Normal Distribution

Many families of distributions include members that show a familiar bell-shaped curve.

R plots of binomial, Poisson, and gamma

```r
source("R-code/plot-distributions.R");
plotBinom(10, .3)
plotPoisson(3)
plotGamma(3)
```
In each case, the underlying curve is a normal or Gaussian density function.

If $X$ has a Gaussian distribution, and

$$E(X) = \mu, \ V(X) = \sigma^2,$$

then the probability density function of $X$ is

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

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

plotGaussian(mu = 2, sigma = 0.5)
The standard normal distribution has $\mu = 0$ and $\sigma = 1$; the probability density function simplifies to

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$$

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

$$Z = \frac{X - \mu}{\sigma},$$

then $Z$ follows the standard normal distribution: $Z \sim N(0, 1)$.

Conversely, if $Z \sim N(0, 1)$ and

$$X = \mu + \sigma Z,$$

then $X \sim N(\mu, \sigma^2)$. 
Cumulative distribution function

The cumulative distribution function of $N(0, 1)$ is written

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)du.$$ 

It cannot be written in closed form, but is tabulated and programmed in most languages, such as the `pnorm()` function in R:

curve(pnorm(x), from = -3, to = 3)
Example: Normally distributed current

The measured current $X$ in a wire is normally distributed with mean $\mu = 10$ mA and standard deviation $\sigma = 2$ mA.

What is the probability that $X > 13$ mA?

\[
P(X > 13) = 1 - P(X \leq 13)
\]

\[
= 1 - P \left( \frac{X - 10}{2} \leq \frac{13 - 10}{2} \right)
\]

\[
= 1 - P(Z \leq 1.5)
\]

\[
= 1 - \Phi(1.5).
\]

\[
1 - \text{pnorm}(1.5)
\]

# or more simply

\[
\text{pnorm}(13, \text{mean} = 10, \text{sd} = 2, \text{lower.tail} = \text{FALSE})
\]
What is the probability that the current is between 9 and 11 mA?

\[
P(9 < X \leq 11) = P(X \leq 11) - P(X \leq 9)
\]
\[
= \Phi \left( \frac{11 - 10}{2} \right) - \Phi \left( \frac{9 - 10}{2} \right)
\]
\[
= \Phi(0.5) - \Phi(-0.5)
\]
\[
= \Phi(0.5) - [1 - \Phi(0.5)]
\]
\[
= 2\Phi(0.5) - 1
\]

2 * pnorm(0.5) - 1
# or, since pnorm covers negative numbers,
pnorm(0.5) - pnorm(-0.5)
# or even
pnorm(11, mean = 10, sd = 2) - pnorm(9, mean = 10, sd = 2)
Normal Approximations

Many families of distributions include members that show a familiar bell-shaped curve.

The Gaussian distribution can be used to make approximate probability statements for the corresponding members of those families.

In each case, we begin by standardizing the variable.
Example: Binomial distribution

Suppose that $X$ has the binomial distribution with parameters $n$ and $p$: $X \sim \text{Bin}(n, p)$.

Then

$$E(X) = np,$$

and

$$V(X) = np(1 - p),$$

so

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

is the corresponding standardized variable, with expected value 0 and standard deviation 1.
Then if \( n \) is large and \( p \) is not close to either 0 or 1, the distribution of \( Z \) is close to the standard normal distribution.

To be precise,

\[
P(Z \leq z) \approx \Phi(z),
\]

and hence

\[
P(X \leq x) = P \left( Z \leq \frac{x - np}{\sqrt{np(1 - p)}} \right) \\
\approx \Phi \left( \frac{x - np}{\sqrt{np(1 - p)}} \right).
\]
Continuity correction

The approximation is usually used with $x$ an integer, and in that case

$$P(X \leq x) = P(X \leq x + 0.5)$$

and

$$P(X \leq x) \approx \Phi \left( \frac{x + 0.5 - np}{\sqrt{np(1 - p)}} \right)$$

is found to give a closer approximation to the exact probability.
Of course, 

\[ P(X > x) = 1 - P(X \leq x) \]

is approximated by

\[ P(X > x) \approx 1 - \Phi \left( \frac{x + 0.5 - np}{\sqrt{np(1 - p)}} \right). \]
But

\[ P(X \geq x) = P(X > x - 1) \]

is approximated by

\[ P(X \geq x) \approx 1 - \Phi \left( \frac{x - 1 + 0.5 - np}{\sqrt{np(1 - p)}} \right) = 1 - \Phi \left( \frac{x - 0.5 - np}{\sqrt{np(1 - p)}} \right). \]

When the inequality is inclusive (≤ or ≥), the sign of the correction term ±0.5 is chosen to increase the approximating probability.
Example: Poisson distribution

Suppose that $X$ has the Poisson distribution with parameter $\theta$: $X \sim \text{Poisson}(\theta)$.

Then

$$E(X) = V(X) = \theta,$$

so

$$Z = \frac{X - \theta}{\sqrt{\theta}}$$

is the corresponding standardized variable, with expected value 0 and standard deviation 1.
Then if $\theta$ is large, the distribution of $Z$ is close to the standard normal distribution.

The approximation is used in the same way as for the binomial distribution, including the use of a continuity correction:

$$P(X \leq x) \approx \Phi \left( \frac{x + 0.5 - \theta}{\sqrt{\theta}} \right),$$

and so on.
How large is large enough?

You will sometimes read that these approximations are good if:

- \( np > 5 \) and \( n(1 - p) > 5 \), in the binomial case;
- \( \theta > 5 \) in the Poisson case.

However, the approximation error can be as great as 0.03, so you are guaranteed no more than a rough approximation.

Modern software makes it easy to use “exact” calculations (or at least much more accurate approximations), so these approximations are more useful in understanding the reason for the ubiquitous bell curve than in actual computation.
The Gaussian distribution in the real world

The central limit theorem (CLT) explains why the Gaussian distribution appears in theory.

It states essentially that any random variable that is the sum of many small, independent contributions is approximately Gaussian.

For instance, a binomial random variable is the sum of the success indicators for each trial.

In many real world measurement systems, the random measurement noise does consist of the sum of many small perturbations, and by the CLT could be expected to be approximately Gaussian.

But it is only an approximation, and often not a good one!
Lognormal Distribution

The Gaussian distribution is often used as a model for a measured physical quantity.

But every Gaussian distribution has a positive probability of negative values, which is a deficiency if the physical quantity is always positive.

One alternative that is often used is the lognormal distribution.
Suppose that $W \sim N(\theta, \omega^2)$, and that $X = \exp(W)$; then conversely $W = \log(X)$, and $X$ has the lognormal distribution with parameters $\theta$ and $\omega$.

**Expected value:**

$$E(X) = e^{\theta + \omega^2 / 2}.$$

**Variance:**

$$V(X) = e^{2\theta + \omega^2} \left( e^{\omega^2} - 1 \right)$$

$$= E(X)^2 \left( e^{\omega^2} - 1 \right).$$
If some observations are thought to have the lognormal distribution, we simply take their logarithms and use the methods for the normal distribution.

The lognormal distribution is often used as a model for environmental variables like concentrations of air pollutants.

Central Limit Theorem
Just as the Gaussian distribution describes measurements where many small independent effects are added together, the lognormal distribution arises when many independent effects are multiplied together.

- Eg: distribution of masses of grains of sand.