Continuous distributions

2. Density functions

Continuous r.v. take values from intervals on the real line.

The **(probability) density function** (p.d.f.) of a continuous r.v. X is the function $f(x)$ such that, for any constants $a \leq b$

$$\Pr(a \leq X \leq b) = \int_a^b f(x)dx \, .$$

Note that $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

The probability density function defines the **distribution** of X.

Mean and variance of a continuous r.v.

The definitions given in the discrete case are readily extended.

The **mean (expected value)** of a continuous r.v. X is

$$E(X) = \int_{-\infty}^{\infty} x\, f(x) dx\,,$$

and the definition is extended to any function g of X

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx\,.$$

This includes the **variance** as a special case.

Two results, quite useful for continuous r.v., apply to a *linear transformation* $a + b\,X$, with a, b constants:

$$\begin{aligned} E(a + b\,X) &= a + b\,E(X) \\ \mathrm{var}(a + b\,X) &= b^2\,\mathrm{var}(X)\,. \end{aligned}$$

Notable continuous random variables

Important continuous distributions include:

- Normal distribution
- Gamma, exponential and χ^2 distribution
- F distribution
- t and Cauchy distributions
- Beta distribution

The normal distribution has a major role in statistics. The χ^2, t and F distributions are *relative* of the normal distribution.

The normal distribution

A r.v. X has a normal (or *Gaussian*) distribution if it has p.d.f.

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

The notation is $X \sim \mathcal{N}(\mu, \sigma^2)$, and $E(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$, $\sigma^2 > 0$, $\mu \in \mathbb{R}$.

An important property is that for any constants a, b

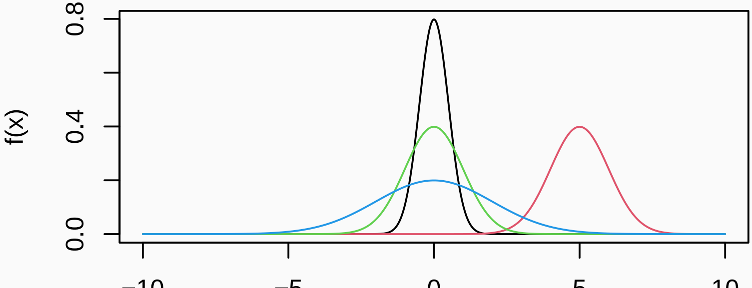
$$a + b\,X \sim \mathcal{N}(a + b\,\mu, b^2\,\sigma^2)\,,$$

so that $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$, the **standard normal distribution**.

Finally, $Y = e^X$ has a **lognormal distribution**, useful for asymmetric variables with occasional right-tail outliers.

R lab: the normal distribution

```
xx <- seq(-10, 10, l=1000)
plot(xx, dnorm(xx, 0, 0.5), xlab ="x", ylab ="f(x)", type ="l")
lines(xx,  dnorm(xx, 5, 1), col = 2)
lines(xx,  dnorm(xx, 0, 1), col = 3)
lines(xx,  dnorm(xx, 0, 2), col = 4)
```



The Gamma and the exponential distributions

A r.v. X has a Gamma distribution if it has the following pdf

$$f(x) = \frac{\lambda^{\alpha} x^{\alpha-1} \quad e^{-\lambda x}}{\Gamma(\alpha)}, \qquad x \geq 0$$

where $\lambda, \alpha > 0$ and $\Gamma(\alpha) = \int_0^{\infty} e^{-x} x^{\alpha-1} dx$.

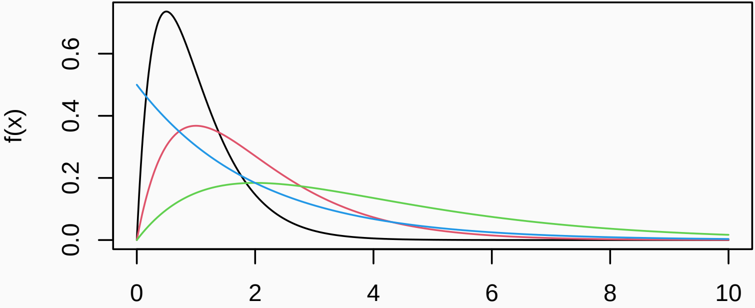
The notation is $X \sim Ga(\alpha, \lambda)$, $E(X) = \frac{\alpha}{\lambda}$ and $\mathrm{var}(X) = \frac{\alpha}{\lambda^2}$.

When α is an integer it is also called **Erlang** distribution.

When $\alpha = 1$ it is called **exponential** distribution. The exponential distribution is related to the Poisson r.v. since X represents the waiting times between two arrivals in a Poisson process (The process which generates the Poisson rv)

Rlab: The Gamma and the exponential distributions

```
xx <- seq(0, 10, l=1000)
plot(xx, dgamma(xx, 2, 2), xlab ="x", ylab ="f(x)", type ="l")
lines(xx,  dgamma(xx, 2, 1), col = 2)
lines(xx,  dgamma(xx, 2, .5), col = 3)
lines(xx,  dgamma(xx, 1, .5), col = 4) # exponential distributio
```



The Beta (and the uniform) distribution

A r.v. X has a Beta distribution if it has the following pdf

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1$$

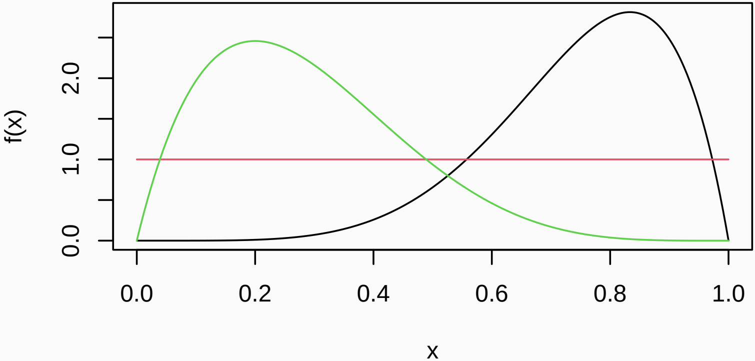
$\alpha, \beta > 0$

The notation is $X \sim Be(\alpha, \beta)$, $E(X) = \frac{\alpha}{\alpha+\beta}$ and $\mathrm{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

The **Uniform** distribution on $[0, 1]$ is a special case when $\alpha = 1$ and $\beta = 1$.

R lab: the Beta distribution

```
xx <- seq(0, 1, l=1000)
plot(xx, dbeta(xx, 6,2), xlab ="x", ylab ="f(x)", type ="l")
lines(xx,  dbeta(xx, 1,1), col = 2)
lines(xx,  dbeta(xx, 2, 5), col = 3)
```



The χ^2 distribution

Let $Z_1, \ldots, Z_k$ be a set of independent $\mathcal{N}(0,1)$ r.v., then $X = \sum_{i=1}^{k} Z_i^2$ is a r.v. with a χ^2 **distribution with k degrees of freedom**.

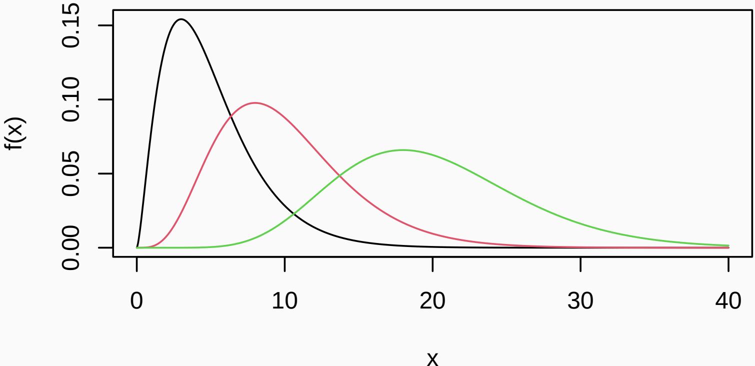
The notation is $X \sim \chi_k^2$, $E(X) = k$ and $\mathrm{var}(X) = 2\,k$.

It is a special case of the Gamma distribution. In fact a χ^2 distribution with k degrees of freedom is a Gamma distribution with parameters $\alpha = k/2$ and $\lambda = 1/2$.

It plays an important role in the theory of hypothesis testing in statistics.

R lab: the χ^2 distribution

```
xx <- seq(0, 40, l=1000)
plot(xx, dchisq(xx, 5), xlab ="x", ylab ="f(x)", type ="l")
lines(xx,  dchisq(xx, 10), col = 2)
lines(xx,  dchisq(xx, 20), col = 3)
```



The F distribution

Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$, independent, then the r.v.

$$F = \frac{X/n}{Y/m}$$

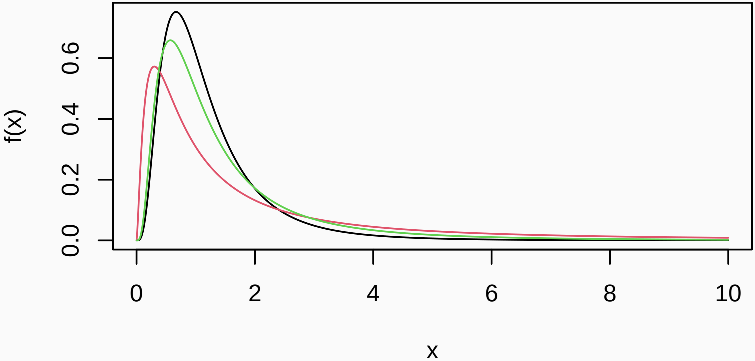
has an F **distribution with** n **and** m **degrees of freedom**.

The notation is $F \sim \mathcal{F}_{n,m}$, and $E(F) = m/(m-2)$ provided that $m > 2$.

The distribution is almost never used as a model for observed data, but it has a central role in hypothesis testing involving linear models.

R lab: the F distribution

```
xx <- seq(0, 10, l=1000)
plot(xx, df(xx, 10, 10), xlab ="x", ylab ="f(x)", type ="l")
lines(xx,  df(xx, 5, 2), col = 2)
lines(xx,  df(xx, 10, 5), col = 3)
```



The t and Cauchy distributions

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_n$, independent, then the r.v.

$$T = \frac{Z}{\sqrt{\frac{X}{n}}}$$

has an t **distribution with** n **degrees of freedom**.

The notation is $T \sim t_n$, and $E(T) = 0$ provided that $n > 1$, whereas $\mathrm{var}(T) = n/(n-2)$ provided that $n > 2$.

t_∞ is $\mathcal{N}(0,1)$, while for n finite the distribution has heavier tails than the standard normal distribution.

The case t_1 is the **Cauchy distribution**.

The distribution has a central role in statistical inference; at times it is used for modelling phenomena presenting *outliers*.

R lab: the t and Cauchy distributions

```
xx <- seq(-5, 5, l=1000)
plot(xx, dnorm(xx, 0, 1), xlab ="x", ylab ="f(x)", type ="l")
lines(xx,  dt(xx, 30), col = 2)
lines(xx,  dt(xx, 5), col = 3)
lines(xx,  dt(xx, 1), col = 4)
```