Probability distributions

Dr Gavin M Abernethy

Contents

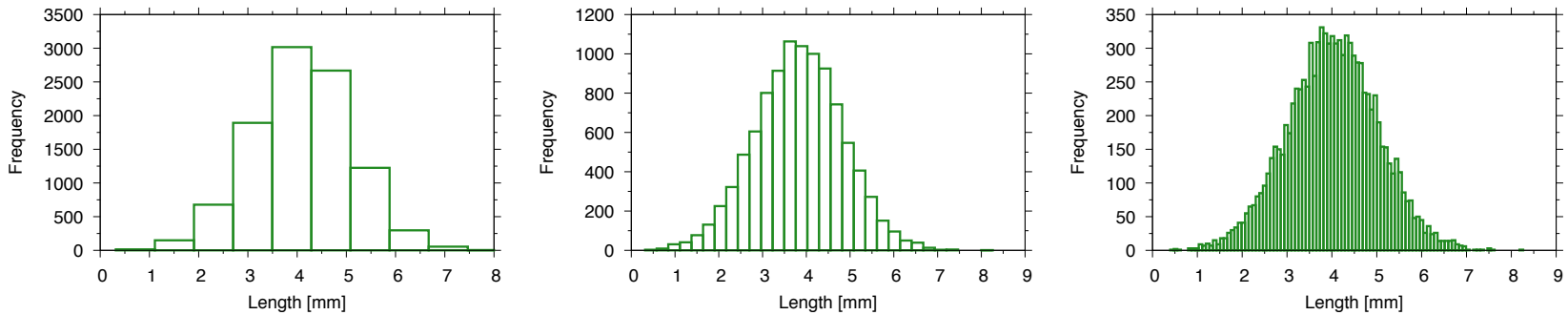
Today we will cover...

- Continuous probability distributions.
- Normal distributions.
- Performing calculations using normal distribution tables.
- Log-normal distributions.

From histograms to probability distributions

Histograms illustrate the frequency distribution of a data set for variable x in discrete bins.

If we place our data in increasingly narrow bins, this discrete distribution approaches a continuous curve, described by a frequency distribution function $g(x)$.



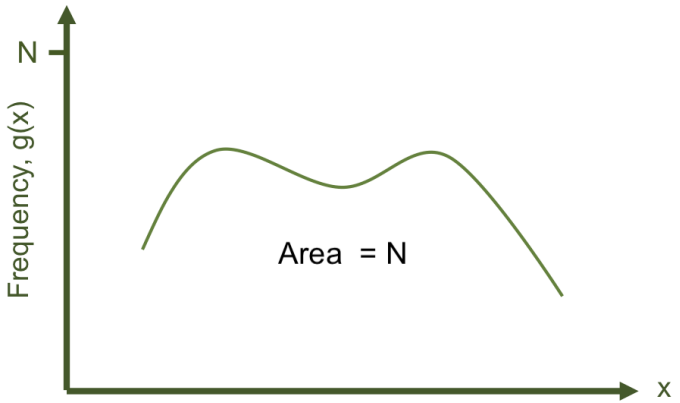
(a) 10 classes (b) 30 classes (c) 100 classes

Figure: Discrete frequency distributions

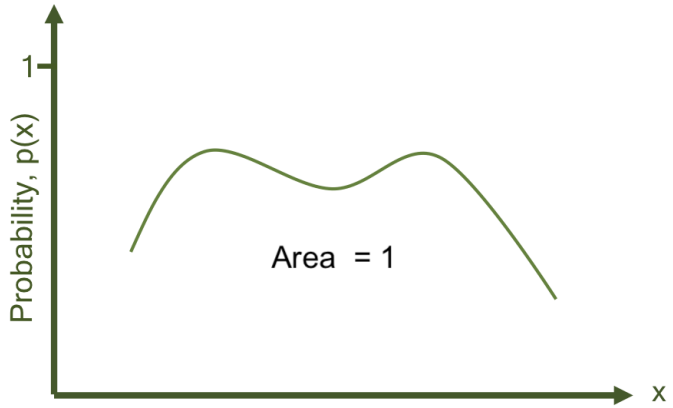
Probability distributions

If we then "normalise" the data, by dividing by the total number N of data points, we instead obtain a probability distribution $p(x)$.

This doesn't change the shape of the distribution, but the height is rescaled from a frequency to a probability (between 0 and 1).



(a) Frequency $g(x)$



(b) Probability $p(x)$

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = N$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Properties of a continuous probability distribution

In a probability distribution, the area under the curve $p(x)$ between $a \leq x \leq b$ gives the probability that the variable x takes a value in this interval. **Probability and area are the same in this context!**

Mean and variance

For a continuous variable x with probability distribution $p(x)$,

Mean:

$$\bar{x} = \int_{-\infty}^{\infty} xp(x)\ \mathrm{d}x$$

Variance:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \bar{x})^2 p(x)\ \mathrm{d}x$$

The Normal Distribution

So instead of frequency tables, we may be able to describe how a random variable (e.g. heights of students) behaves, with a function that is the particular probability distribution.

Many natural properties (such as heights of people), obey the **Normal distribution** - also known as a Gaussian distribution or a Bell curve. It is often useful to fit such a distribution to a data set so that the standard deviation and mean can be easily estimated.

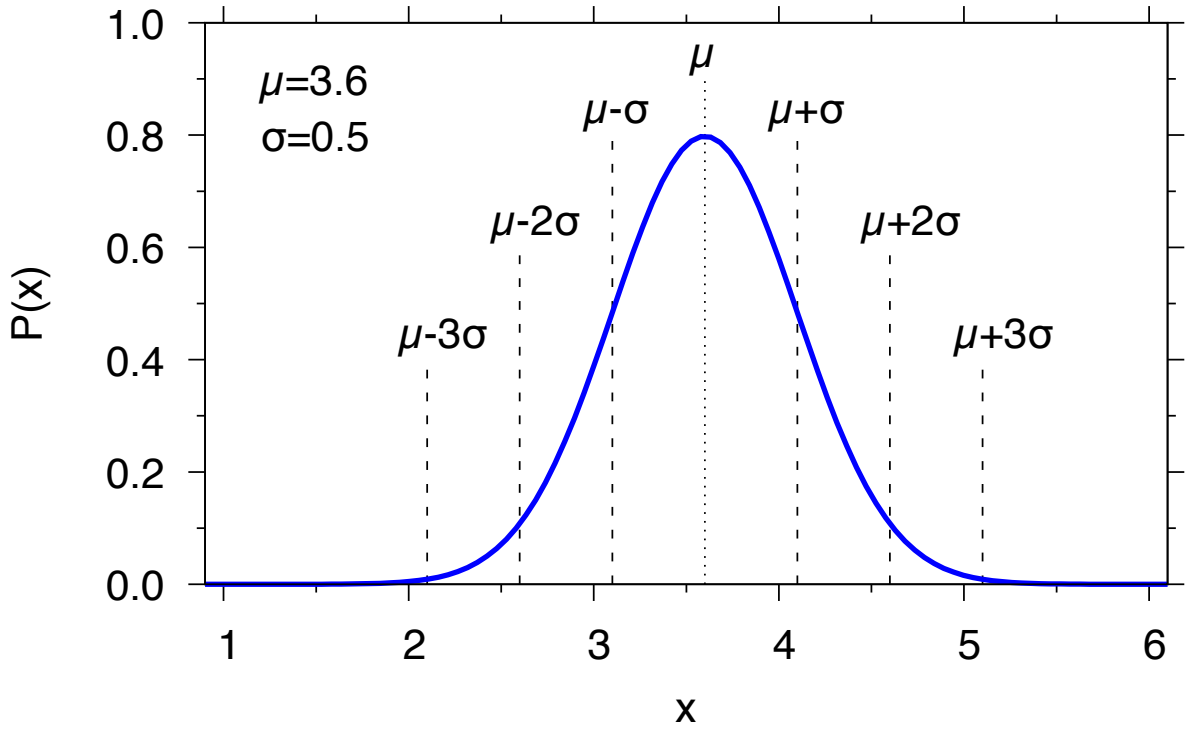
The **normal probability distribution function**, takes the form:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ is the *mean* and σ is the *standard deviation*.

The Normal Distribution

Sketching this probability distribution function $P(x)$:



Around 68% of the area under a normal distribution curve will be within 1 standard deviation (from $\mu - \sigma$ to $\mu + \sigma$). Around 95% within 2 standard deviations and 99% within 3 standard deviations.

The Normal Distribution

In general, the mean of a variable with any continuous probability distribution function $p(x)$ is:

$$\bar{x} = \int_{-\infty}^{\infty} xp(x)\ \mathrm{d}x$$

This will also hold for this *particular* probability distribution function:

$$\begin{aligned} \bar{x} &= \int_{-\infty}^{\infty} xP(x)\ \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ \mathrm{d}x \\ &= \mu \end{aligned}$$

Notation

If X is a normally-distributed variable with a mean μ, and standard deviation σ, then we write this formally as:

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

In words, this says: "Variable X obeys a normal distribution with mean, μ, and variance σ^2."

Note: Be very careful not to mix up the variance and the standard deviation!

Standard normal distribution and the tables

To solve problems concerning a normally-distributed variable, we convert them to the **standard normal distribution**, satisfying:

$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

It is often denoted, Z, so we would write:

$$Z \sim N(0, 1)$$

By converting to this distribution, we can obtain solutions from its table of values. The diagram shows the area that we want, from the centre of the curve (zero) to some value z which is measured in how many standard deviations it is from the mean. This number to 1 d.p. is located on the left hand side, and the columns of the table (0-9) correspond to the second decimal place.

Standard normal distribution and the tables

So, to find the probability that variable $X \sim N(\mu, \sigma^2)$ lies in the range $\mu < X < x$, we convert to the standard distribution using:

$$z = \frac{|x - \mu|}{\sigma}$$

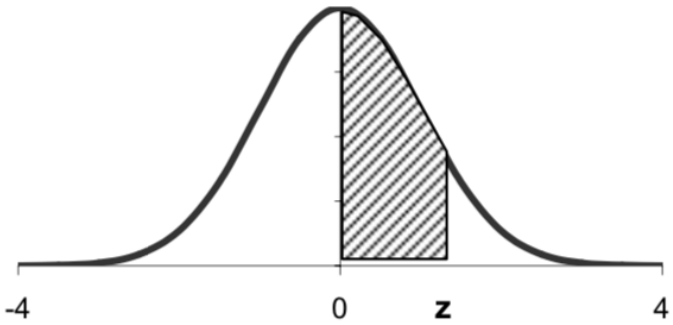
Then look up z in the table to obtain the probability/area under $0 < Z < z$ in the standard distribution, which corresponds to the probability $P(\mu < X < x)$ in our original distribution.

To find other ranges, such as $x < X < +\infty$, we may need to use some extra steps using the fact that exactly half of the distribution is above and below the mean.

Standard normal distribution table

Standard Normal Distribution

Areas under the Standard Normal Curve from 0 to z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830

Example: Normal distribution (I/VI)

A company manufactures a microprocessor to control industrial robots. Existing data indicates that the life span of the microprocessors is described by a normal distribution with mean $\mu = 4000$ hours and standard deviation $\sigma = 200$ hours.

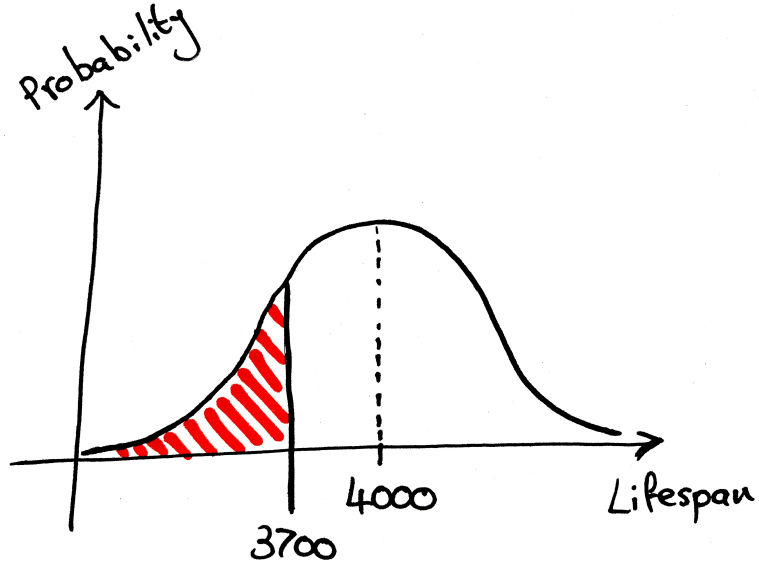
Determine the probability that the life span of such a microprocessor is:

1. Less than 3700 hours.
2. Between 3700 hours and 4250 hours.
3. More than 4250 hours.

In each case, it will be helpful to draw the curve first.

Example: Normal distribution (II/VI)

(i) We want the area below 3700:



Let's call the lifespan variable X, then $X \sim N(4000, 200^2)$ and first we want to find the probability $P(X < 3700)$.

This can be found by calculating the area from 3700 to 4000 (the same as that from 4000 to 4300) and subtracting it from 1/2.

Example: Normal distribution (III/VI)

Calculate the number of standard deviations from the mean:

$$\text{No. Standard Deviations} = \frac{|3700 - \mu|}{\sigma} = \frac{4000 - 3700}{200} = 1.50$$

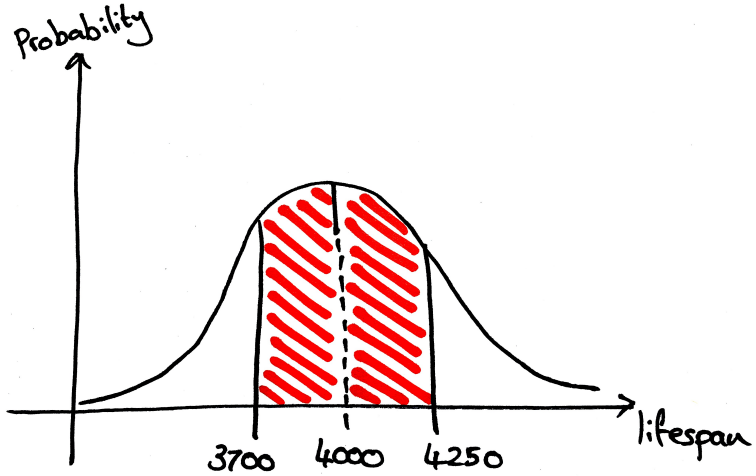
From looking at the table, this value is 0.4332.

Thus, our desired area (probability) is:

$$\begin{aligned} P(X < 3700) &= 0.5 - P(3700 < X < 4000) \\ &= 0.5 - 0.4332 \\ &= 0.0668 \end{aligned}$$

Example: Normal distribution (IV/VI)

(ii) We want the area between 3700 and 4250:



We already have the area from 3700 to 4000 (0.4332), now we need the area from 4000 to 4250.

Example: Normal distribution (V/VI)

Calculate the number of standard deviations from the average:

$$\text{No. Standard Deviations} = \frac{|4250 - \mu|}{\sigma} = \frac{4250 - 4000}{200} = 1.25$$

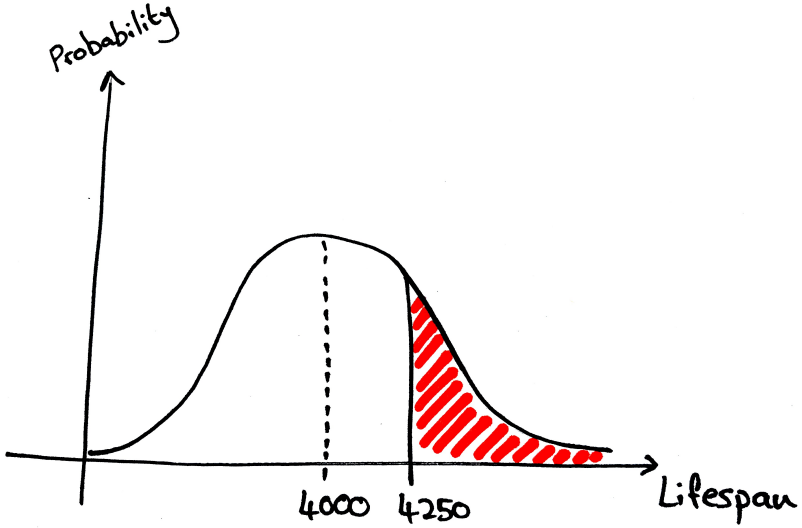
From looking at the table, this value is 0.3944.

Thus, our total area is:

$$\begin{aligned} P(3700 < X < 4250) &= P(3700 < X < 4000) + P(4000 < X < 4250) \\ &= 0.4332 + 0.3944 \\ &= 0.8276 \end{aligned}$$

Example: Normal distribution (VI/VI)

(iii) We want the area above 4250:



We already found that the area under $4000 < X < 4250$ is 0.3944.
So all we need to do is subtract this from 0.5:

$$\begin{aligned} P(X > 4250) &= 0.5 - P(4000 < X < 4250) \\ &= 0.5 - 0.3944 \\ &= 0.1056 \end{aligned}$$

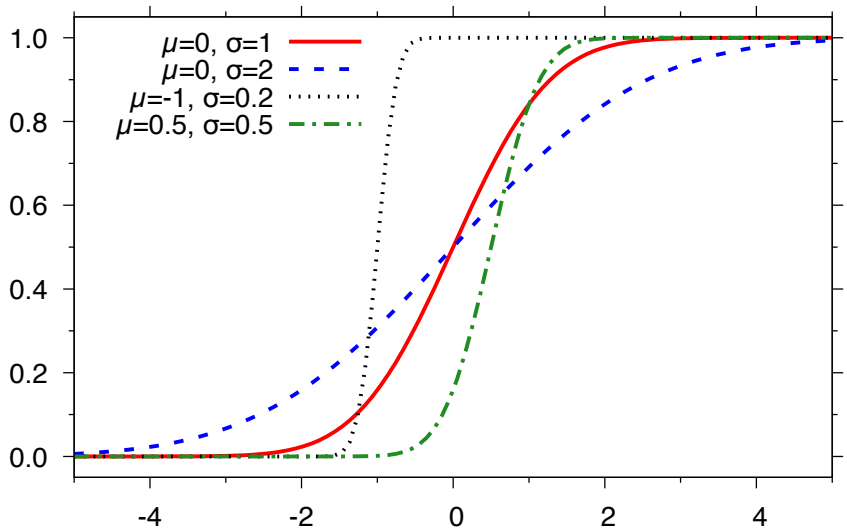
Cumulative Frequency Distribution

An alternative representation of a distribution is the cumulative frequency distribution (CFD). This is the sum of the area under the probability distribution curve up to that value of x, so it tends to 1.

For the normal distribution, this is given by:

$$\Phi(x) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \quad \text{where} \quad erf(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-t^2}\,\mathrm{d}t$$

This is illustrated below for a range of μ and σ:



Summary: Normal Distribution

To summarise the important properties of a normal distribution $X \sim N(\mu, \sigma^2)$:

Probability distribution function	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Cumulative distribution function	$\frac{1}{2} + \frac{1}{2} erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)$

Log-Normal Distribution

A related common distribution is the log-normal distribution, which describes grain sizes in the polycrystalline material $Cu(In,Ga)Se_2$.

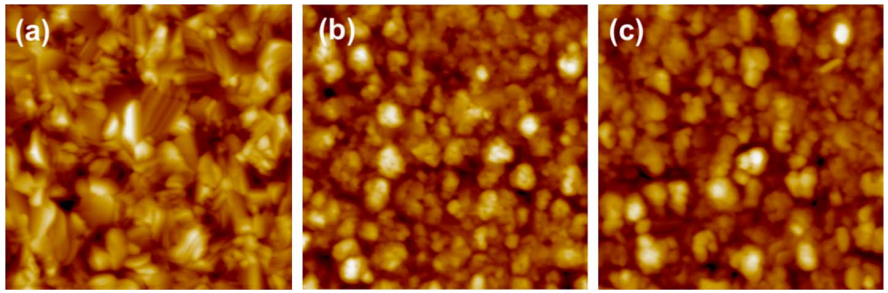


Figure: AFM images of three semiconductor devices taken from Microscopy Today, Volume 26, Issue 3 pages 32-39 (2018).

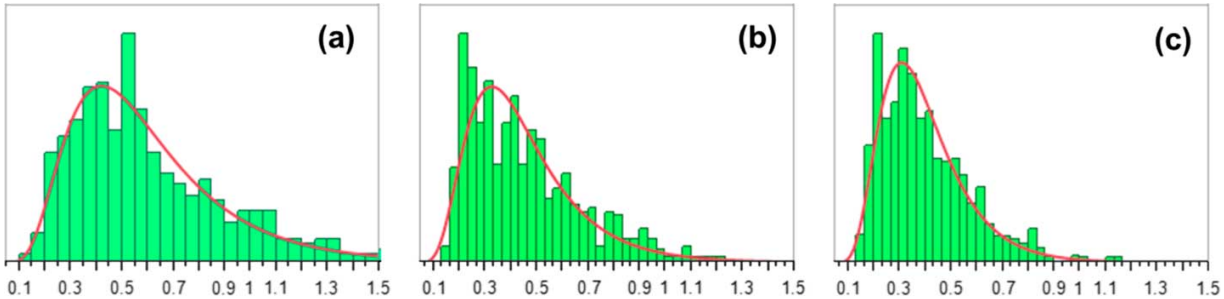


Figure: Grain size distributions of three devices taken from Microscopy Today, Volume 26, Issue 3 pages 32-39 (2018).

Log-Normal Distribution

A log-normally distributed variable x has pdf:

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

This formula is somewhat similar in form to the normal distribution. Unlike a normal distribution, whose mean is μ (the middle of the curve) and the variance just given by σ^2, the log-normal distribution has the following properties:

Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Mode	$\exp(\mu - \sigma^2)$
Variance	$[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$
Cumulative distribution function	$\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(x)-\mu}{\sqrt{2}\sigma}\right)$

Log-Normal Distribution

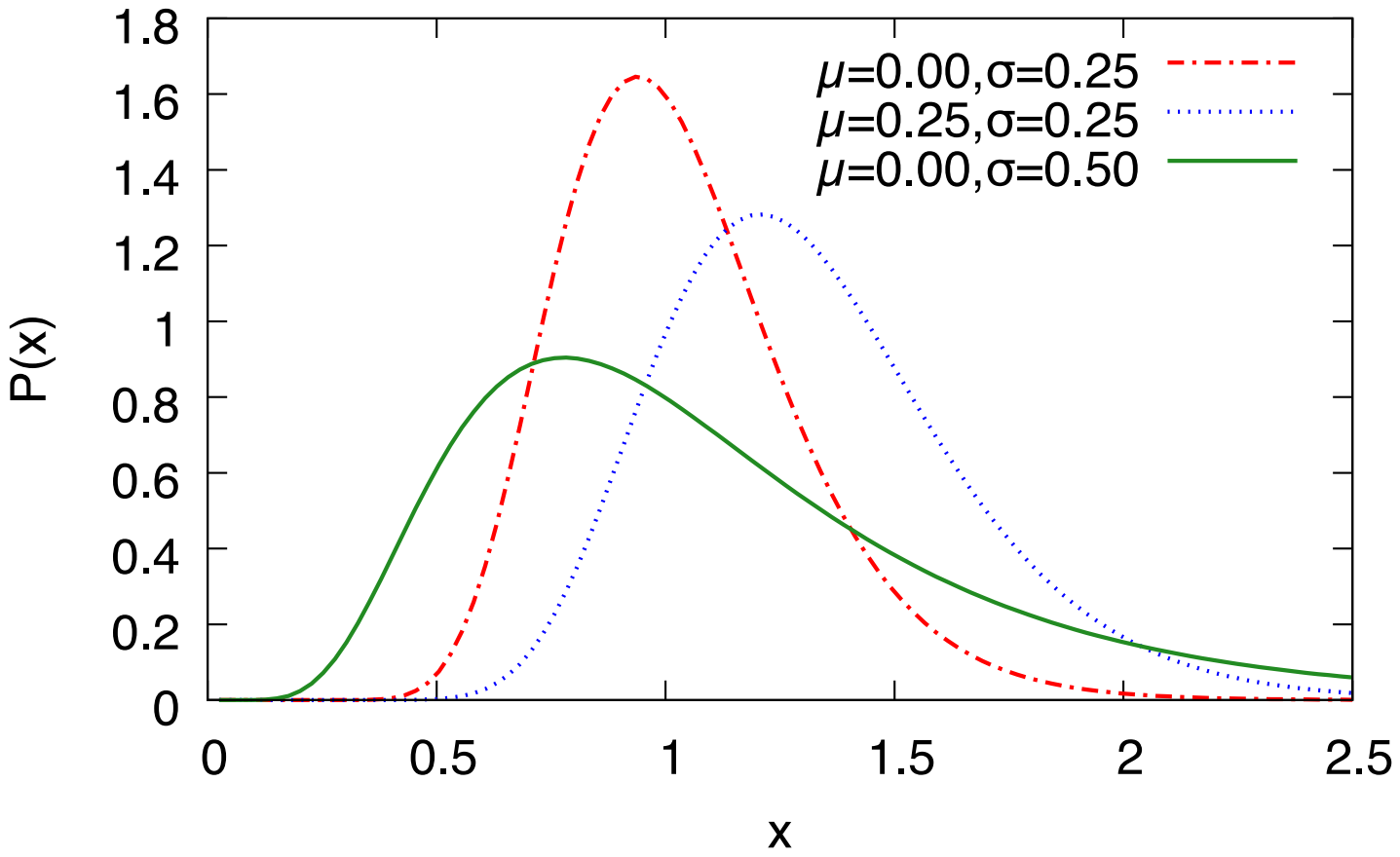


Figure: Examples of log-normal probability distributions