at 0.8 of the standard normal

ans =

    0.7881

>> normcdf(42,40,2.5) % cdf at 42 of a normal with mu = 40 and sigma = 2.5

ans =

    0.7881

>> normcdf(42,40,2.5)-normcdf(37,40,2.5) % area between 37 and 42

ans =

    0.6731

>>
```

Percentage Points of the Normal

- $P(Z < z) = 0.75$
- What is the value of z?
- Normal Table A.2
  - $P(Z < 0.67) = 0.7486$
  - $P(Z < 0.68) = 0.7517$
  - Interpolating, $z = 0.6745$
- Bottom line of Table A.3
  - 0.674
- MATLAB or another computer program
  - 0.6745
  - 0.6744898

Percentage Points of the Normal/Gaussian Distribution

$$P(Z > z_\alpha) = \alpha$$

$$z_{0.25} = 0.6745 \approx 2/3$$

$$P(-z_{0.25} < Z < z_{0.25}) = 0.50$$

$$P(-2/3 < Z < 2/3) = 0.4950$$

$$z_{0.025} = 1.9600 \approx 2$$

$$P(-z_{0.025} < Z < z_{0.025}) = 0.95$$

$$P(-2 < Z < 2) = 0.9545$$

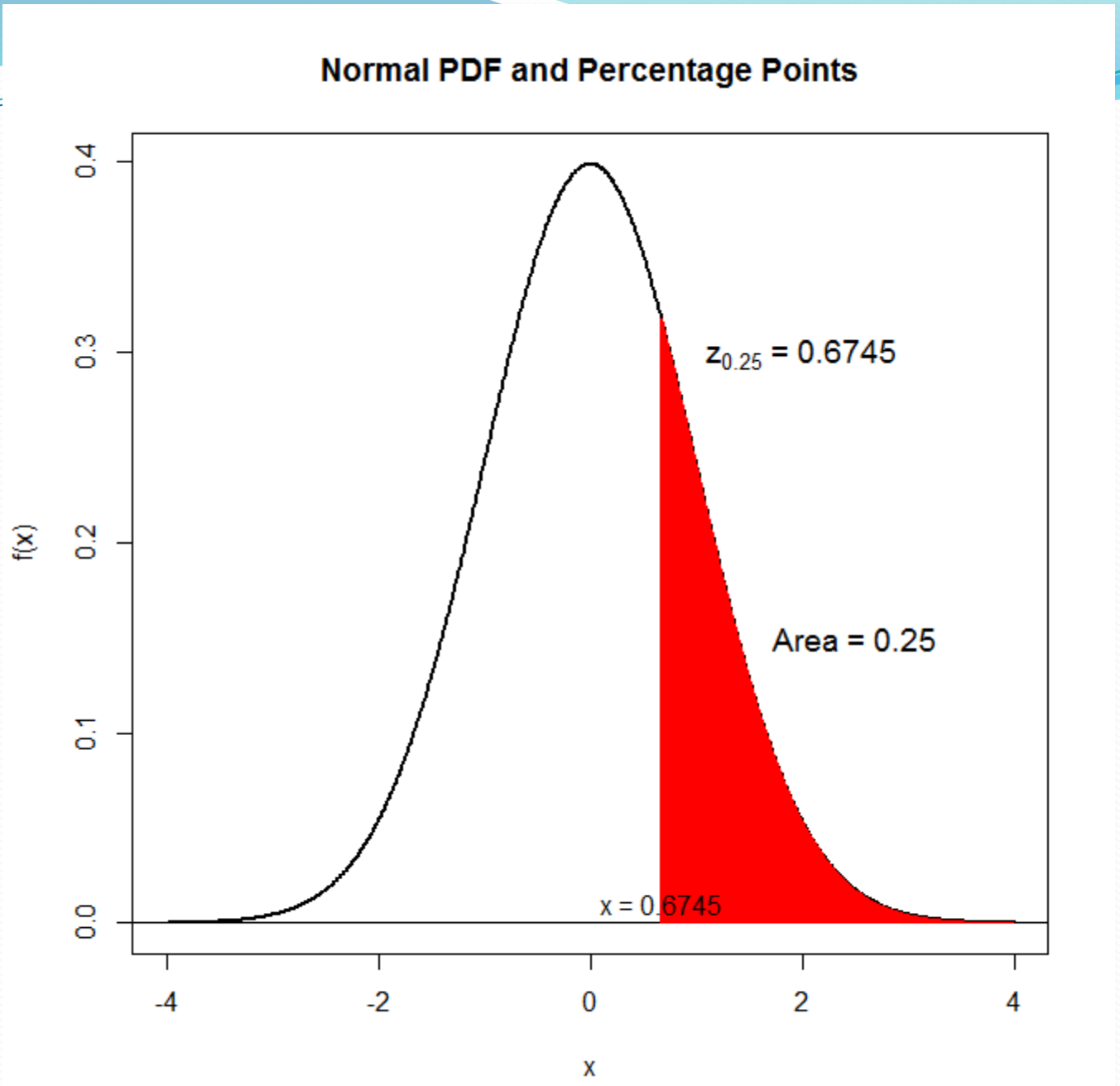
$$z_{0.001} = 3.0902 \approx 3$$

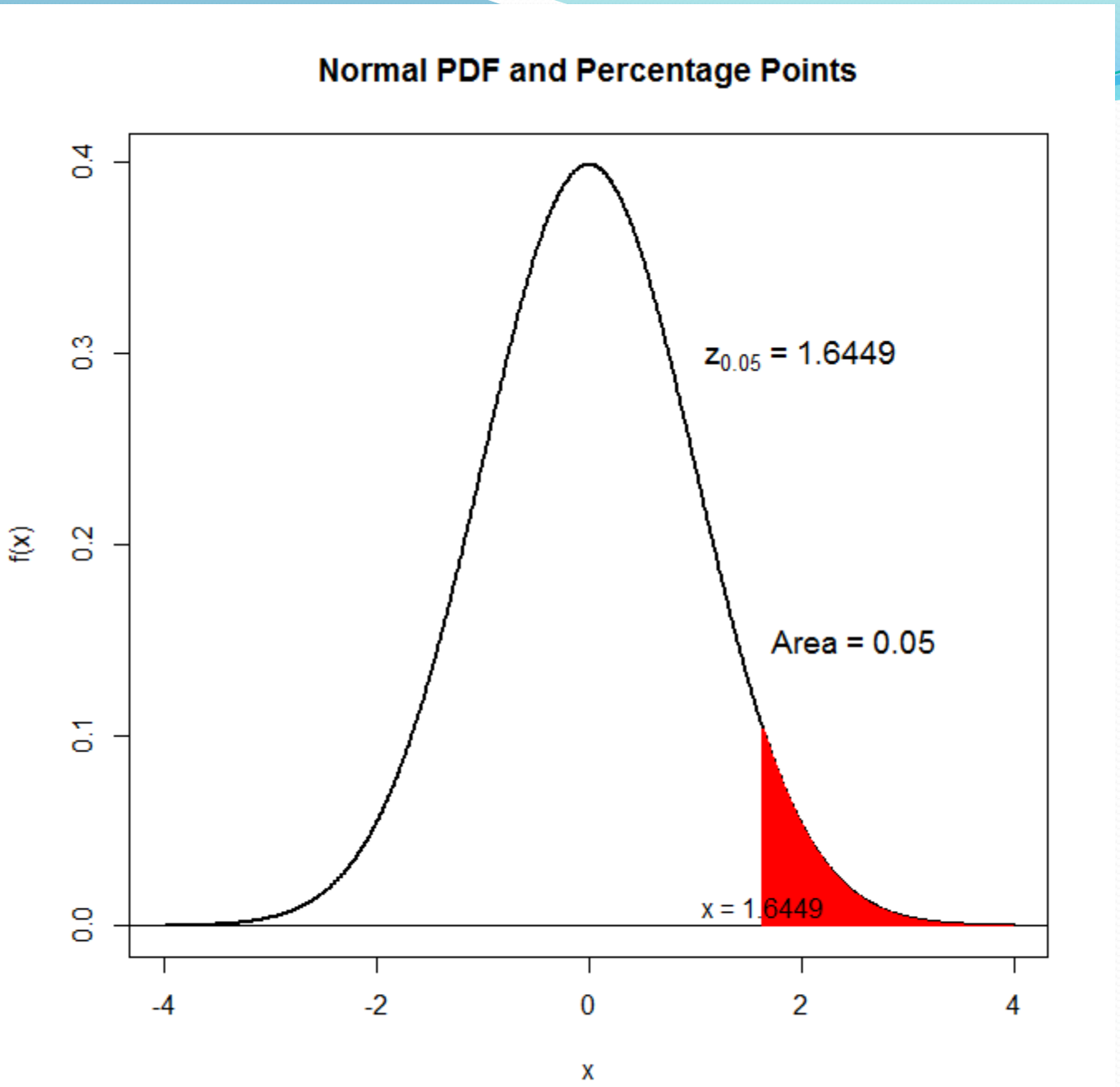
$$P(-z_{0.001} < Z < z_{0.001}) = 0.998$$

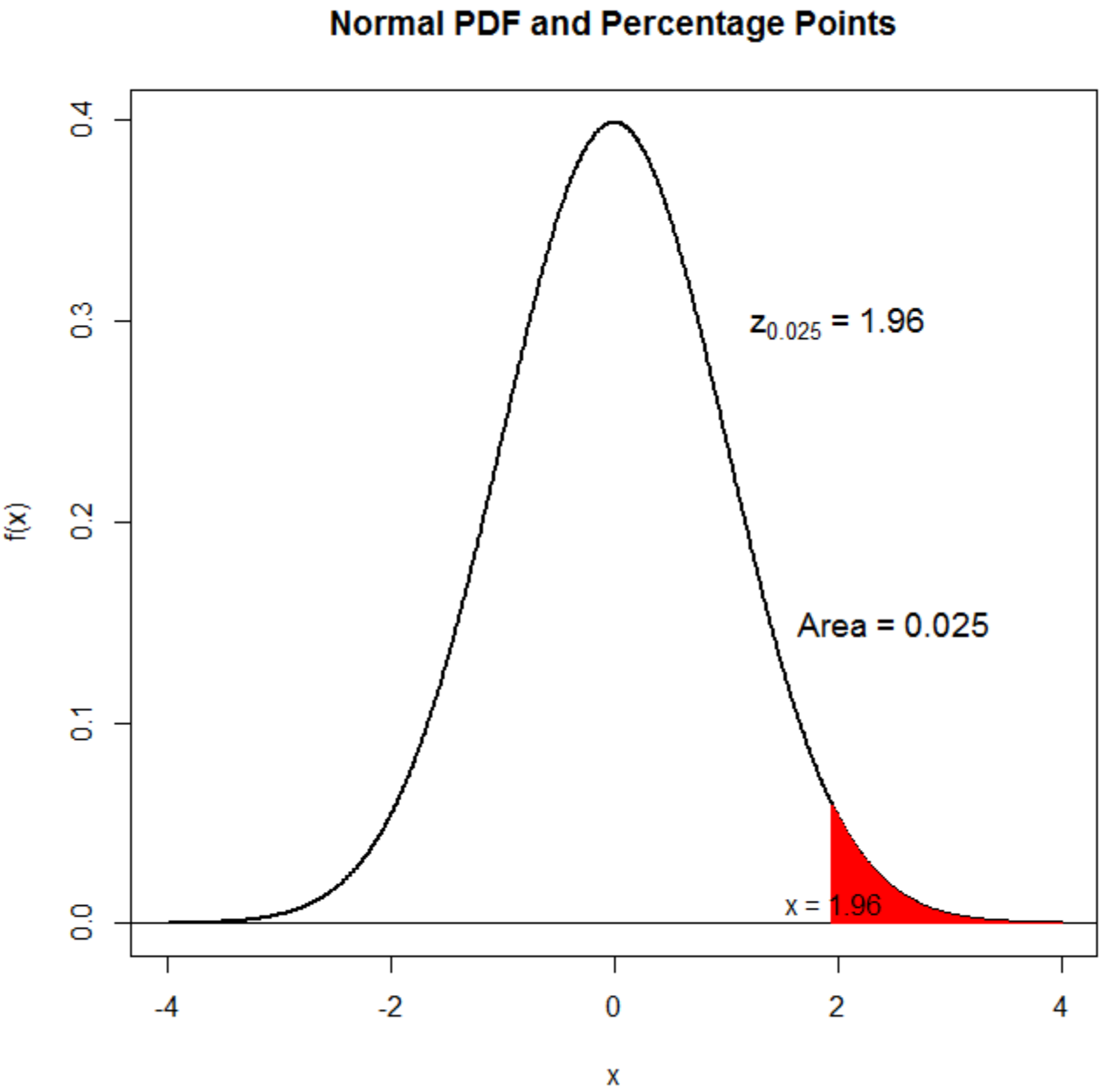
$$P(-3 < Z < 3) = 0.9973$$

Normal Inverse CDF in MATLAB

```
>> norminv(.025) % Point on standard normal w/ 0.025 to the left

ans =

   -1.9600

>> norminv(.975) % Point on standard normal w/ 0.025 to the
                 % right, or 0.975 to the left. Z_0.025
ans =

    1.9600

>> norminv(.975,100,10) % 0.025 point when mu = 100, sigma = 10

ans =

  119.5996
```


Normal PDF and Percentage Points
f(x)
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
$z_{0.25}$ = 0.6745
Area = 0.25
x = 0.6745


Normal PDF and Percentage Points
f(x)
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
$z_{0.05}$ = 1.6449
Area = 0.05
x = 1.6449


Normal PDF and Percentage Points
f(x)
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
$z_{0.025}$ = 1.96
Area = 0.025
x = 1.96

- Suppose the daily electricity cost for a bioprocessing plant has mean $832 and standard deviation $126.
- If the daily cost is normally distributed, what is the cost level that will not be exceeded more than 10% of the time?
- $z_{0.10} = 1.2816$
- Cost = $832 + (1.2816)($126) = $993.48

Combinations of Normal Random Variables

If X is a normal/Gaussian random variable, then so is $aX + b$

for any constants a and b

The sum or difference of any number of normal/Gaussian random variables

is also a normal/Gaussian random variable

$$E(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1E(X_1) + c_2E(X_2) + \cdots + c_nE(X_n)$$

$$V(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_n^2V(X_n) \quad \text{(if the } X_i \text{ are independent)}$$

Suppose that the X_i are independent random variables and

$$X_i \sim N(\mu, \sigma^2)$$

Then

$$E(X_{\text{sum}}) = E(X_1 + X_2 + \cdots + X_n) = n\mu$$

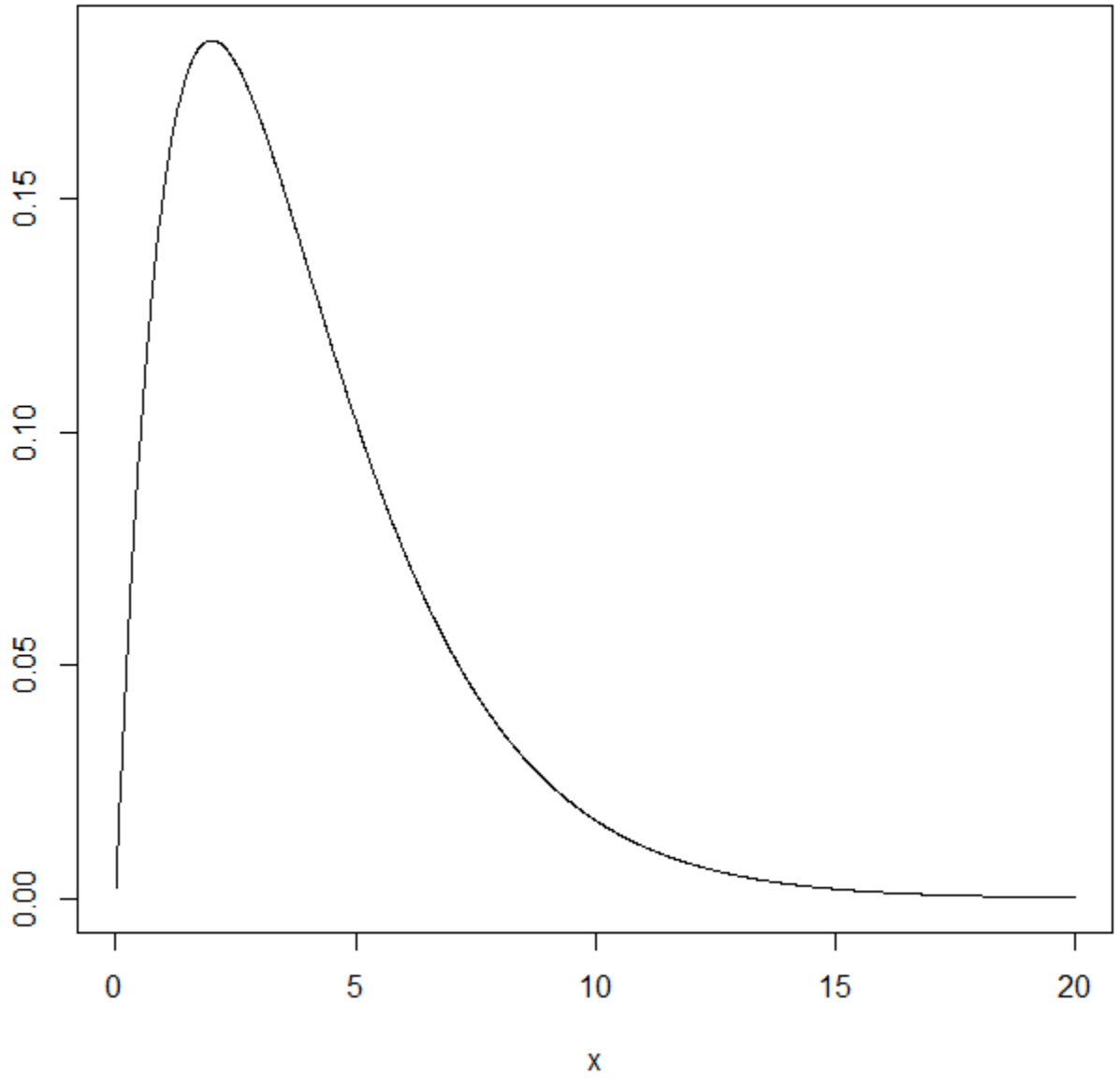
$$V(X_{\text{sum}}) = V(X_1 + X_2 + \cdots + X_n) = n\sigma^2$$

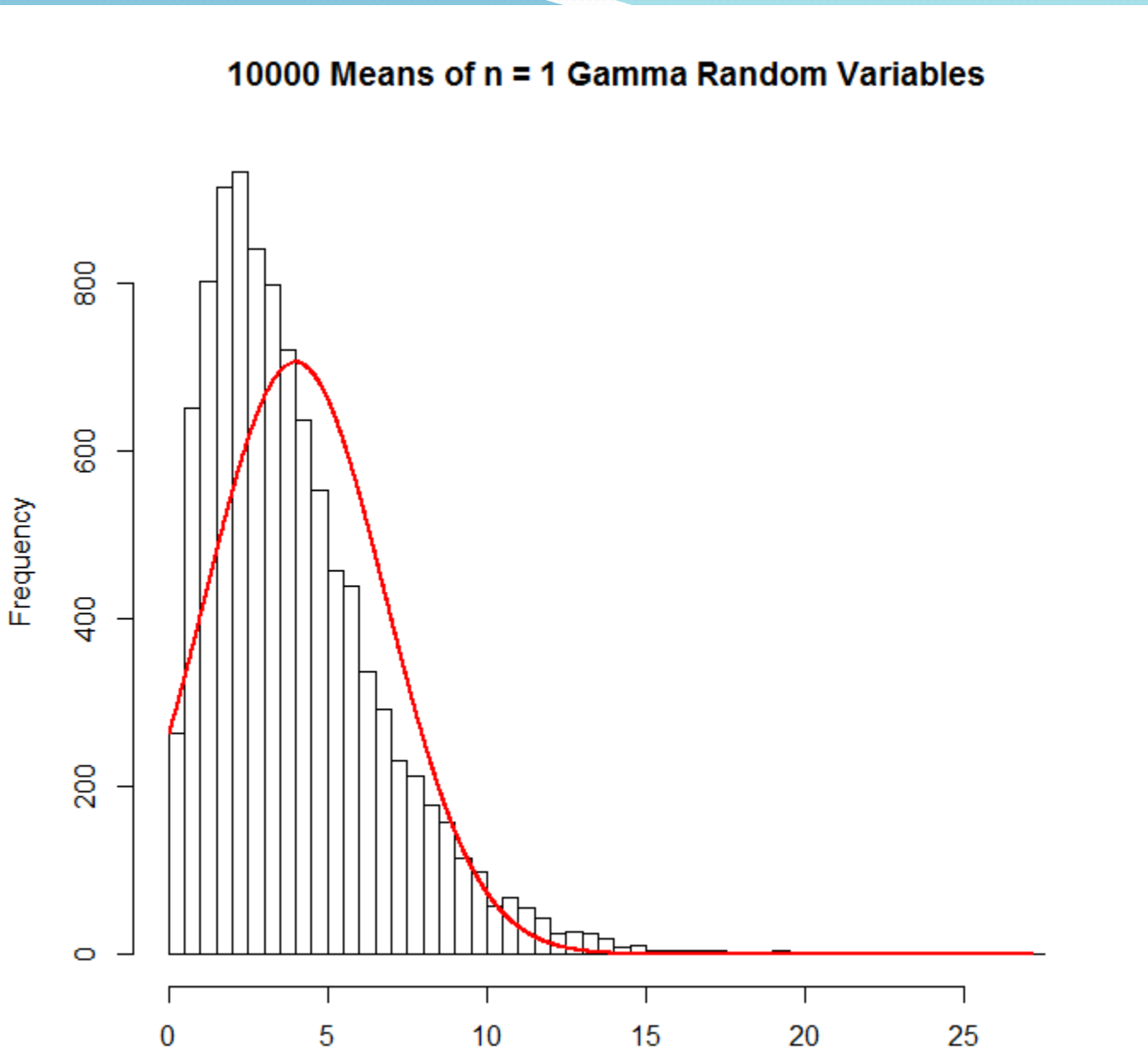
$$E(\bar{X}) = E(n^{-1}X_{\text{sum}}) = n^{-1}n\mu = \mu$$

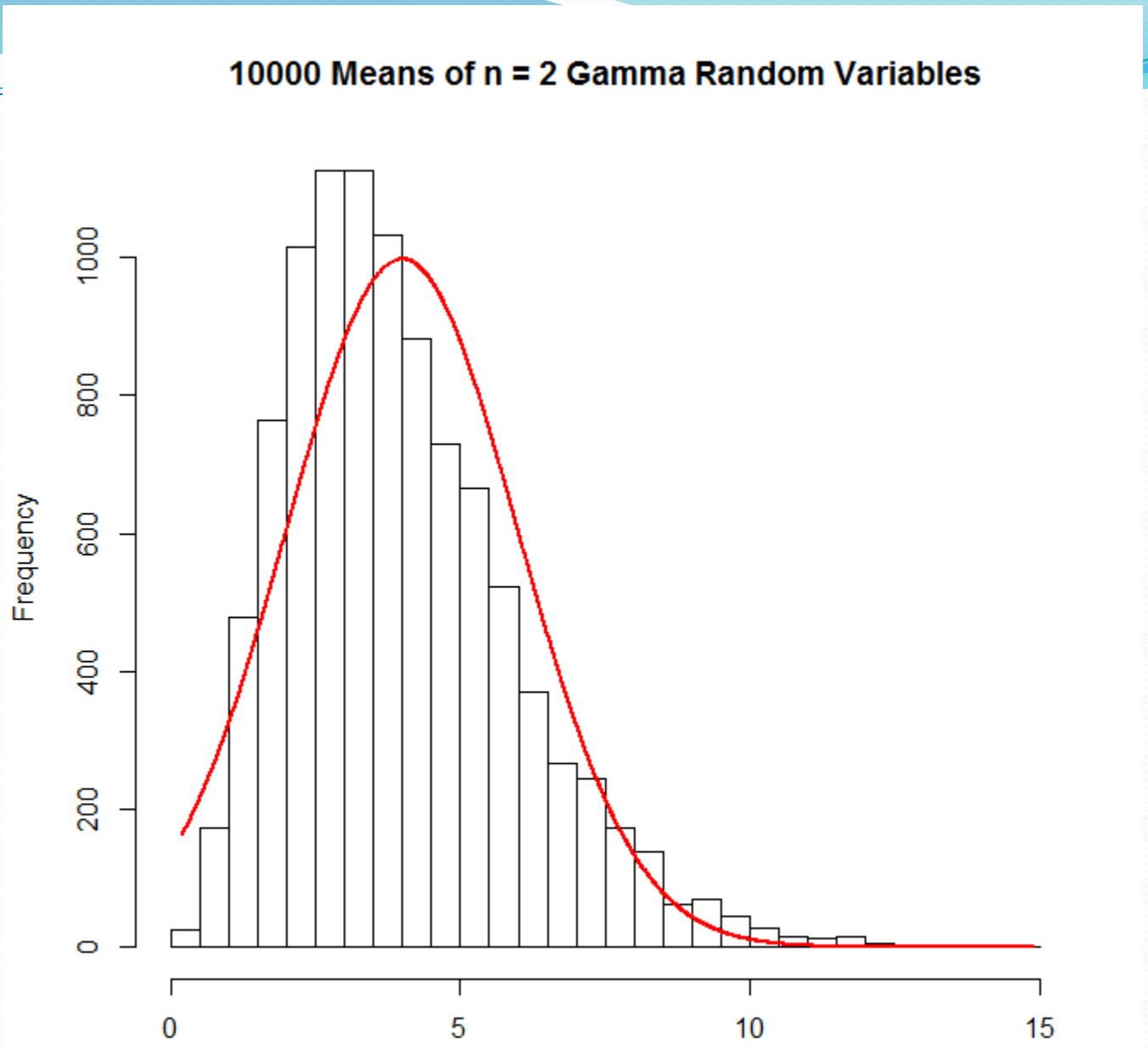
$$V(\bar{X}) = V(n^{-1}X_{\text{sum}}) = n^{-2}n\sigma^2 = \sigma^2 / n$$

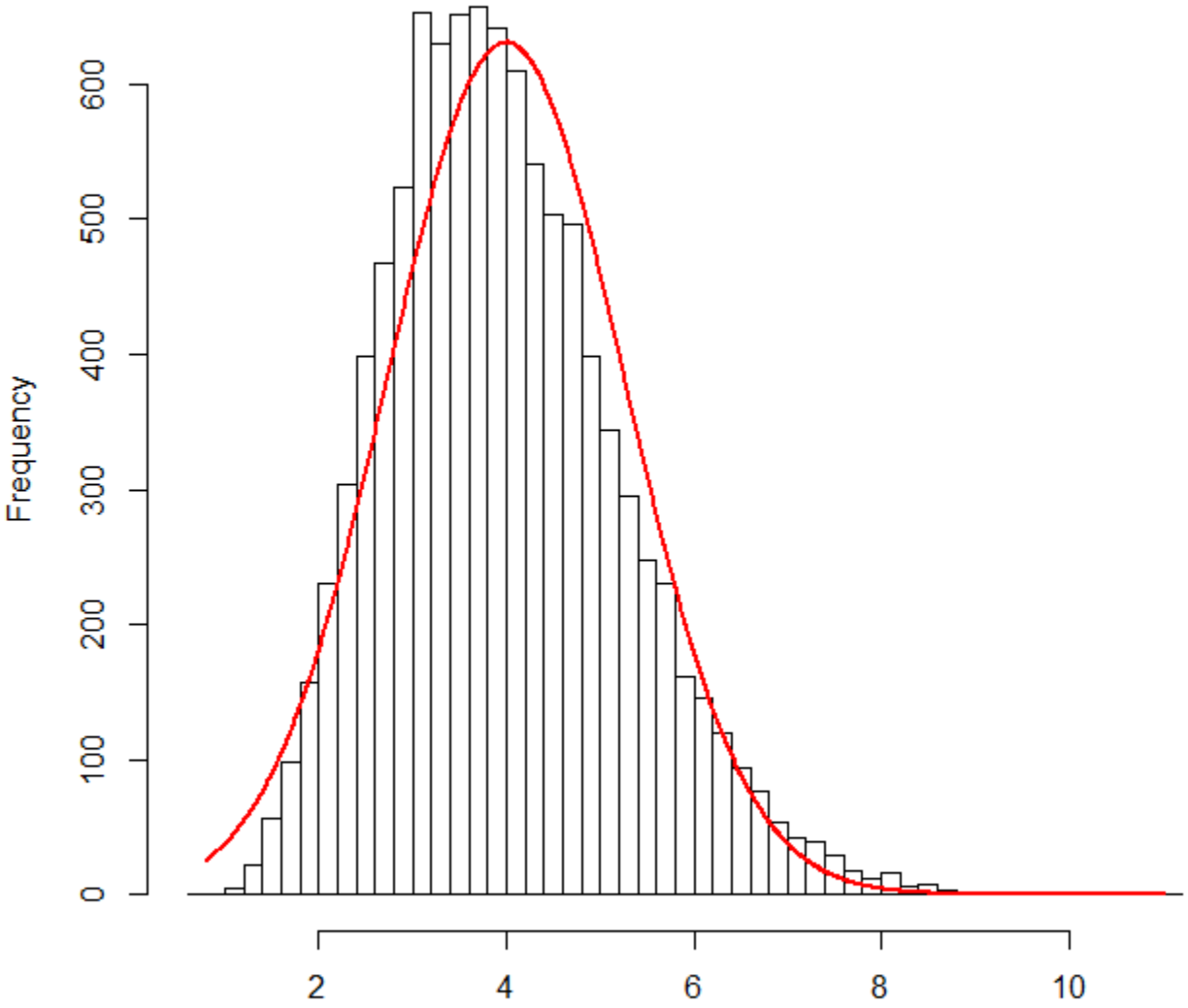
Why is the Normal Distribution Important?

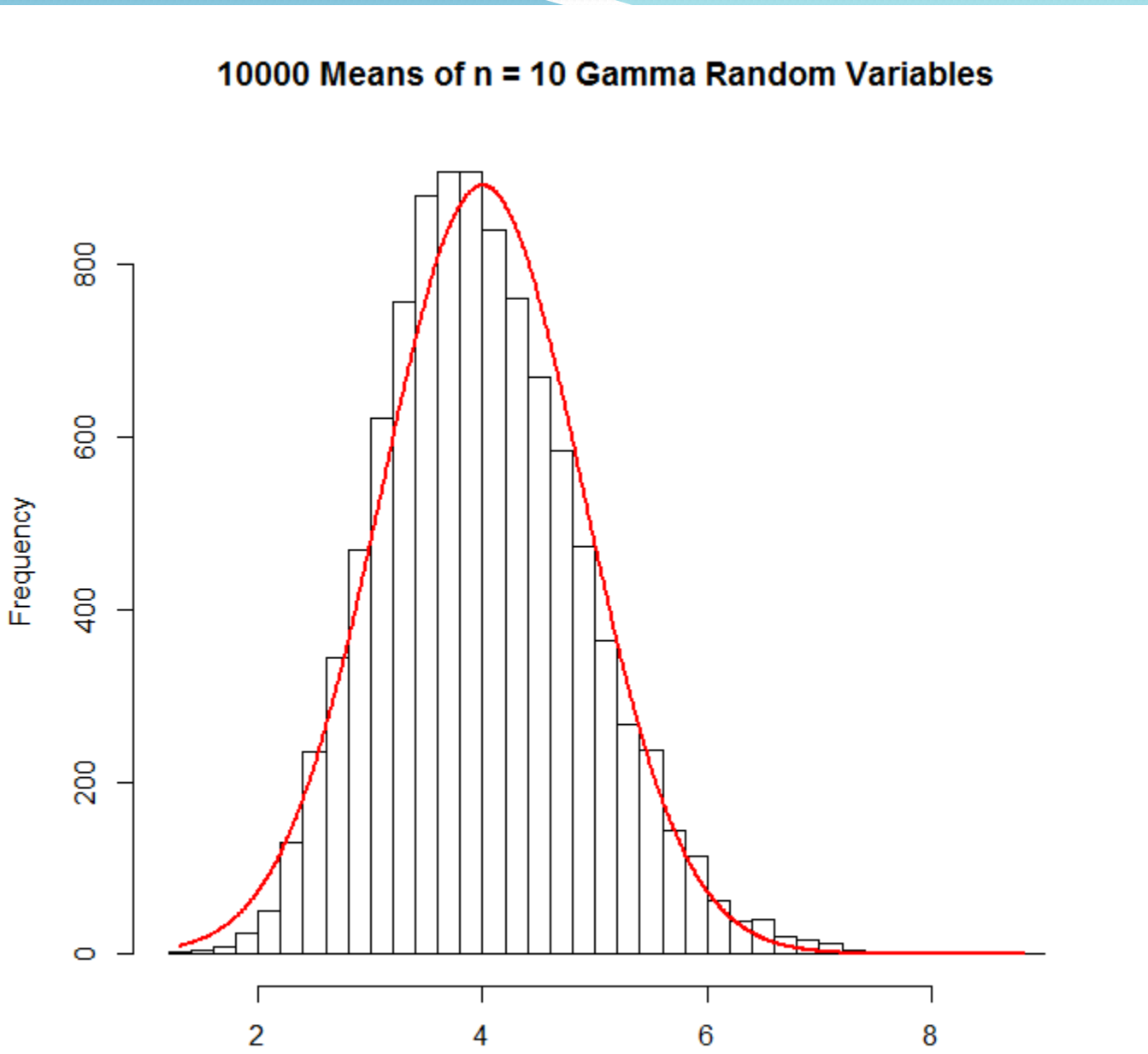
- Most of the time, we do not have data that are normally distributed, nor is it mostly important, with a few exceptions.
- If we have data that are more or less symmetric, and don't have large outliers, then we are generally ok.
- The main reason the normal distribution is important is not that data are usually normally distributed, but that means and other statistics calculated from data are approximately normally distributed. This is called the central limit theorem.
- **The average of a large number of data points is approximately normally distributed.**

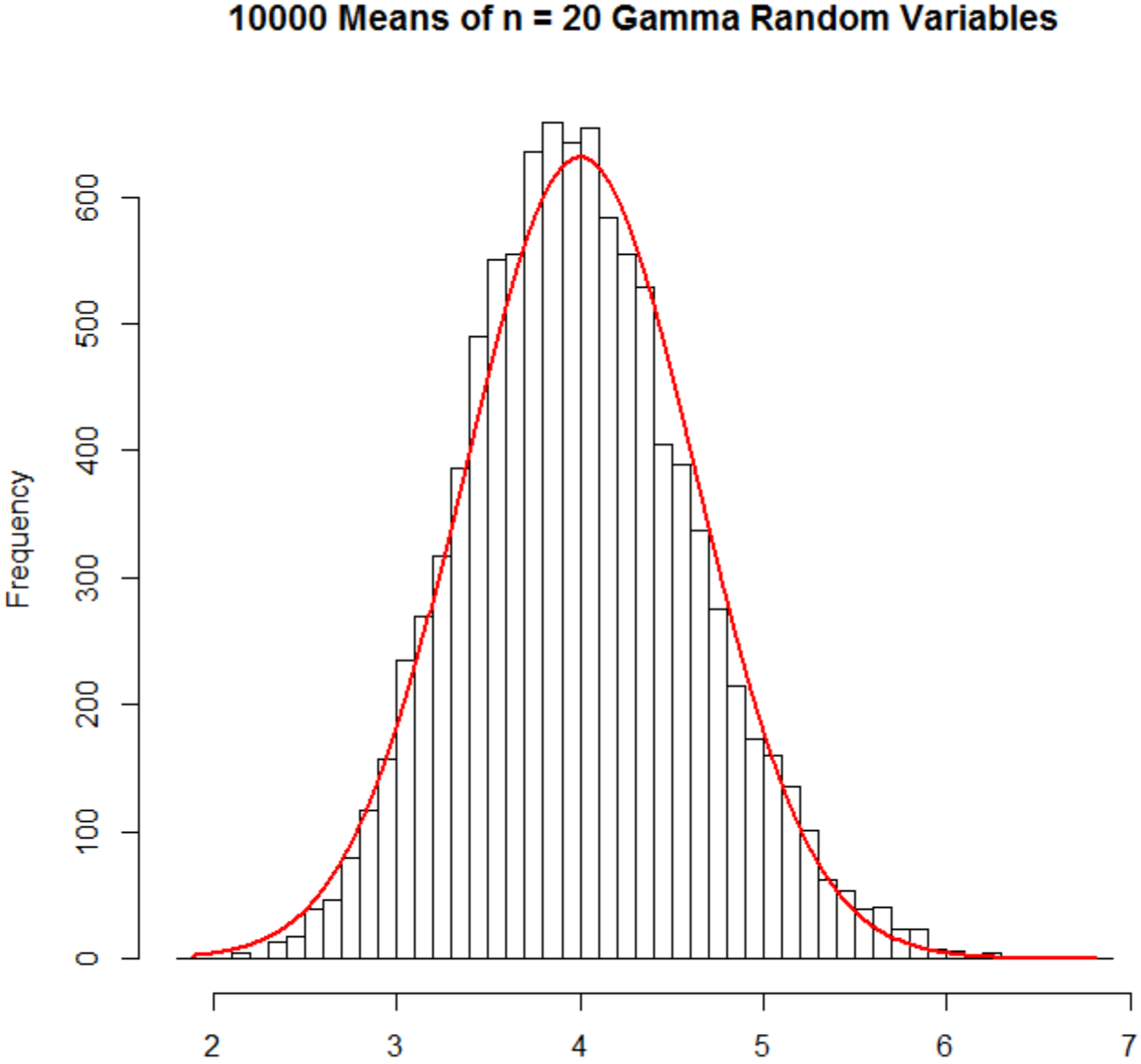

Gamma Distribution with Shape 2 and Scale 2
0.00
0.05
0.10
0.15
0
5
10
15
20
x

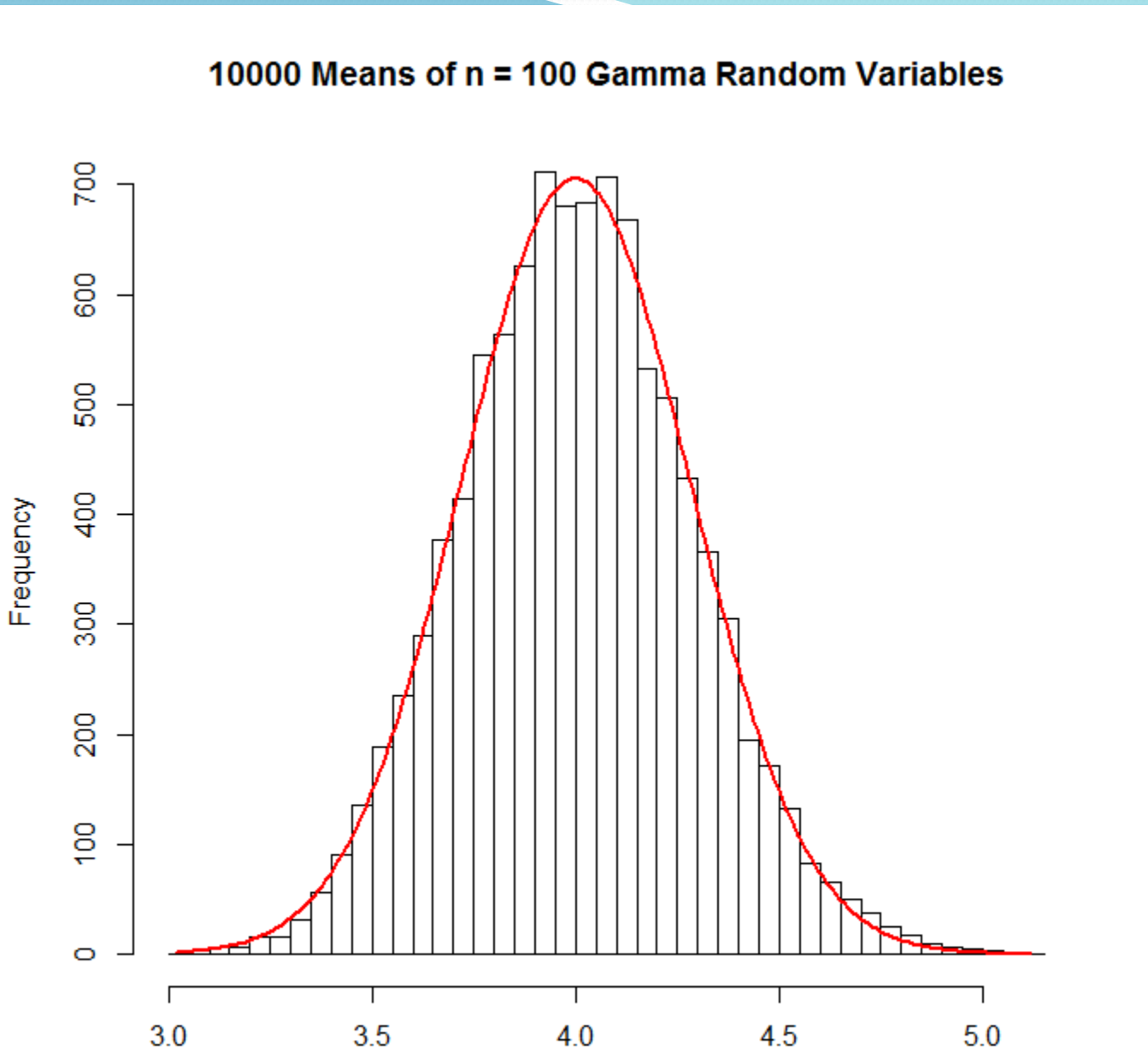

10000 Means of n = 1 Gamma Random Variables
Frequency
0
200
400
600
800
0
5
10
15
20
25


10000 Means of n = 2 Gamma Random Variables
Frequency
0
200
400
600
800
1000
0
5
10
15


10000 Means of n = 5 Gamma Random Variables
Frequency
0
100
200
300
400
500
600
2
4
6
8
10

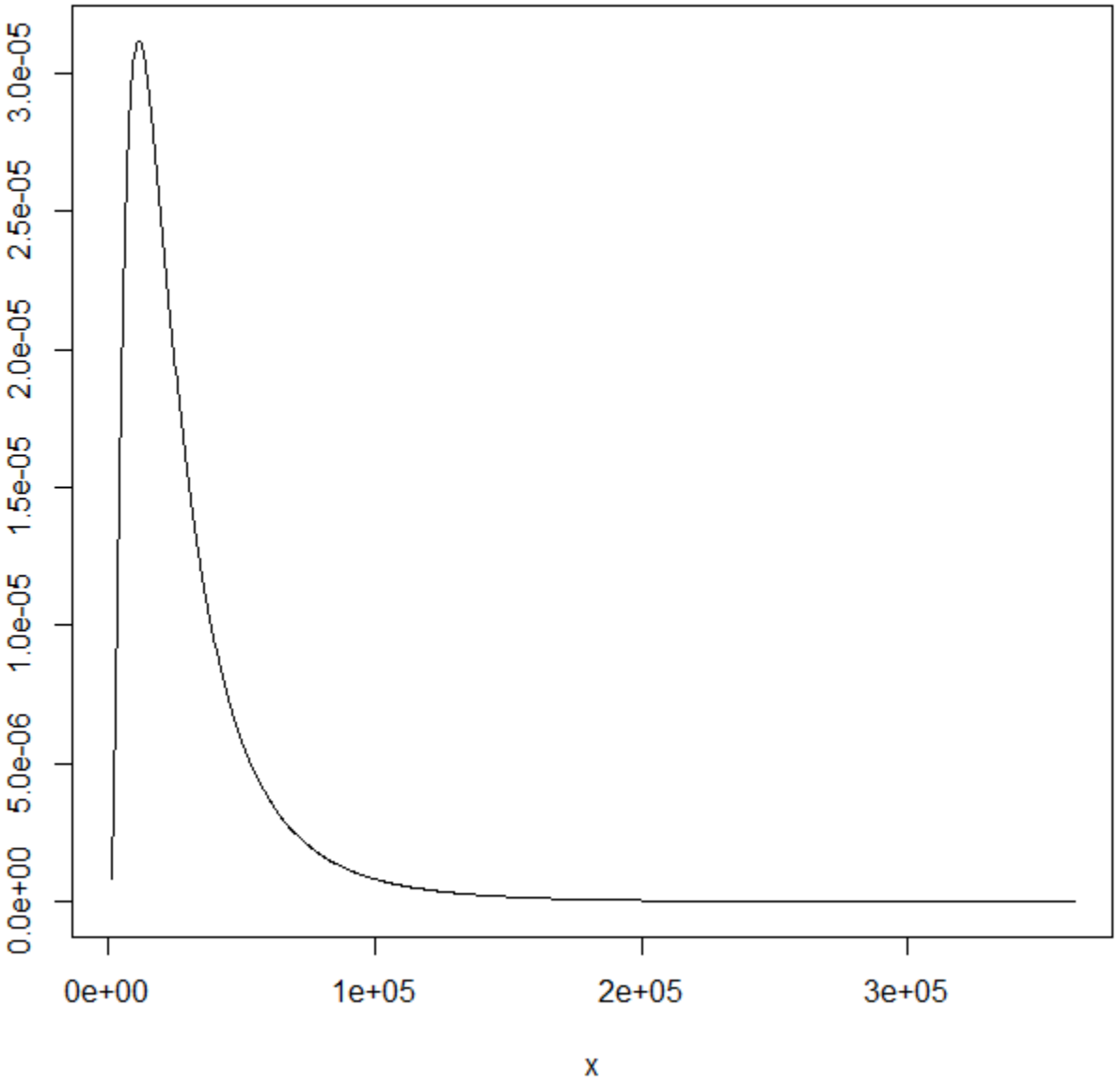

10000 Means of n = 10 Gamma Random Variables
Frequency
0
200
400
600
800
2
4
6
8

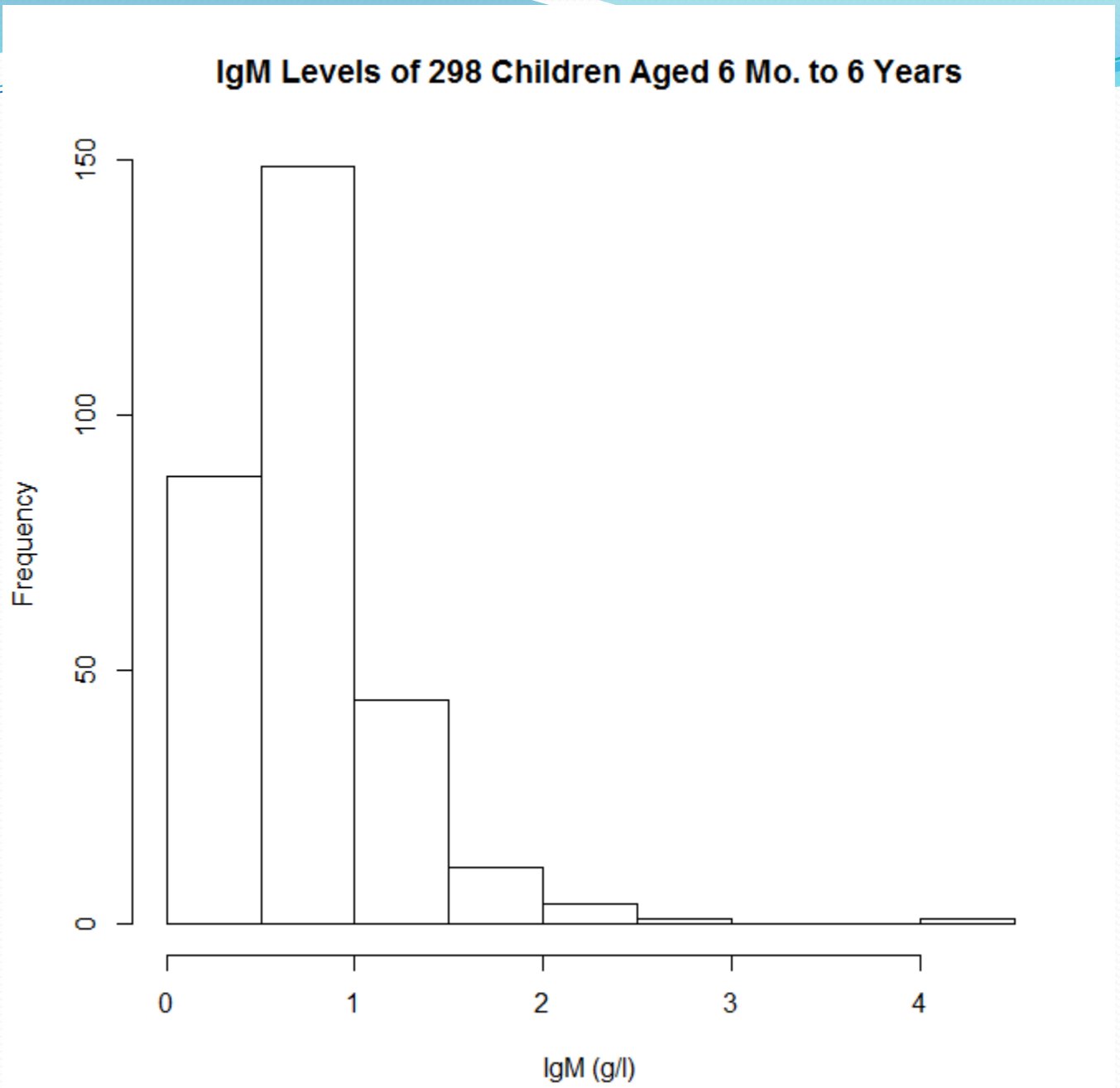

10000 Means of n = 20 Gamma Random Variables
Frequency
0
100
200
300
400
500
600
2
3
4
5
6
7

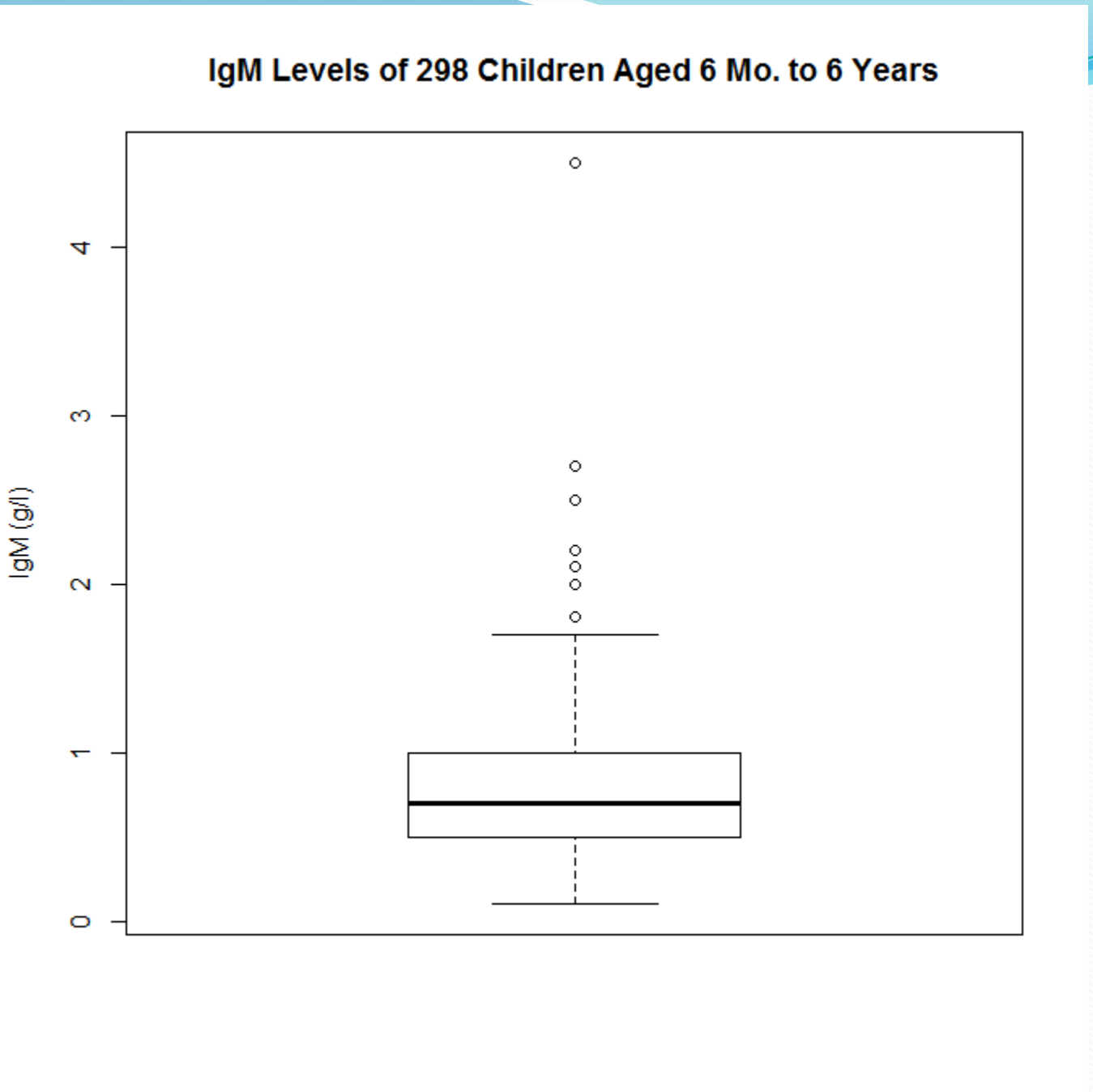

10000 Means of n = 100 Gamma Random Variables
Frequency
0
100
200
300
400
500
600
700
3.0
3.5
4.0
4.5
5.0

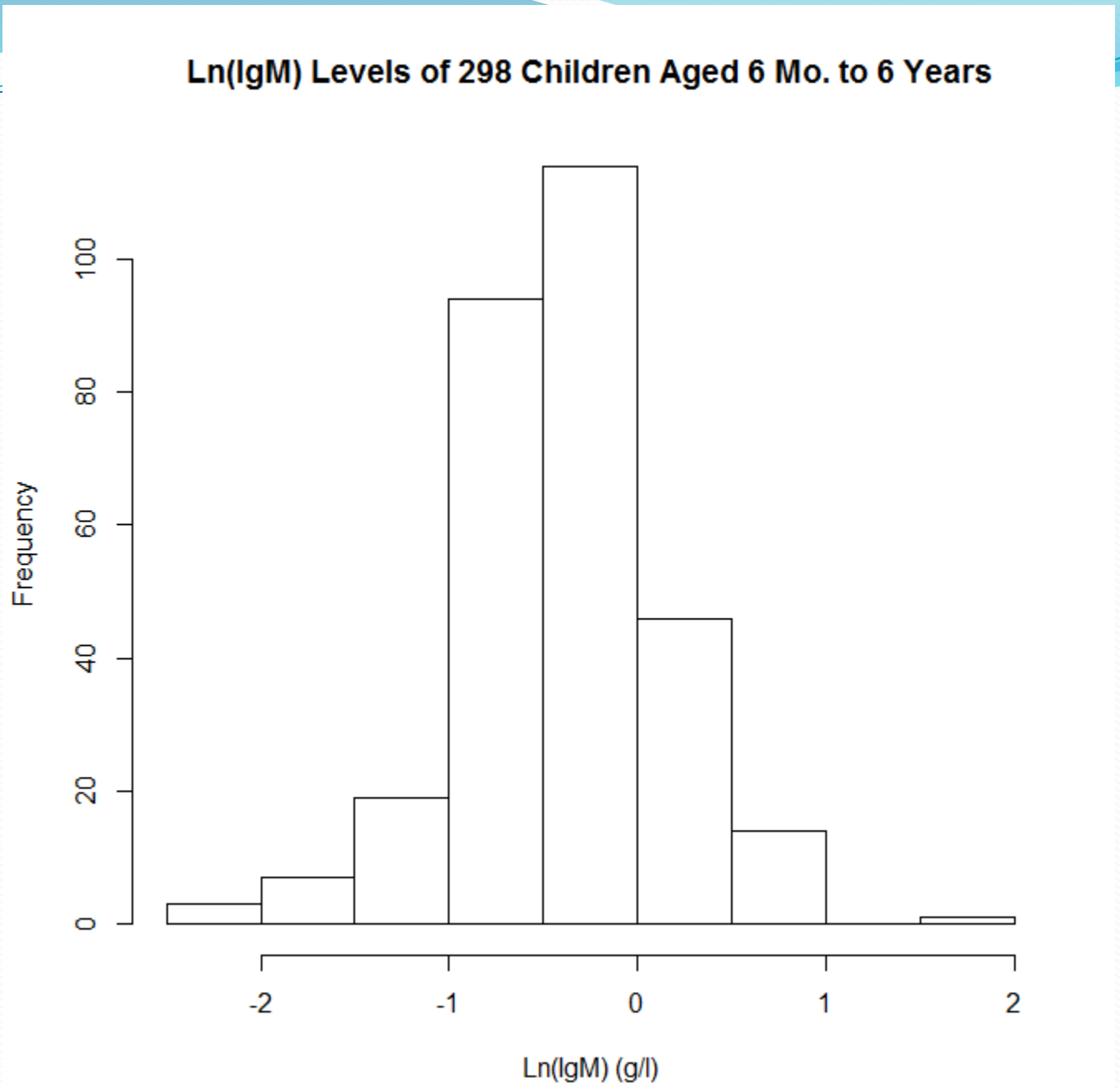
The Lognormal Distribution

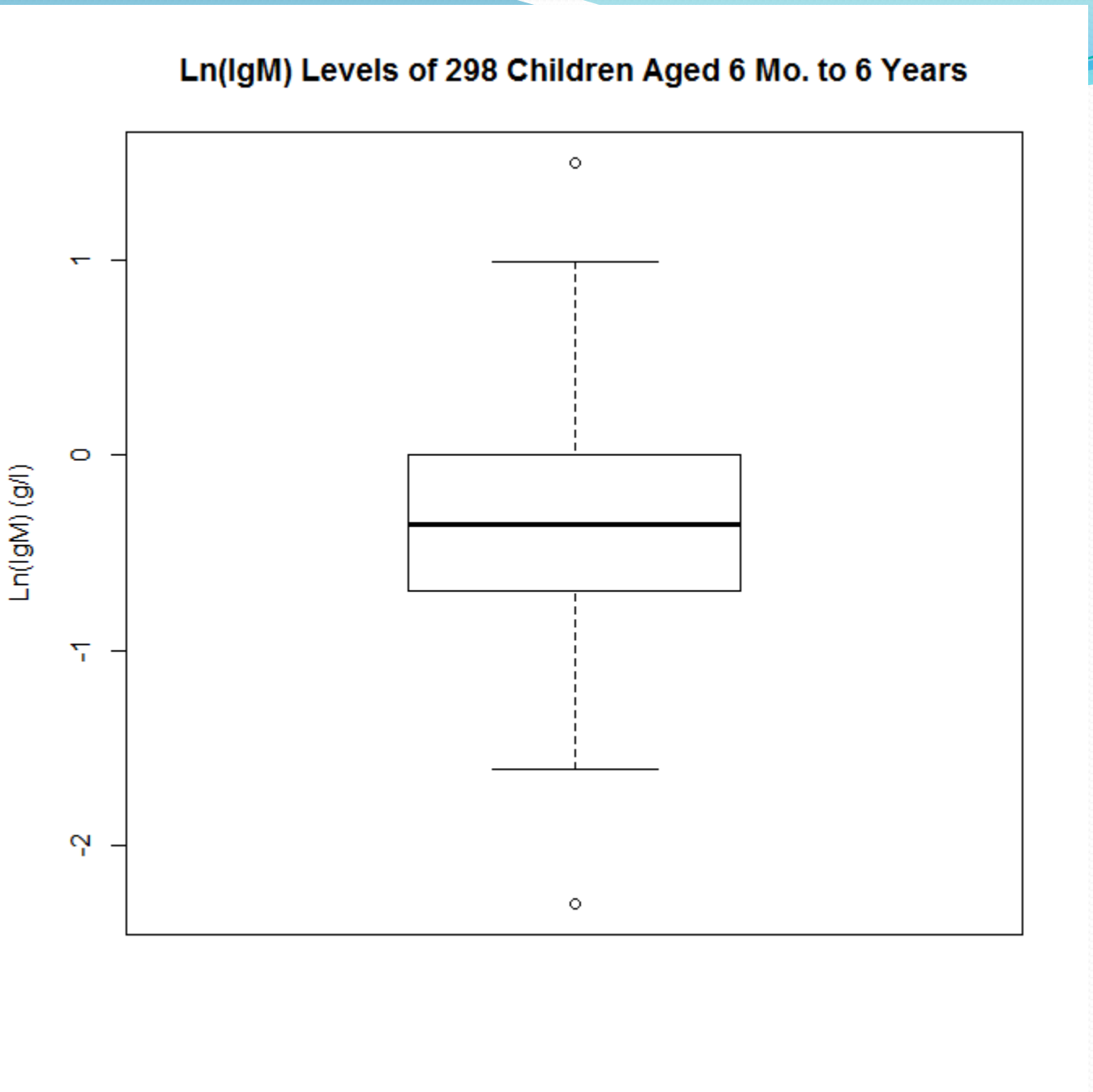
- If X is normally distributed, then e^X is lognormally distributed.
- Mostly, we don't use this distribution directly.
- Instead, if we have data that are right skewed, we often take the log of the data before we analyze it, take the mean, etc.


Lognormal Distribution from N(10 0.64)
0.0e+00
5.0e-06
1.0e-05
1.5e-05
2.0e-05
2.5e-05
3.0e-05
0e+00
1e+05
2e+05
3e+05
x


IgM Levels of 298 Children Aged 6 Mo. to 6 Years
Frequency
0
50
100
150
0
1
2
3
4
IgM (g/l)


IgM Levels of 298 Children Aged 6 Mo. to 6 Years
IgM (g/l)
0
1
2
3
4


Ln(IgM) Levels of 298 Children Aged 6 Mo. to 6 Years
Frequency
0
20
40
60
80
100
-2
-1
0
1
2
Ln(IgM) (g/l)

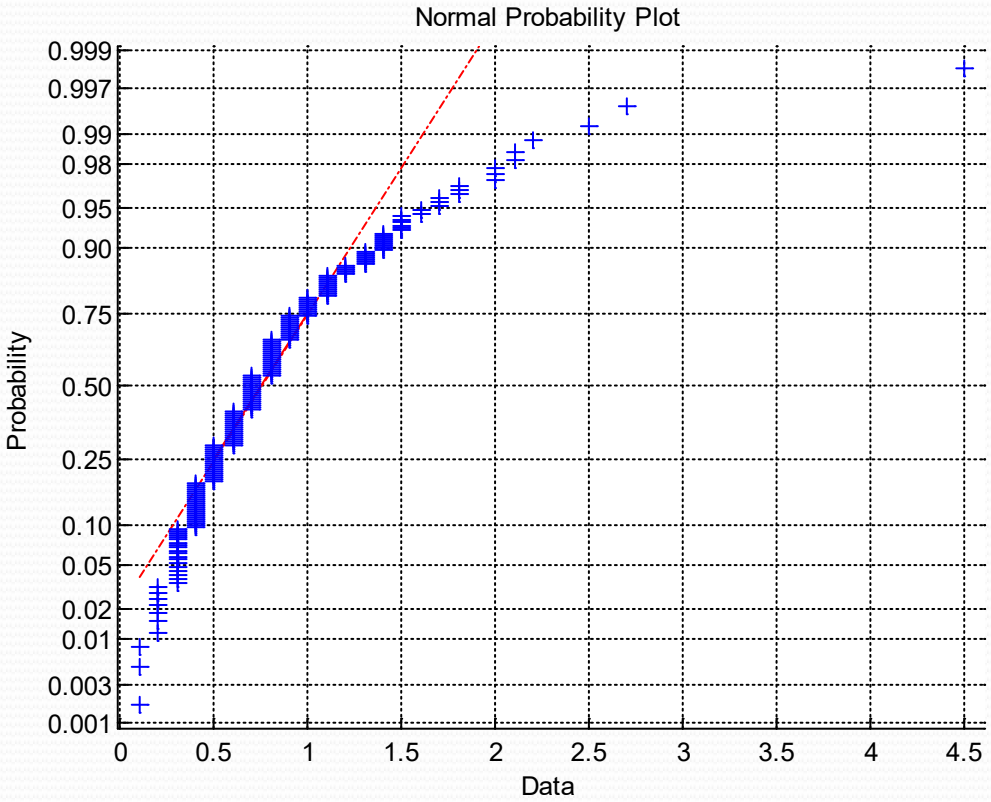

Ln(IgM) Levels of 298 Children Aged 6 Mo. to 6 Years
Ln(IgM) (g/l)
1
0
-1
-2

Probability Plots

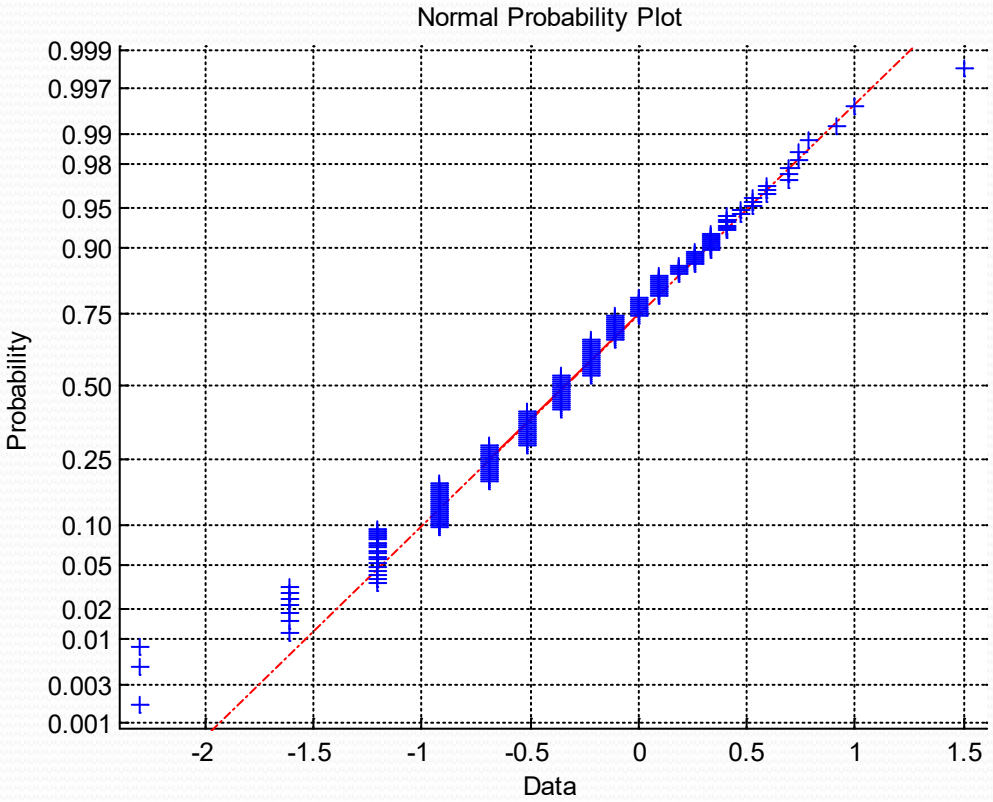
- If we have a sample of size n and want to compare the distribution with a possible theoretical distribution from which it may have come, we have several options
  - Histogram (+ overlaid fitted distribution)
  - Boxplot
  - Probability Plot
- A probability plot puts each point with x-axis value equal to the sample value, and with y-axis value equal to an estimate of what that should have been based on the theoretical distribution.

- Suppose we have a sample of 5 observations.
- The smallest of these can be compared to a typical value for the smallest observation out of 5 from a normal distribution, and similarly for the others.
- We can't use $\Phi^{-1}(1/5)$ for the smallest, because we would then have to use $\Phi^{-1}(5/5) = \Phi^{-1}(1) = \infty$ for the largest.
- The book suggests that we use percentage points $(i - 0.5)/n = .1, .3, .5, .7, .9$, and this is reasonable, though there are other possible choices.
- It does not matter whether we use a standard normal, or one with the same mean and standard deviation as the data, since the percentage points move linearly with the mean and variance, and all that matters is whether the data lie on a straight line.
- For a normal probability plot, use `normplot()` in MATLAB

Normal Probability Plot of IgM Data



Normal Probability Plot of Log(IgM) Data



The Exponential Distribution

- The exponential distribution is often used to model time until an event.
- The domain is $[0, \infty)$ and the PDF and CDF are
  - $f(x) = \lambda e^{-\lambda x}$
  - $F(x) = 1 - e^{-\lambda x}$
- The mean and the variance are
  - $\mu = 1/\lambda$
  - $\sigma^2 = 1/\lambda^2$
  - $\sigma = 1/\lambda$

Hazard Rate

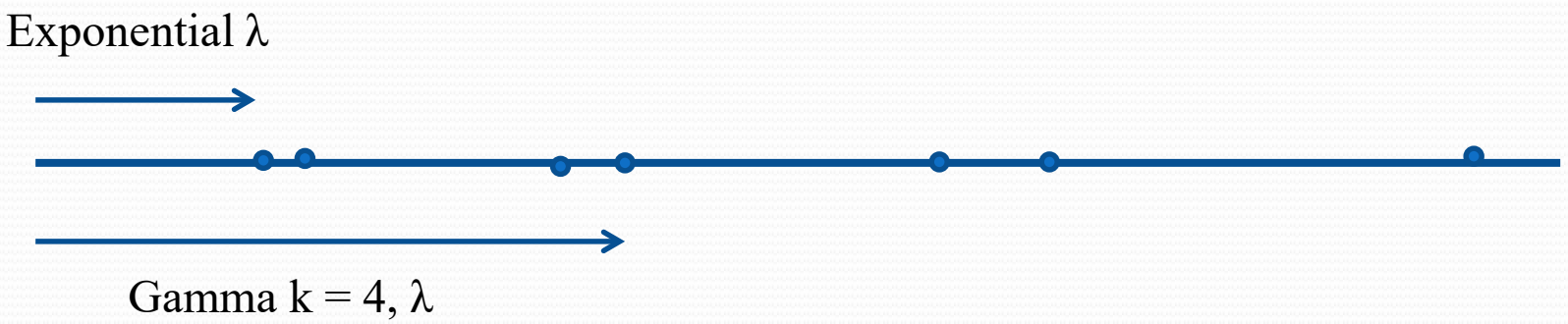
- The hazard rate for a distribution used to model time to an event is the chance that an event will occur at time *t* if an event has not yet happened.
- The chance that the event has not yet happened is $1 - F(t) = e^{-\lambda x}$
- The hazard rate is
  - $f(t)/(1 - F(t)) = \lambda e^{-\lambda x}/e^{-\lambda x} = \lambda$
  - So the exponential has constant hazard

Lack of Memory

- If the waiting time until an event has an exponential distribution with parameter λ and if the event has not happened yet at time t, then the further waiting time after *t* until the event occurs is exponential with parameter λ.
- Suppose the waiting time in minutes for a bus is exponential with parameter $\lambda = 0.1$.
- The mean waiting time is 10 minutes.
- If it has been 10 minutes and the bus has not yet arrived, then the mean further waiting time is 10 minutes.

Connection with the Poisson

- A Poisson process with rate parameter λ produces points on a line so that
  - On any finite interval [a, b], the number of points within the interval is a Poisson random variable with parameter $\lambda(b-a)$.
  - The two random variables associated with disjoint intervals [a, b] and [c, d] are statistically independent.
- If events follow a Poisson process on the line with parameter λ, the interval between any point on the line and the next event has an exponential distribution with parameter λ.
- The waiting time until the k^{th} event follows a gamma distribution with parameters k and λ.



The gamma distribution with parameters k and λ has

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$$

where

$$\Gamma(k) = (k-1)!$$

When $k = 1$

$$f(x) = \lambda e^{-\lambda x}$$

which is the exponential distribution

Other Distributions

- The uniform is used when the random variable falls into a fixed interval with equal likelihood everywhere.
- The beta distribution is an uneven distribution over a fixed interval.
- The Weibull is often used for reliability when the hazard is not constant. Many items have a “bathtub” hazard rate, which needs an even more complex model.

Distributions in MATLAB

```
normcdf(x, mu, sigma)
norminv(p, mu, sigma)
normcdf(x)          #standard normal
norminv(p)
expcdf(x, mu)
expinv(p, mu)
gamcdf(x, a, b)          #a = k, b = 1/λ
unifcdf(x, a, b)
betacdf(x, a, b)
wblcdf(x, a, b)
# and many others
```

Estimators and Estimates

- An *estimator* is a method of computing a number from data, The result is called an *estimate*.
- If we are trying to estimate a parameter, it is best if the estimate is close to the parameter value.
- Since the estimate is a random variable, “close” must be defined as an average distance from the true value, in some sense of average.
- This will depend on the estimator, the population, and the sample size, among other things.

Figures of Merit

Suppose we have a population with parameter θ

We have a random variable $X = \hat{\theta}$ that is meant to estimate θ

We would like to choose X from among available alternatives

so that X is on the average as close as possible to θ

1. We want X not to be too far to one side or the other

   $E(X-\theta) = \mu_X - \theta \approx 0$ Bias

2. We want the distance between X and θ not to be too large

   $E(X-\theta)^2$ is small

$$E(X-\theta)^2 = E(X-\mu_X+\mu_X-\theta)^2 = E(X-\mu_X)^2 - 2E(X-\mu_X)(\mu_X-\theta) + E(\mu_X-\theta)^2$$
$$= V(X) + E(\mu_X-\theta)^2$$

The mean square error of an estimate is the variance plus the square of the bias.

The Sample Mean

Suppose that $X_1, X_2, ..., X_n$ are a random sample from a population and that

$$E(X_i) = \mu_X$$

$$V(X_i) = \sigma_X^2$$

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$$

We previously saw that

$$E(\bar{X}) = \mu_X$$

$$V(\bar{X}) = \sigma^2 / n$$

So the bias of $\bar{X}$ is

$$E(\bar{X} - \mu_X) = \mu_X - \mu_X = 0$$

and the variance (and MSE) of $\bar{X}$ is

$$V(\bar{X}) = \sigma^2 / n$$

The Sample Proportion

Suppose that $X \sim \text{Bin}(n, p)$ and consider $\hat{p} = X / n$

We previously saw that

$$E(\hat{p}) = np / n = p$$

$$V(\hat{p}) = np(1-p) / n^2 = p(1-p) / n$$

So the bias of $\hat{p}$ is

$$E(\hat{p} - p) = p - p = 0$$

and the variance (and MSE) of $\hat{p}$ is

$$V(\hat{p}) = p(1-p) / n$$

Unbiased vs. Minimum MSE

- Unbiased sounds like a good thing, but it may not be the most important characteristic.
- A small amount of bias relative to the variability of the estimator may not be important.
- The sample variance s^2 (with denominator $n - 1$) is an unbiased estimate of the population variance σ^2.
- But the sample standard deviation s is not unbiased for the population standard deviation σ.
- We use it anyway, because it is optimal in other ways.

Consistency/Convergence

Another important property of a point estimator is that

as the sample size gets larger, the estimate gets closer to the true value.

This is called *consistency*.

For our purposes, this simply means that the MSE $\rightarrow 0$ as $n \rightarrow \infty$

For the sample mean

$\text{MSE}(\bar{X}) = \sigma^2 / n \rightarrow 0$

For the sample proportion

$\text{MSE}(\hat{p}) = p(1-p) / n \rightarrow 0$