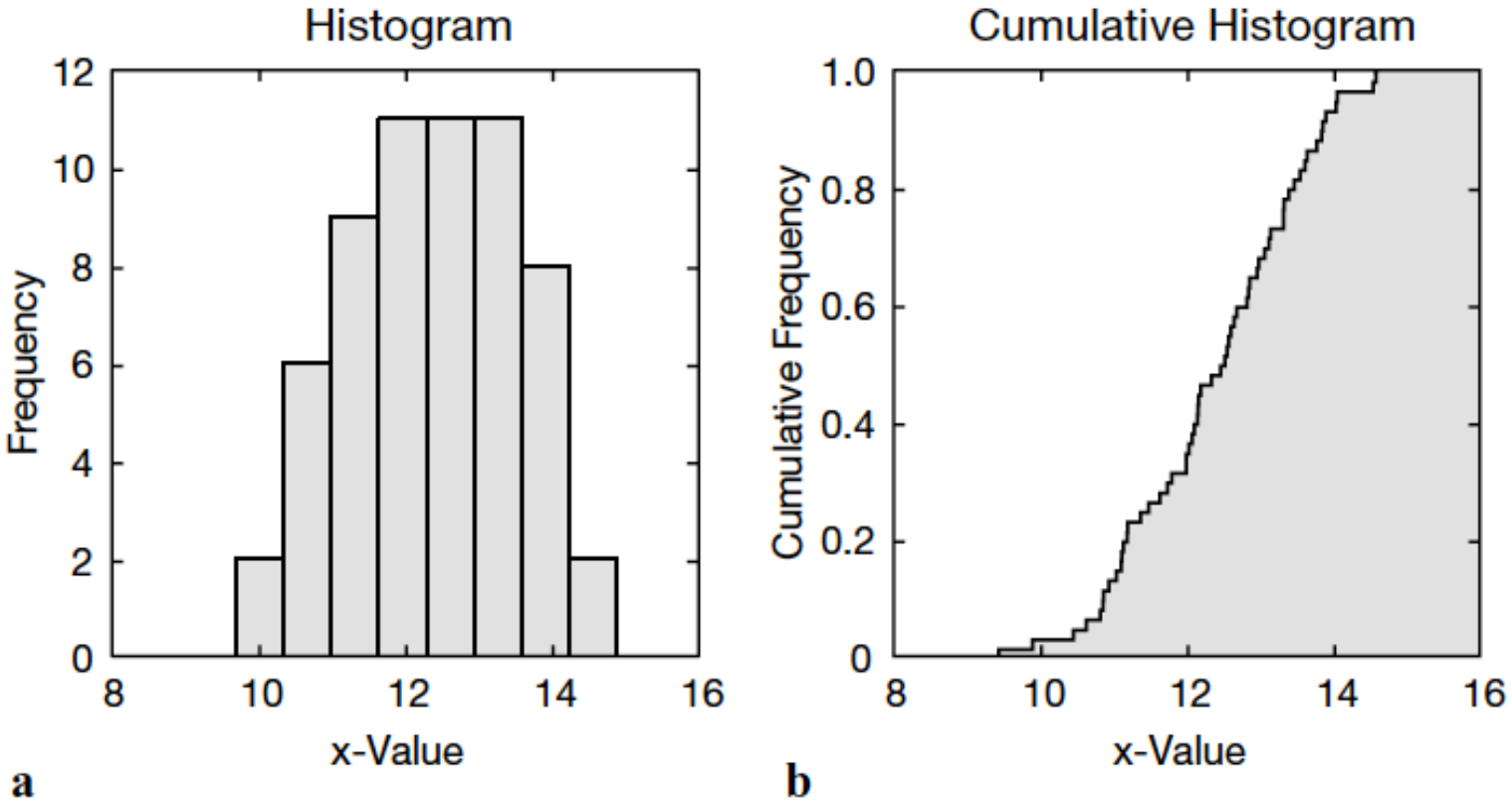
$$x = (x_1, x_2, \ldots, x_N)$$

Measures of Central Tendency

Parameters of central tendency or location represent the most important measures for characterizing an empirical distribution (Fig. 3.2). These values help locate the data on a linear scale. They represent a typical or best value that describes the data. The most popular indicator of central tendency is the *arithmetic mean*, which is the sum of all data points divided by the number of observations:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The arithmetic mean can also be called the mean or the average of an univariate data set. The sample mean is often used as an estimate of the population mean μ for the underlying theoretical distribution. The arithmetic mean is sensitive to outliers, i.e., extreme values that may be very different from the majority of the data.



Graphical representation of an empirical frequency distribution. **a** In a *histogram*, the frequencies are organized in classes and plotted as a bar plot. **b** The *cumulative histogram* of a frequency distribution displays the counts of all classes lower and equal than a certain value. The cumulative histogram is normalized to a total number of observations of one.

from the majority of the data. Therefore, the *median* is often used as an alternative measure of central tendency. The median is the x-value which is in the middle of the data, i.e., 50% of the observations are larger than the median and 50% are smaller. The median of a data set sorted in ascending order is defined as

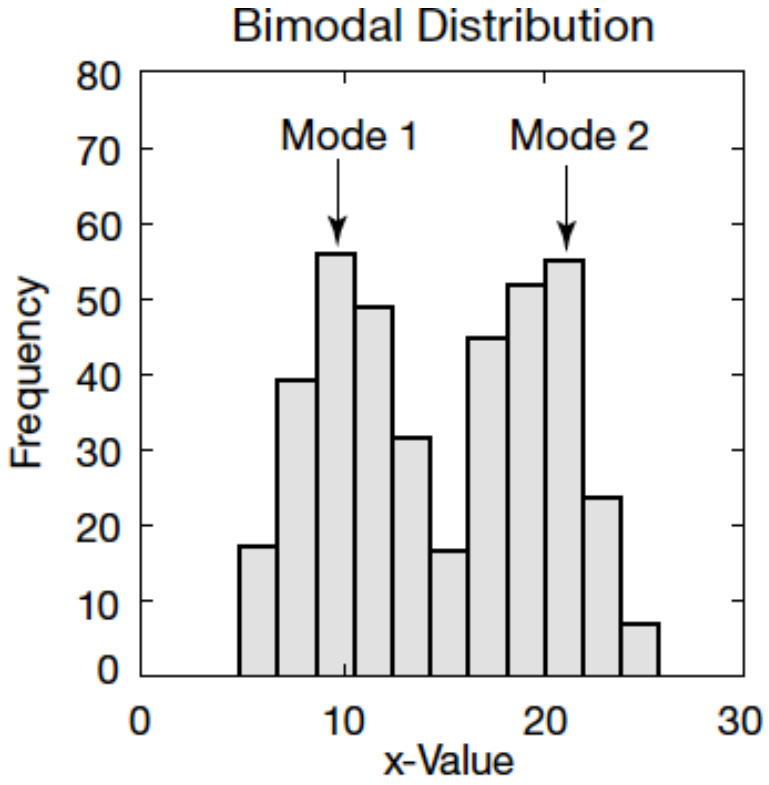
$$\tilde{x} = x_{(N+1)/2}$$

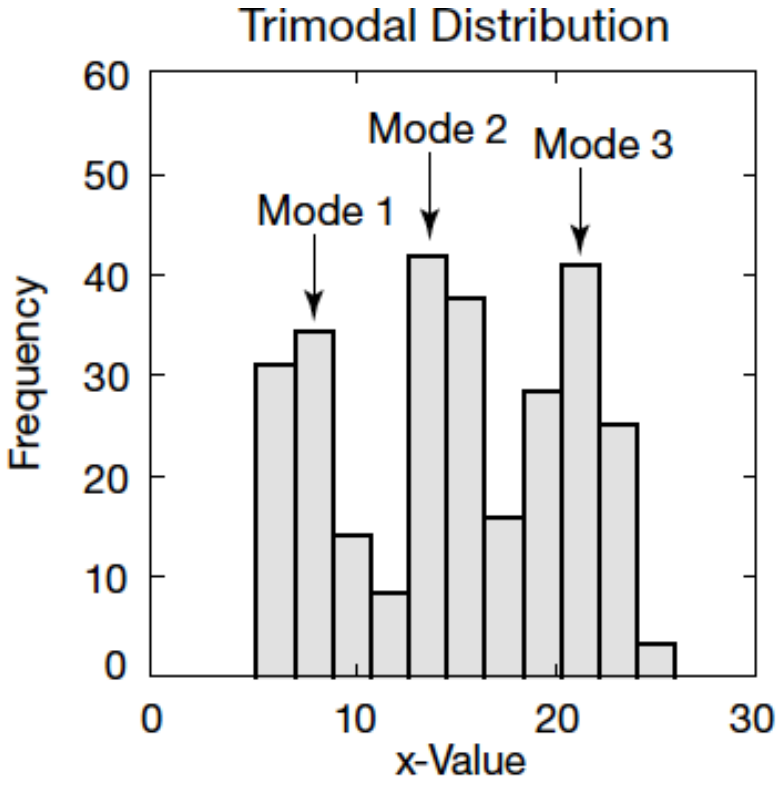
if N is odd and

$$\tilde{x} = \left(x_{(N/2)} + x_{(N/2)+1}\right)/2$$

if N is even. Although outliers also affect the median, their absolute values do not influence it. *Quantiles* are a more general way of dividing the data sample into groups containing equal numbers of observations. For example, *quartiles* divide the data into four groups, *quintiles* divide the observations in five groups and *percentiles* define one hundred groups.

The third important measure for central tendency is the *mode*. The mode is the most frequent x value or – if the data are grouped in classes – the center of the class with the largest number of observations. The data have no mode if there aren't any values that appear more frequently than any of the other values. Frequency distributions with one mode are called *unimodal*, but there may also be two modes (*bimodal*), three modes (*trimodal*) or four or more modes (*multimodal*).





The measures mean, median and mode are used when several quantities add together to produce a total, whereas the *geometric mean* is often used if these quantities are multiplied. Let us assume that the population of an organism increases by 10% in the first year, 25% in the second year, then 60% in the last year. The average increase rate is not the arithmetic mean, since the number of individuals is multiplied by (not added to) 1.10 in the first year, by 1.375 in the second year and 2.20 in the last year. The average growth of the population is calculated by the geometric mean:

$$\overline{x}_G = \left(x_1 \cdot x_2 \cdot \ldots \cdot x_N\right)^{1/N}$$

The average growth of these values is 1.4929 suggesting a ~49% growth of the population. The arithmetic mean would result in an erroneous value of 1.5583 or ~56% growth. The geometric mean is also an useful measure of central tendency for skewed or log-normally distributed data. In other words, the logarithms of the observations follow a gaussian distribution. The geometric mean, however, is not calculated for data sets containing negative values.

negative values. Finally, the *harmonic mean*

$$\bar{x}_H = N / \left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N} \right)$$

is used to take the mean of asymmetric or log-normally distributed data, similar to the geometric mean, but they are both not robust to outliers. The harmonic mean is a better average when the numbers are defined in relation to some unit. The common example is averaging velocity. The harmonic mean is also used to calculate the mean of samples sizes.

Measures of Dispersion

Another important property of a distribution is the dispersion. Some of the parameters that can be used to quantify dispersion are illustrated in Figure 3.3. The simplest way to describe the dispersion of a data set is the *range*, which is the difference between the highest and lowest value in the data set given by

$$\Delta x = x_{\max} - x_{\min}$$

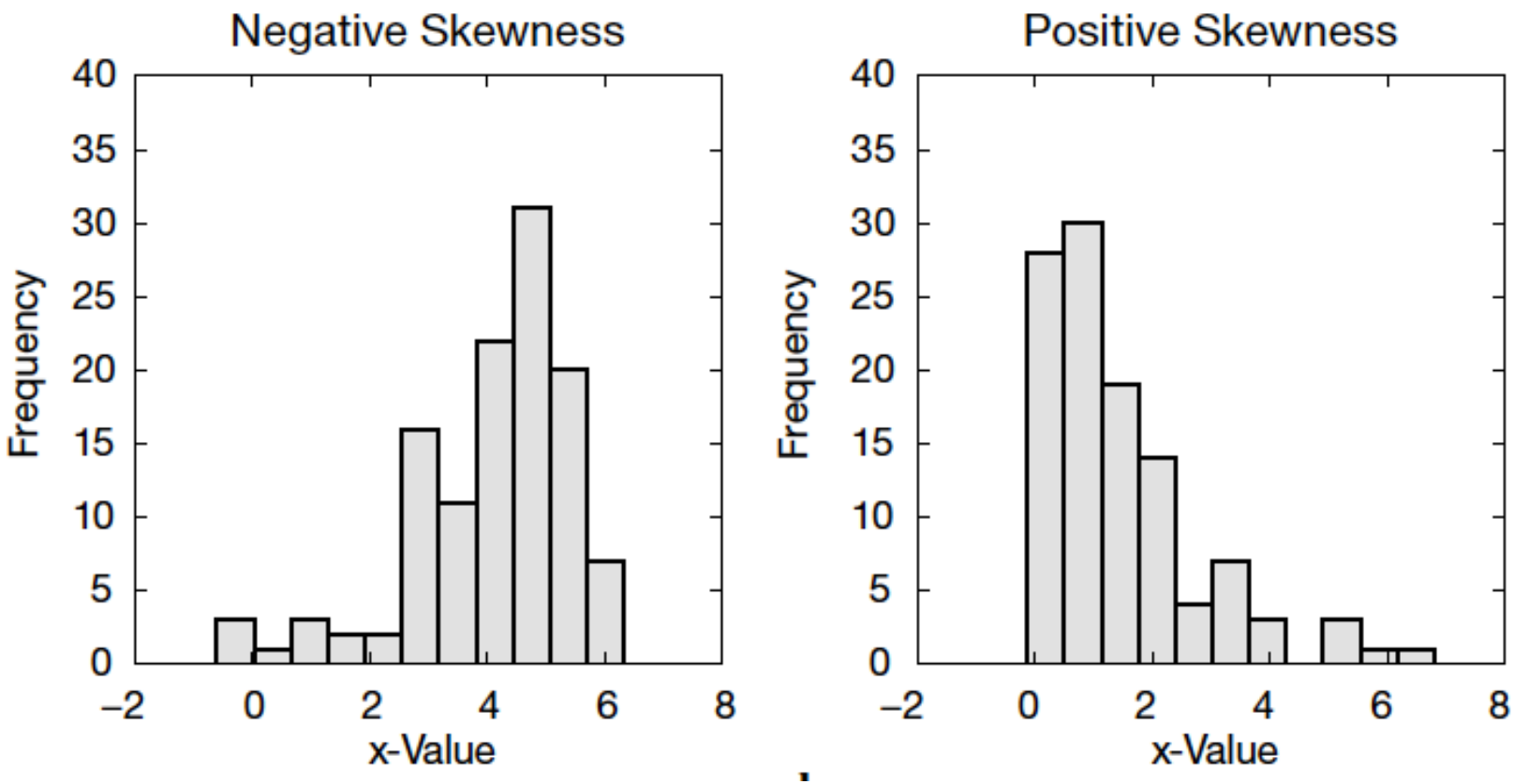
Since the range is defined by the two extreme data points, it is very susceptible to outliers. Hence, it is not a reliable measure of dispersion in most cases. Using the interquartile range of the data, i.e., the middle 50% of the data attempts to overcome this. A most useful measure for dispersion is the *standard deviation*.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The standard deviation is the average deviation of each data point from the mean. The standard deviation of an empirical distribution is often used as an estimate for the population standard deviation σ. The formula of the population standard deviation uses N instead of $N-1$ in the denominator. The sample standard deviation s is computed with $N-1$ instead of N since it uses the sample mean instead of the unknown population mean. The sample mean, however, is computed from the data x_i, which reduces the degrees of freedom by one. The *degrees of freedom* are the number of values in a distribution that are free to be varied. Dividing the average deviation of the data from the mean by N would therefore underestimate the population standard deviation σ.

The *variance* is the third important measure of dispersion. The variance is simply the square of the standard deviation.


Negative Skewness
Frequency
x-Value
0
5
10
15
20
25
30
35
40
–2
0
2
4
6
8
Positive Skewness
Frequency
x-Value
0
5
10
15
20
25
30
35
40
–2
0
2
4
6
8

Furthermore, both *skewness* and *kurtosis* can be used to describe the shape of a frequency distribution. Skewness is a measure of asymmetry of the tails of a distribution. The most popular way to compute the asymmetry of a distribution is Pearson's mode skewness:

skewness = (mean-mode) / standard deviation

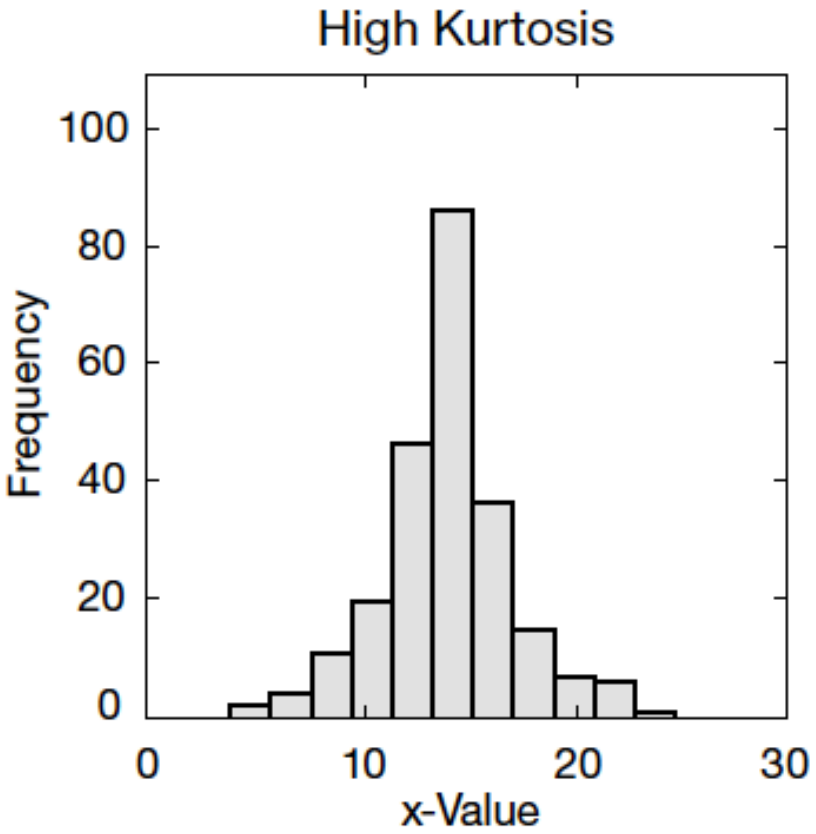
A negative skew indicates that the distribution is spread out more to the left of the mean value, assuming increasing values on the axis to the right. The sample mean is smaller than the mode. Distributions with positive skewness have large tails that extend to the right. The skewness of the symmetric normal distribution is zero. Although Pearson's measure is a useful one, the following formula by Fisher for calculating the skewness is often used instead, including the corresponding MATLAB function.

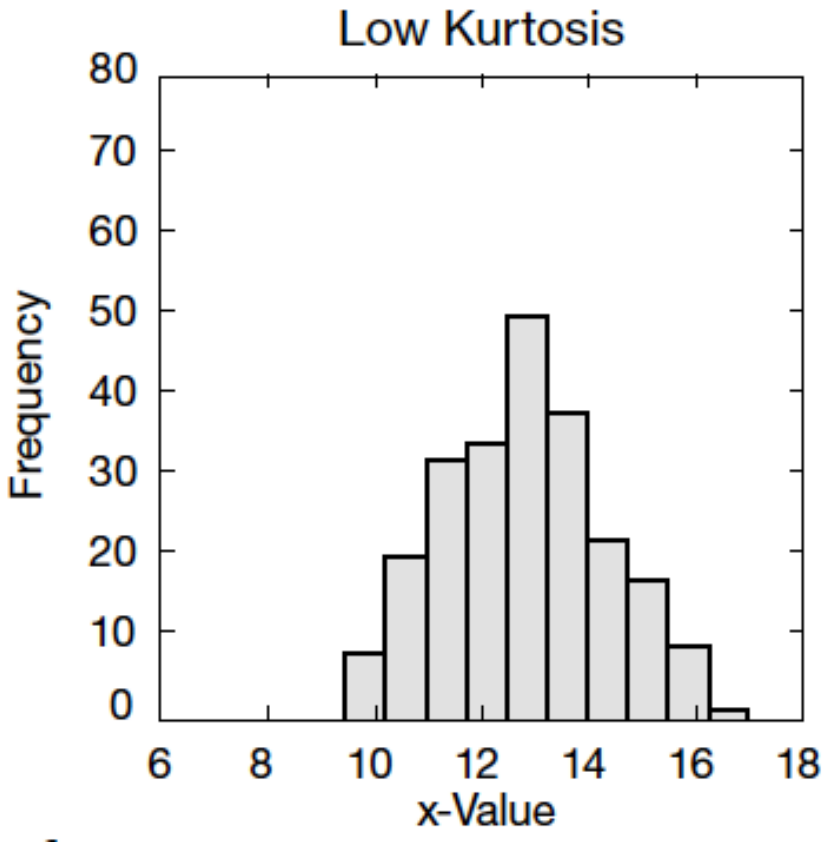
$$skewness = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^3}{s^3}$$

The second important measure for the shape of a distribution is the *kurtosis*. Again, numerous formulas to compute the kurtosis are available.

$$kurtosis = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^4}{s^4}$$

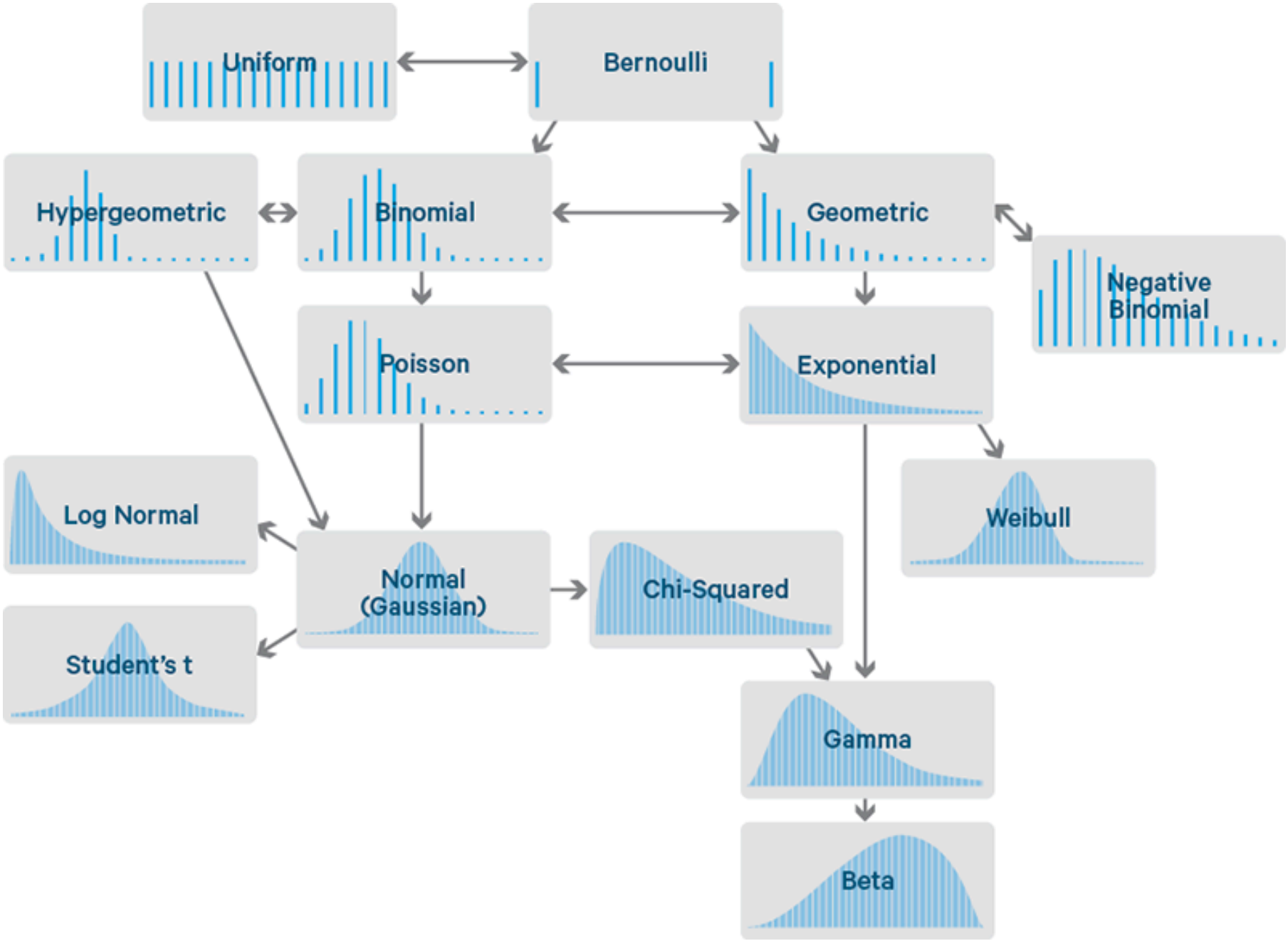
The kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. A high kurtosis indicates that the distribution has a distinct peak near the mean, whereas a distribution characterized by a low kurtosis shows a flat top near the mean and heavy tails. Higher peakedness of a distribution is resulting from rare extreme deviations, whereas a low kurtosis is caused by frequent moderate deviations. A normal distribution has a kurtosis of three. Therefore, some definitions for kurtosis subtract three from the above term in order to set the kurtosis of the normal distribution to zero.


High Kurtosis
Frequency
x-Value
0
20
40
60
80
100
0
10
20
30


Low Kurtosis
Frequency
x-Value
0
10
20
30
40
50
60
70
80
6
8
10
12
14
16
18

Theoretical Distributions

Now we have described the empirical frequency distribution of our sample. A histogram is a convenient way to picture the frequency distribution of the variable x. If we sample the variable sufficiently often and the output ranges are narrow, we obtain a very smooth version of the histogram. An infinite number of measurements $N \to \infty$ and an infinite small class width produce the random variable's *probability density function* (PDF). The probability distribution density $f(x)$ defines the probability that the variate has the value equal to x. The integral of $f(x)$ is normalized to unity, i.e., the total number of observations is one. The *cumulative distribution function* (CDF) is the sum of a discrete PDF or the integral of a continuous PDF. The cumulative distribution function $F(x)$ is the probability that the variable takes a value less than or equal x.


Uniform
Bernoulli
Hypergeometric
Binomial
Geometric
Negative Binomial
Poisson
Exponential
Log Normal
Weibull
Normal (Gaussian)
Chi-Squared
Student's t
Gamma
Beta

Uniform Distribution

A *uniform* or *rectangular distribution* is a distribution that has a constant probability (Fig. 3.4). The corresponding probability density function is

$$f(x) = 1/N = const.$$

where the random variable x has any of N possible values. The cumulative distribution function is

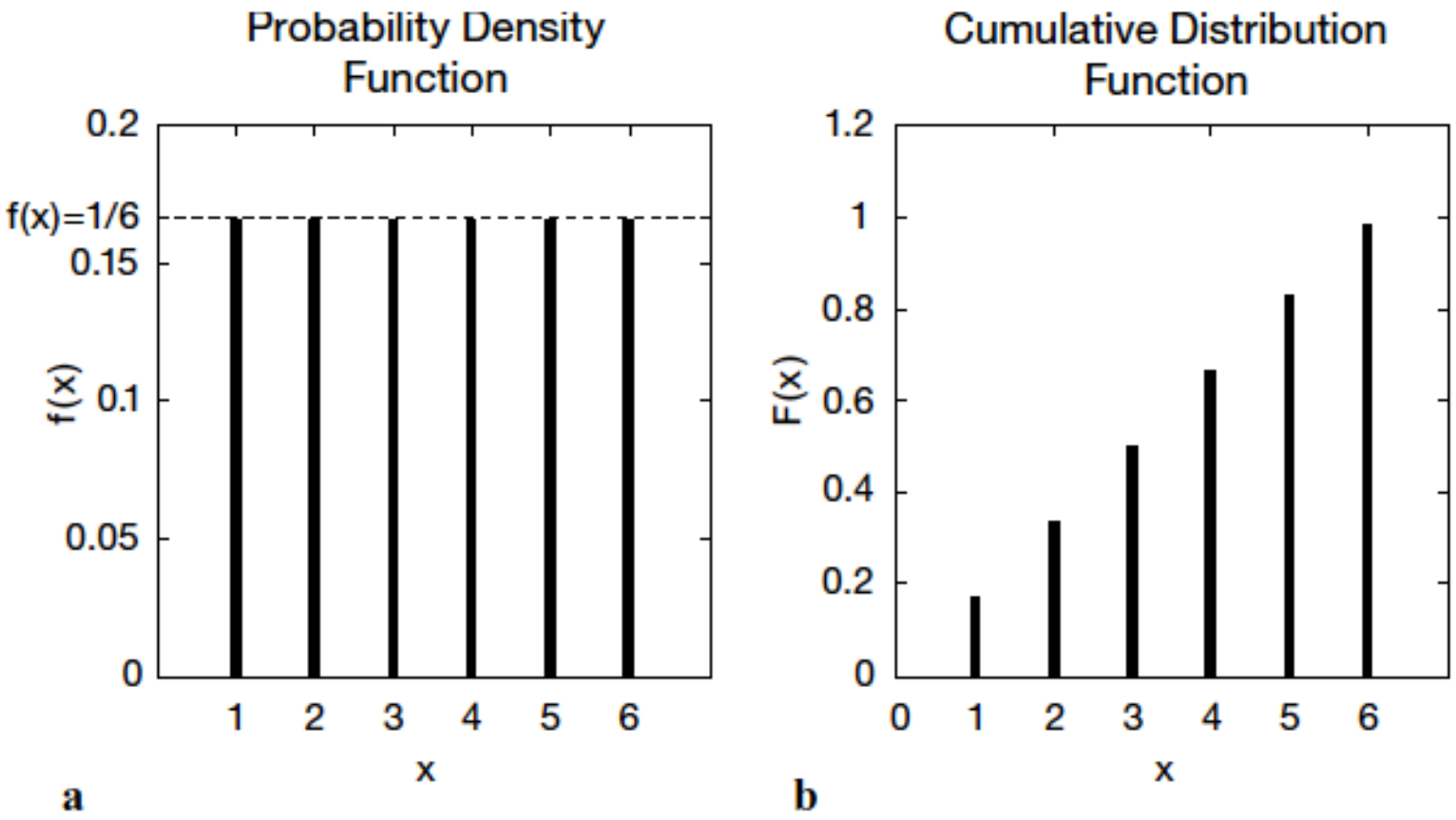
$$F(x) = x \cdot 1/N$$

The probability density function is normalized to unity

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

i.e., the sum of probabilities is one. Therefore, the maximum value of the cumulative distribution function is one.

$$F(x)_{\max} = 1$$



a Probability density function $f(x)$ and **b** cumulative distribution function $F(x)$ of a uniform distribution with $N=6$. The 6 discrete values of the variable x have the same probability of 1/6.

Normal or Gaussian Distribution

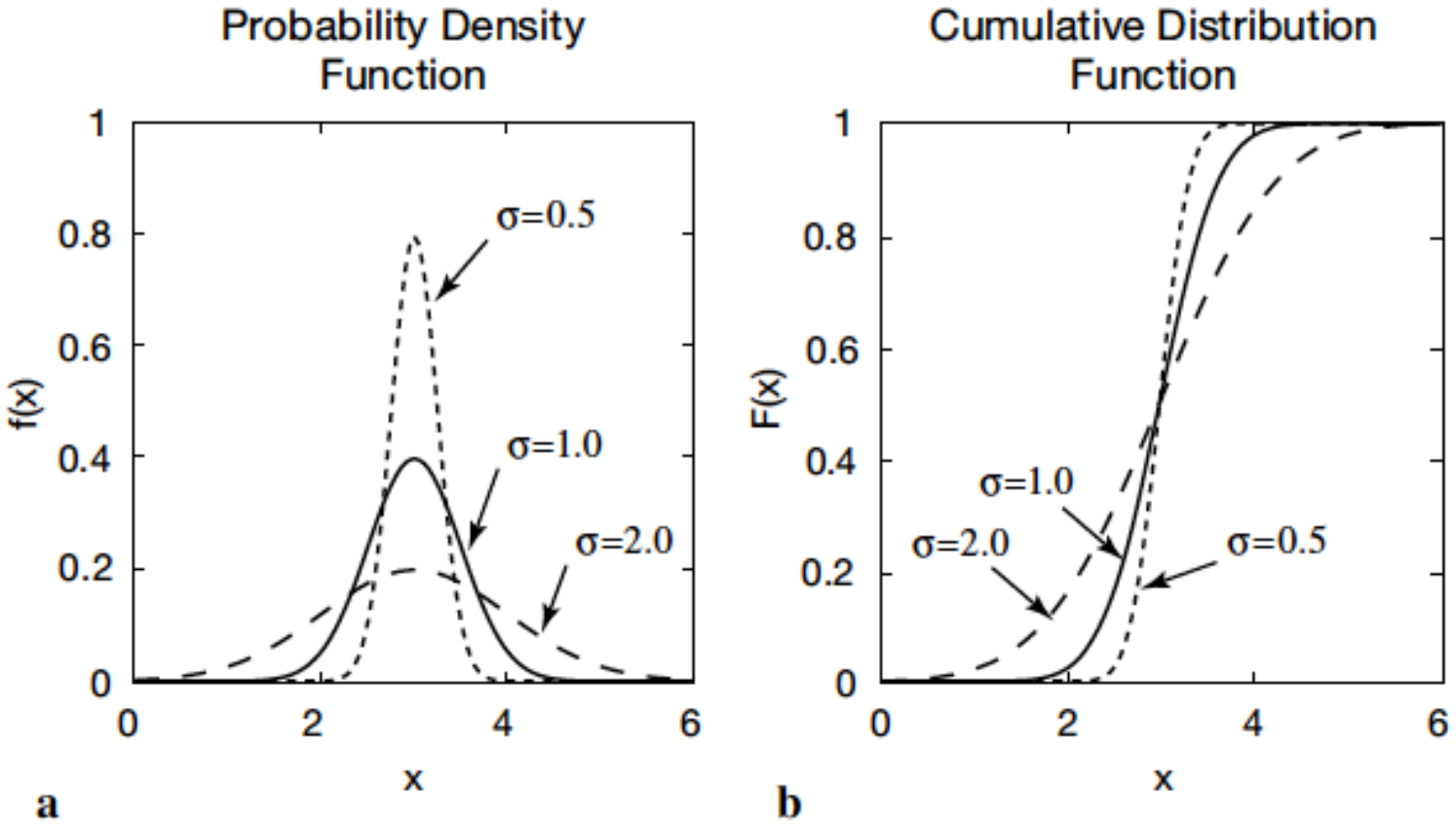
When $p = 0.5$ (symmetric, no skew) and $N \to \infty$, the binomial distribution approaches the *normal* or *gaussian distribution* with the parameters mean μ and standard deviation σ . The probability density function of a normal distribution in the continuous case is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

and the cumulative distribution function is

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) dy$$

The normal distribution is used when the mean is the most frequent and most likely value. The probability of deviations is equal towards both directions and decrease with increasing distance from the mean.

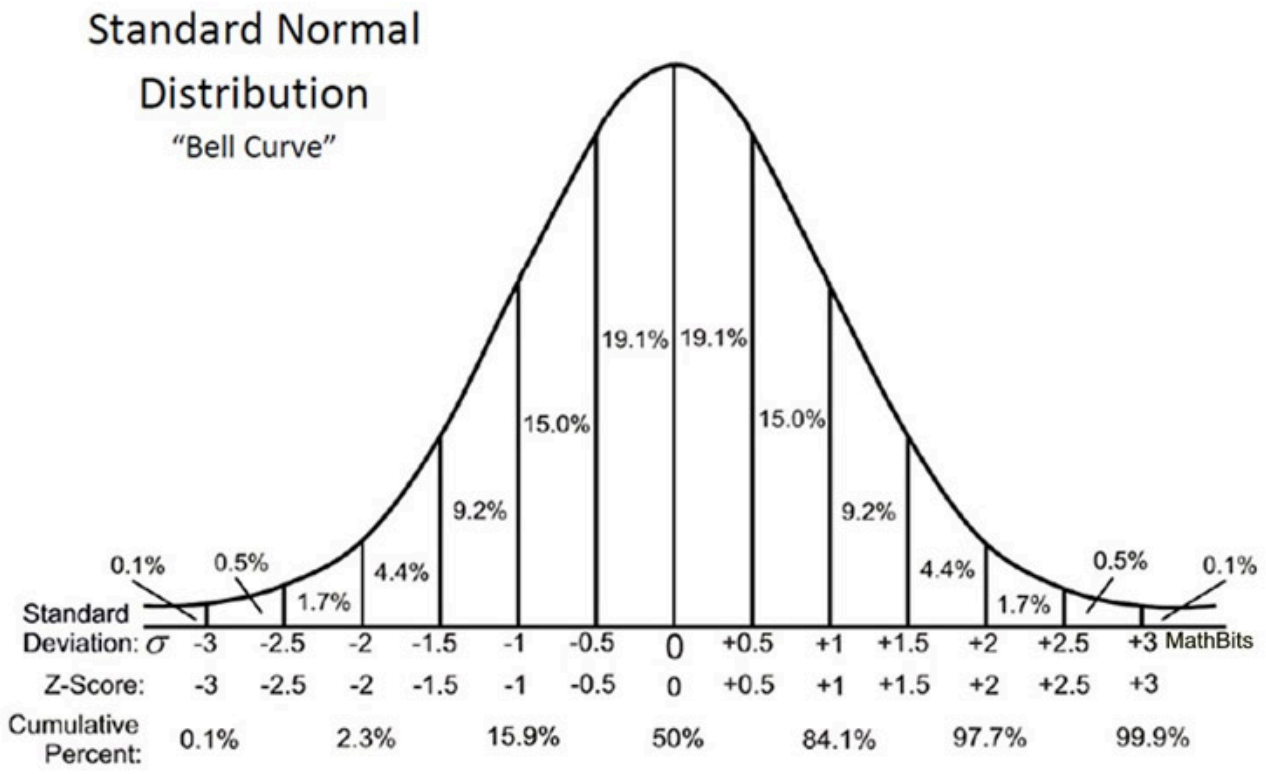


a Probability density function $f(x)$ and **b** cumulative distribution function $F(x)$ of a gaussian or normal distribution with mean $\mu=3$ and different values for standard deviation σ.

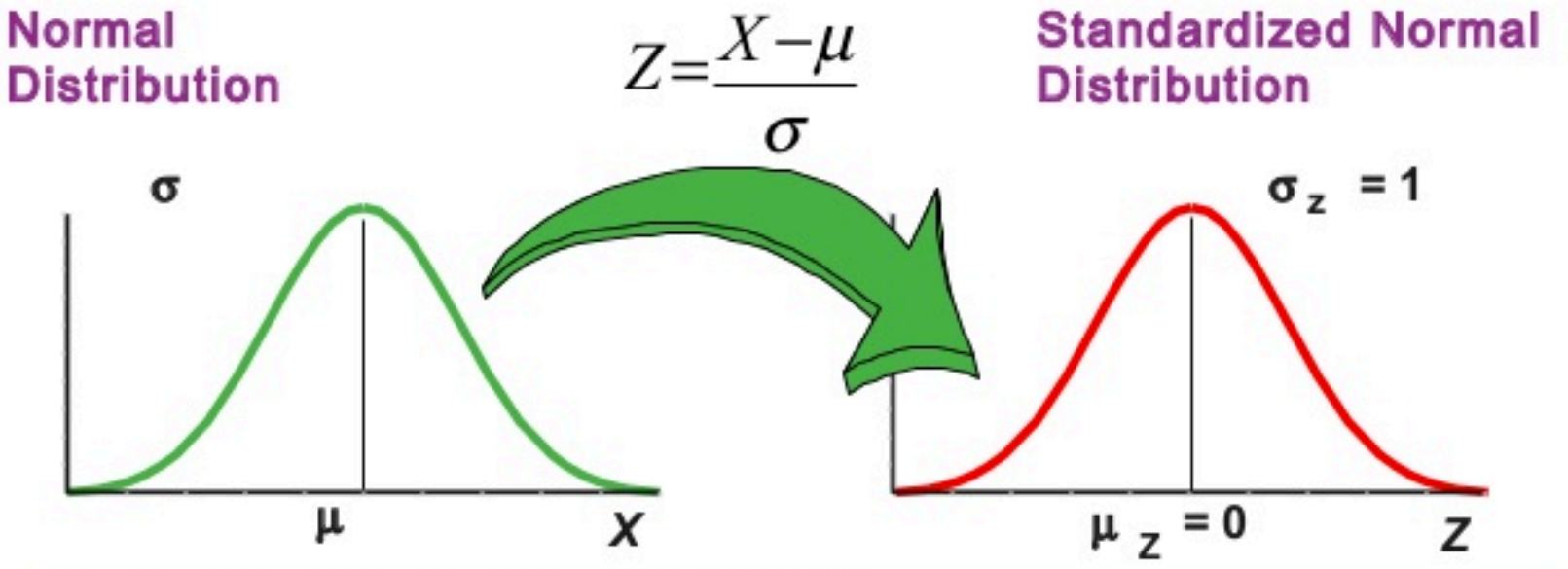
The *standard normal distribution* is a special member of the normal family that has a mean of *zero* and a standard deviation of *one*. We transform the equation of the normal distribution by substitute $z=(x-\mu)/\sigma$. The probability density function of this distribution is

$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$$

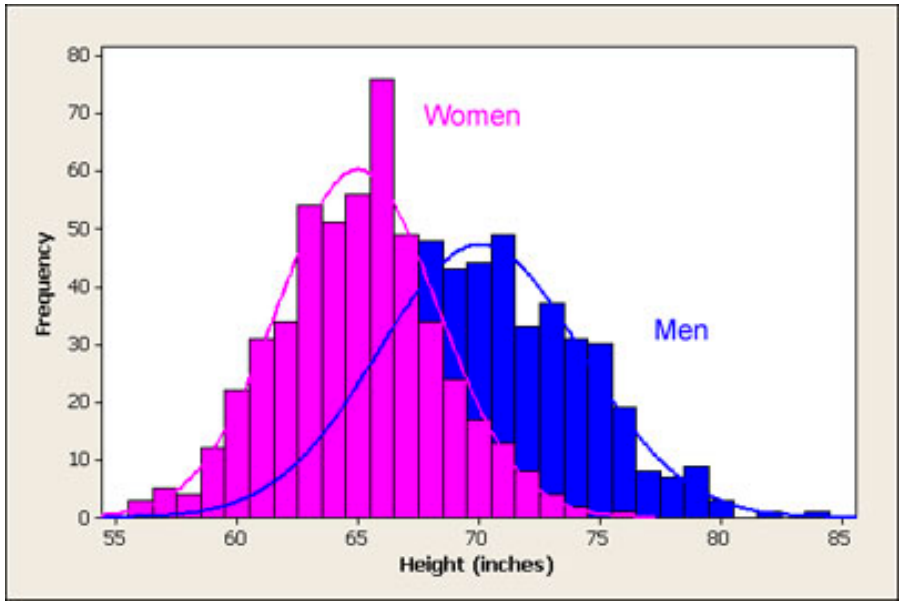
This definition of the normal distribution is often called *z distribution*.

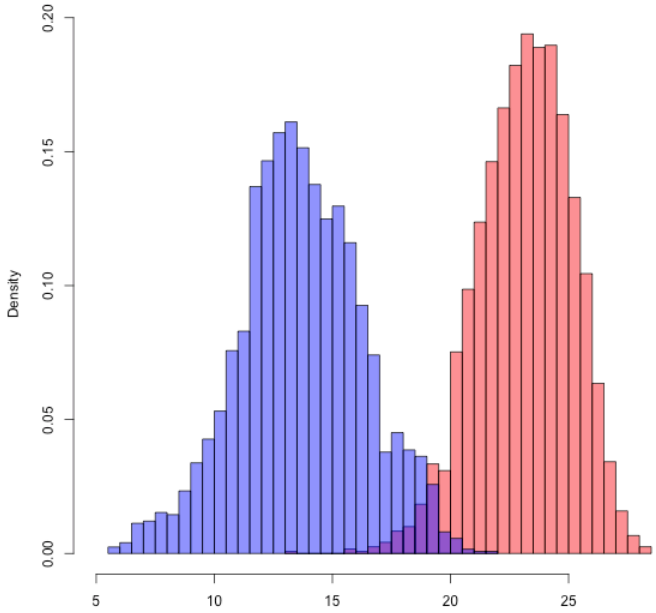


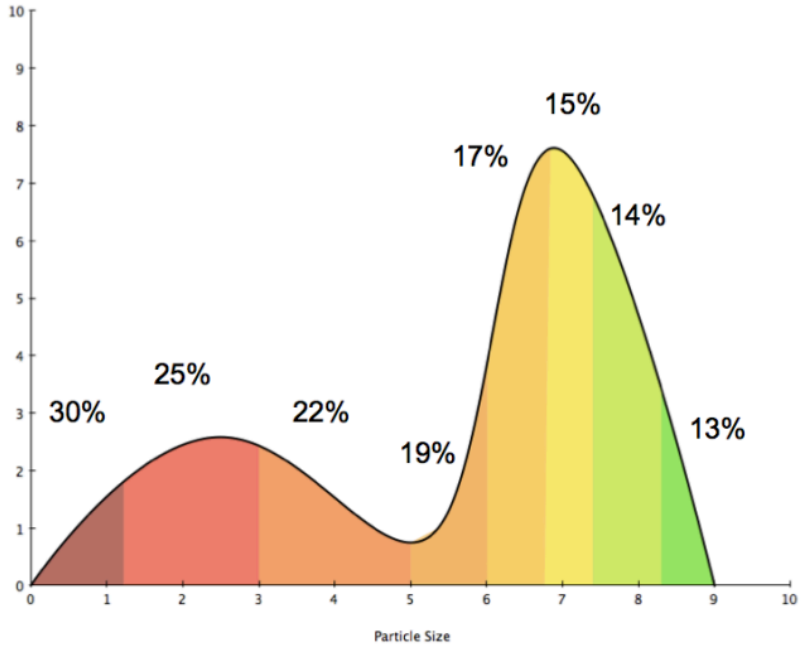
Standardize the Normal Distribution

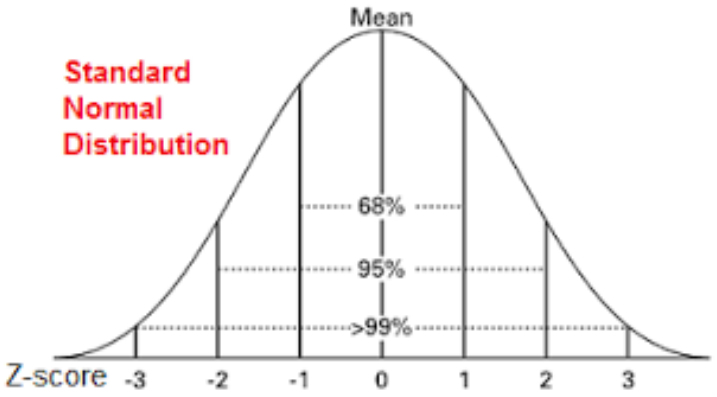


Because we can transform any normal random variable into standard normal random variable, we need only one table!


Women
Men
Frequency
Height (inches)
0
10
20
30
40
50
60
70
80
55
60
65
70
75
80
85


Density
0.00
0.05
0.10
0.15
0.20
5
10
15
20
25


30%
25%
22%
19%
17%
15%
14%
13%
Particle Size
0
1
2
3
4
5
6
7
8
9
10


Standard
Normal
Distribution
Mean
68%
95%
>99%
Z-score
-3
-2
-1
0
1
2
3

Logarithmic Normal or Log-Normal Distribution

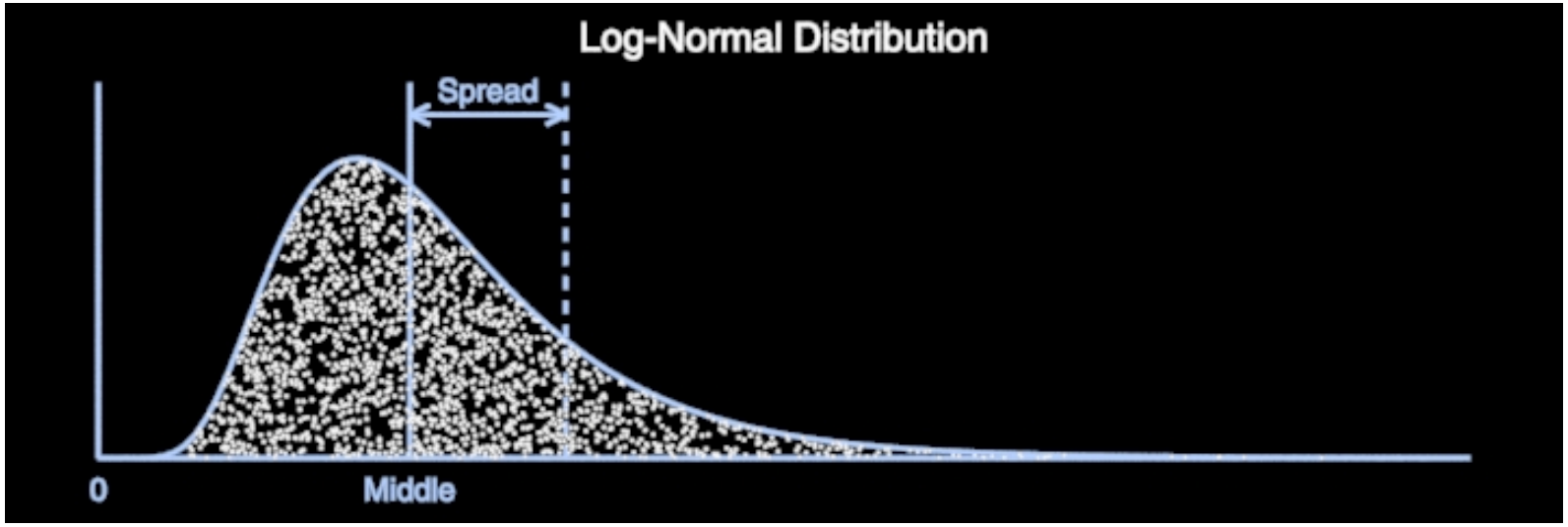
The *logarithmic normal distribution* is used when the data have a lower limit, e.g., mean-annual precipitation or the frequency of earthquakes
In such cases, distributions are usually characterized by significant skewness, which is best described by a logarithmic normal distribution. The probability density function of this distribution is

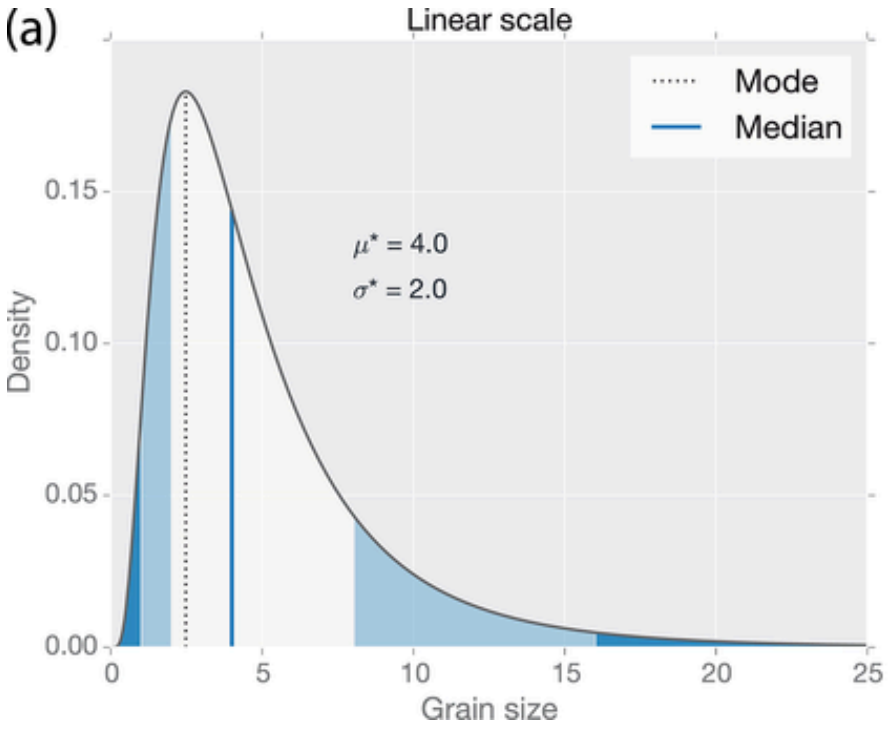
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}x}\exp\left(-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right)$$

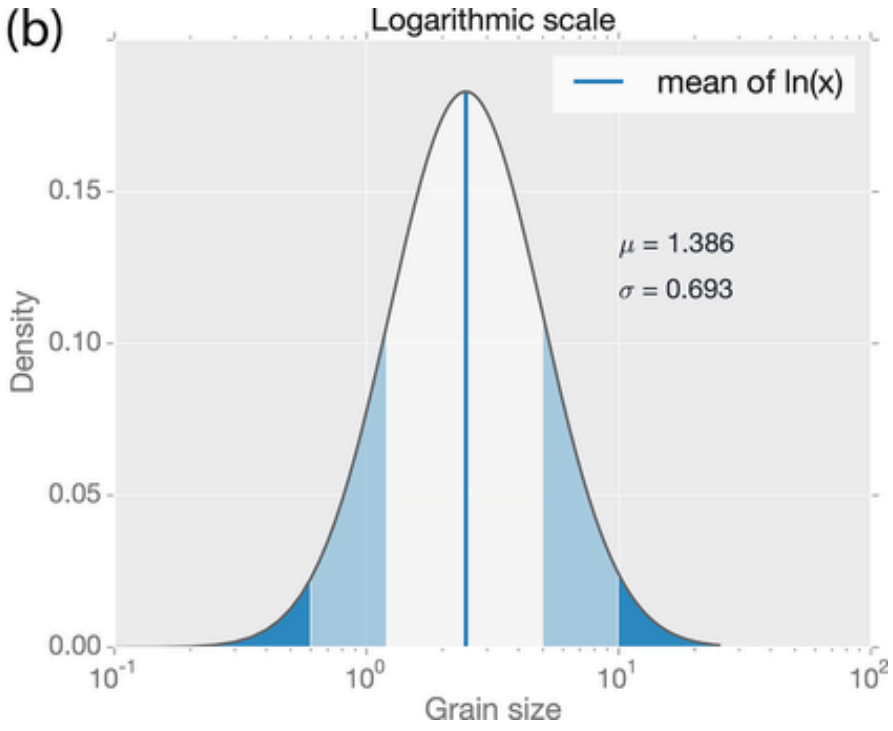
and the cumulative distribution function is

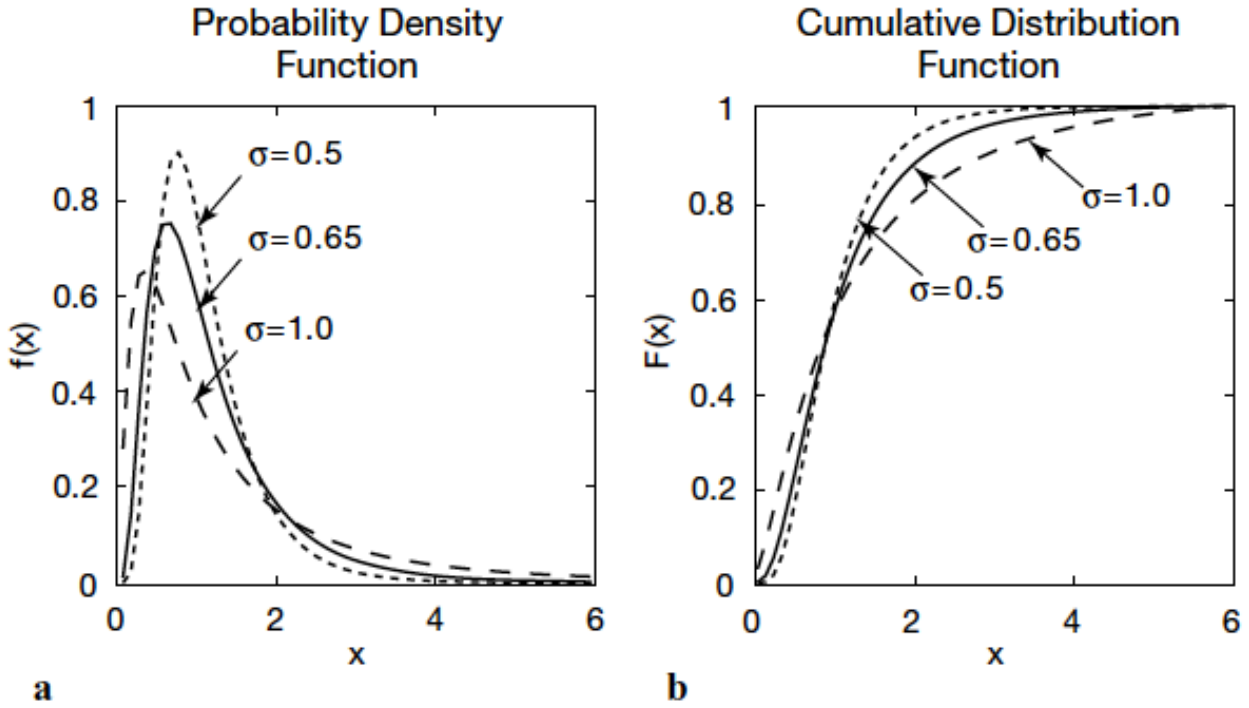
$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\frac{1}{y}\exp\left(-\frac{1}{2}\left(\frac{\ln y-\mu}{\sigma}\right)^2\right)dy$$

where $x>0$.


Log-Normal Distribution
Spread
0
Middle


(a)
Linear scale
Mode
Median
$\mu^* = 4.0$
$\sigma^* = 2.0$
Density
0.15
0.10
0.05
0.00
0
5
10
15
20
25
Grain size


(b)
Logarithmic scale
mean of ln(x)
$\mu = 1.386$
$\sigma = 0.693$
Density
0.15
0.10
0.05
0.00
10^{-1}
10^{0}
10^{1}
10^{2}
Grain size



a Probability density function $f(x)$ and **b** cumulative distribution function $F(x)$ of a logarithmic normal distribution with mean $\mu = 0$ and with different values for σ.

The distribution can be described by the two parameters mean μ and variance σ^2. The formulas for the mean and the variance, however, are different from the ones used for normal distributions. In practice, the values of x are logarithmized, the mean and the variance are computed using the formulas for the normal distribution and the empirical distribution is compared with a normal distribution.