Review of some probability concepts: random variables

(A quick tour)

L. Egidi

Fall 2021

University of Trieste

Random variables

Discrete distributions

Continuous distributions

C.d.f. and quantile functions

Random variables

Random variables

Statistics is about the extraction of information from data that contain an *unpredictable* component.

Random variables (r.v.) are the mathematical device employed to build *models* of this variability.

A r.v. takes a different value at random each time is observed.

Distribution of a r.v.

The main tools used to describe the **distribution** of values taken by a r.v. are:

1. Probability functions
2. Density functions
3. Cumulative distribution functions
4. Quantile functions

Discrete distributions

1. Probability functions

Discrete r.v. take values in a discrete set.

The **probability (mass) function** of a discrete r.v. X is the function $f(x)$ such that

$$f(x) = \Pr(X = x)\,.$$

with $0 \leq f(x) \leq 1$ and $\sum_i f(x_i) = 1$.

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

Mean and variance of a discrete r.v.

For many purposes, the first two moments of a distribution provide a useful summary.

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

$$E(X) = \sum_i x_i \, f(x_i)\,,$$

and the definition is extended to any function g of X

$$E\{g(X)\} = \sum_i g(x_i)\, f(x_i)\,.$$

The special case $g(X) = (X-\mu)^2$, with $\mu = E(X)$, is the **variance** of X

$$\mathrm{var}(X) = E\{(X-\mu)^2\} = E(X^2) - \mu^2.$$

The **standard deviation** is just given by $\sqrt{\mathrm{var}(X)}$.

Notable discrete random variables

Discrete r.v. often used in applications:

- Binomial (and Bernoulli) distribution
- Poisson distribution
- Negative binomial distribution
- Geometric distribution
- Hypergeometric distribution

Let us give a closer look to some of them.

The binomial distribution

Consider n independent binary trials each with success probability p, $0 < p < 1$. The r.v. X that counts the number of successes has **binomial distribution** with probability function

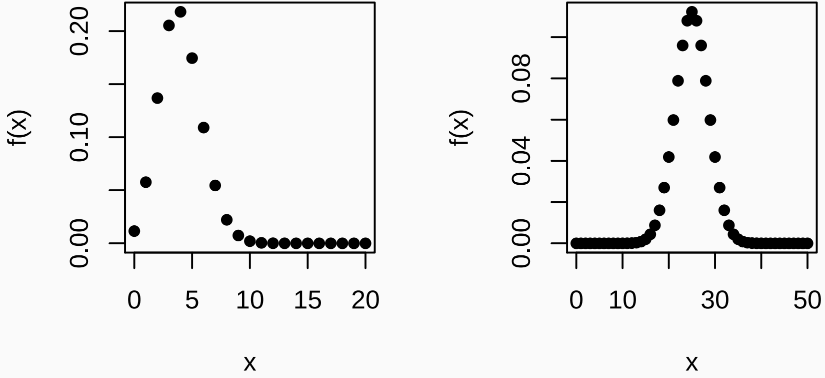
$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

The notation is $X \sim \mathcal{B}_i(n, p)$, and $E(X) = np$, $\mathrm{var}(X) = np(1-p)$.

The case when $n = 1$ is known as **Bernoulli distribution** and a single binary trial is called **Bernoulli trial**.

R lab: the binomial distribution

```r
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dbinom(0:20, 20, 0.2), xlab = "x", ylab = "f(x)")
plot(0:50, dbinom(0:50, 50, 0.5), xlab ="x", ylab = "f(x)")
```



The Poisson distribution

The special case of the binomial distribution with $n \to \infty$ and $p \to 0$, while their product is held constant at $\lambda = np$, yields the **Poisson distribution**.

The probability function is

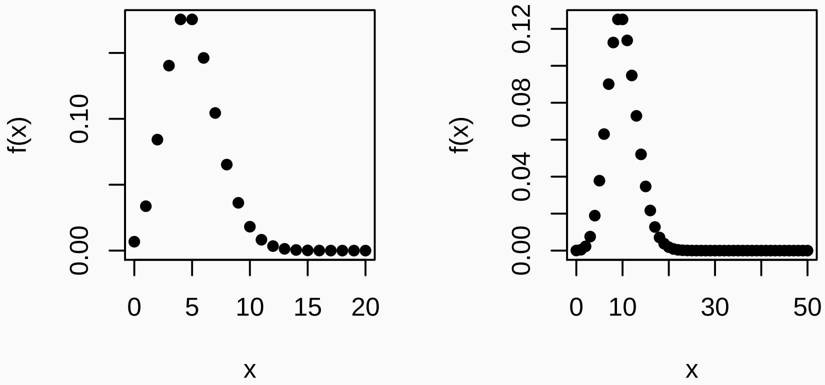
$$\Pr(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}\,, \qquad x = 0, 1, 2, \ldots .$$

with $\lambda > 0$.

The notation is $X \sim \mathcal{P}(\lambda)$, and $E(X) = \mathrm{var}(X) = \lambda$.

R lab: the Poisson distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dpois(0:20, 5), xlab = "x", ylab = "f(x)")
plot(0:50, dpois(0:50, 10), xlab ="x", ylab = "f(x)")
```



Negative binomial distribution

Let us consider a sequence of independent Bernoulli trials with success probability p, let X be the count of trials necessary to observe the r-th success. Then X has a **Negative binomial** (or Pascal) distribution with parameters p and r.

The probability function is

$$\Pr(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad x = r, r+1, r+2, \ldots.$$

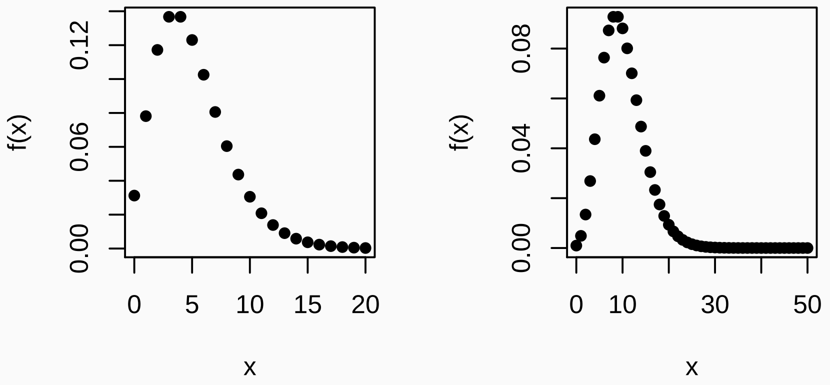
The notation is $X \sim \mathcal{NB}_i(p, r)$, and $E(X) = \frac{r}{p}$, $\mathrm{var}(X) = \frac{r(1-p)}{p^2}$.

It can also be defined with support the Natural numbers by simply considering the variable $Y = X - r$

The case for $r = 1$ is known as the **Geometric** distribution.

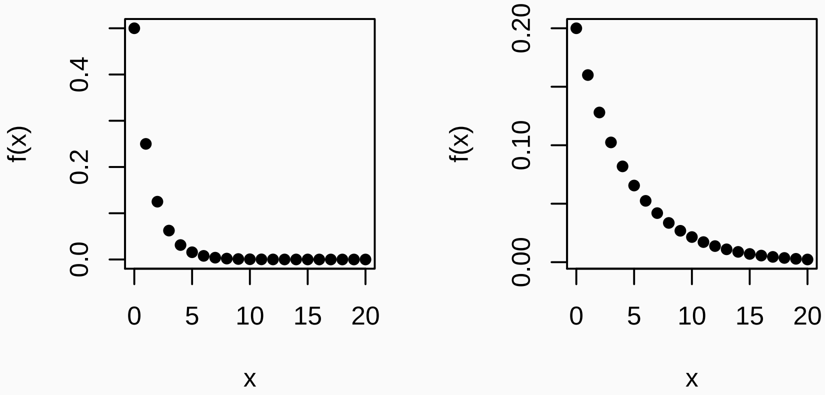
R lab: the Negative Binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dnbinom(0:20, 5, 0.5), xlab = "x", ylab = "f(x)")
plot(0:50, dnbinom(0:50, 10, 0.5), xlab ="x", ylab = "f(x)")
```



R lab: the Geometric distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dnbinom(0:20, 1, 0.5), xlab = "x", ylab = "f(x)")
plot(0:20, dnbinom(0:20, 1, 0.2), xlab ="x", ylab = "f(x)")
```



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

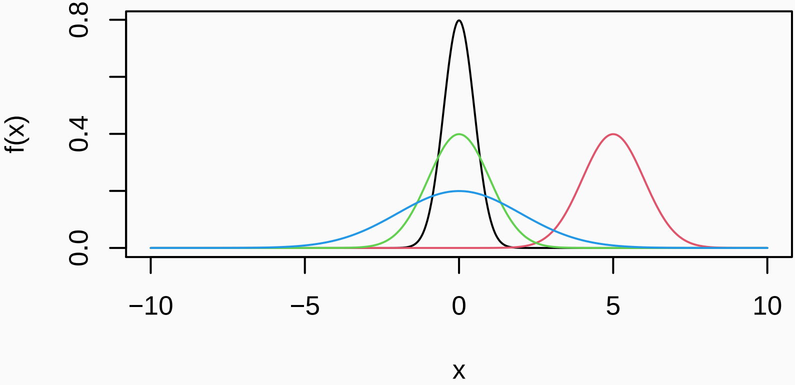
$$a + b\,X \sim \mathcal{N}(a + b\,\mu, b^2\,\sigma^2)\,,$$

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

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

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

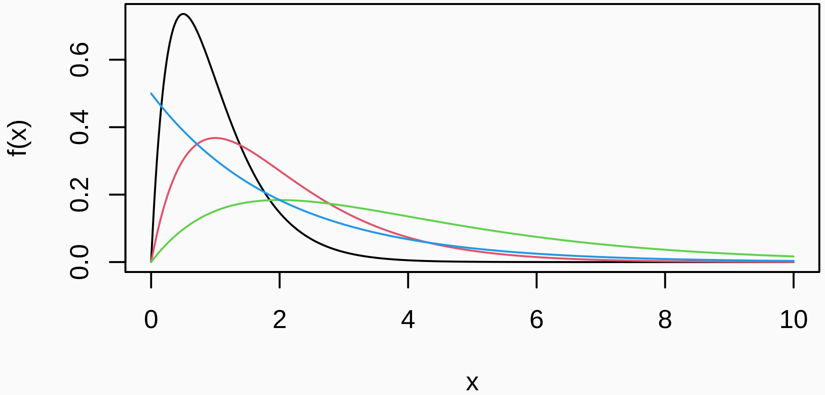
The notation is $X \sim Ga(\alpha, \lambda)$, $E(X) = \frac{\alpha}{\lambda}$ and $\mathrm{var}(X) = \frac{\alpha}{\lambda^2}$.

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

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

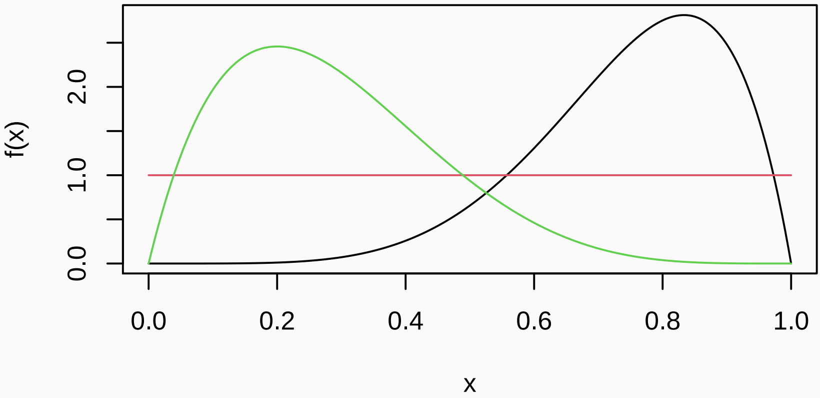
$\alpha, \beta > 0$

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**.

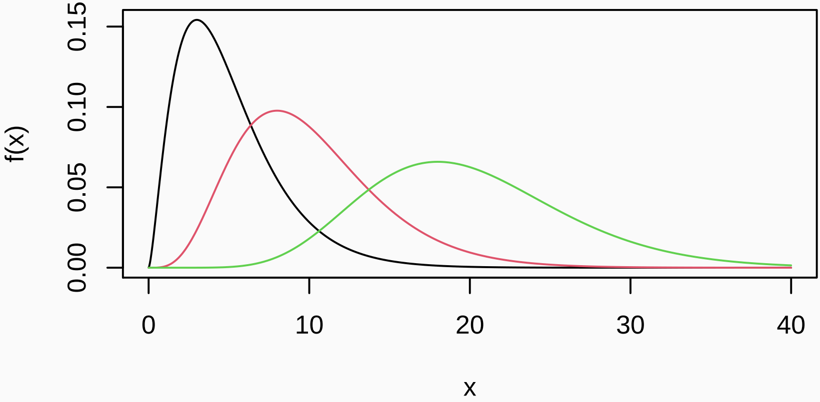
The notation is $X \sim \chi^2_k$, $E(X) = k$ and $\mathrm{var}(X) = 2k$.

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.

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

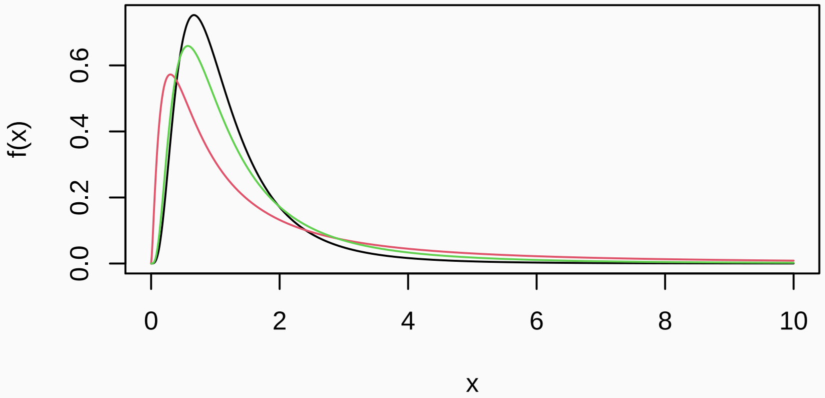
has an F **distribution with** n **and** m **degrees of freedom**.

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 $n > 2$.

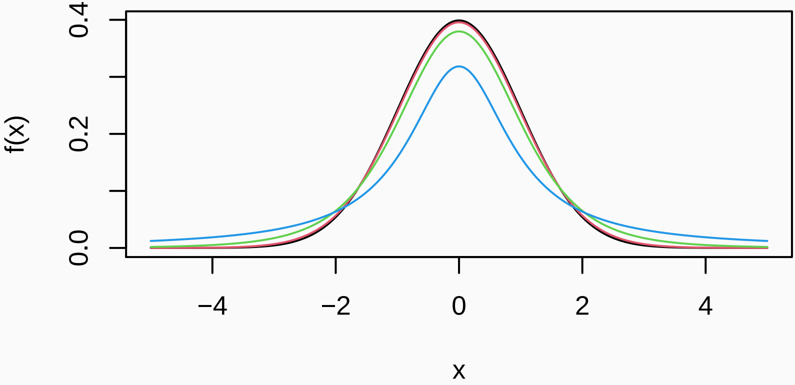
t_∞ is $\mathcal{N}(0, 1)$, while for n finite the distribution has heavier tails than the standard normal distribution.

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)
```



C.d.f. and quantile functions

3. Cumulative distribution functions

The **cumulative distribution function** (c.d.f.) of a r.v. X is the function $F(x)$ such that

$$F(x) = \mathrm{Pr}(X \leq x)\,,$$

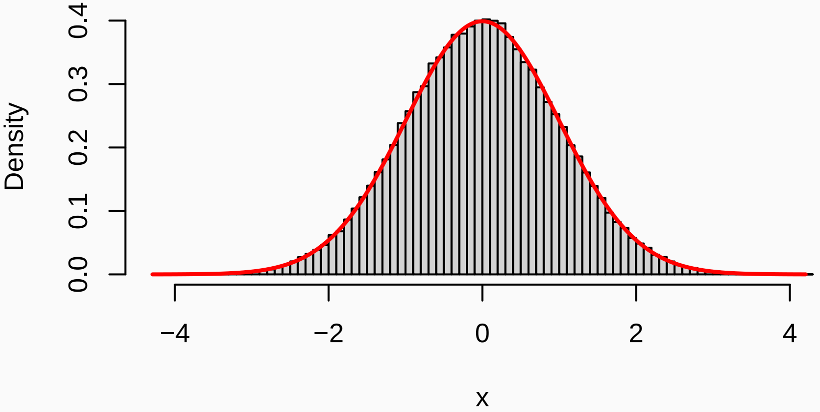
and it can be obtained from the probability function or the density function: the c.d.f. *identifies* the distribution.

From the definition of F it follows that $F(-\infty) = 0$, $F(\infty) = 1$, $F(x)$ is monotonic.

A useful property is that if F is a continuous function then $U = F(X)$ has a uniform distribution.

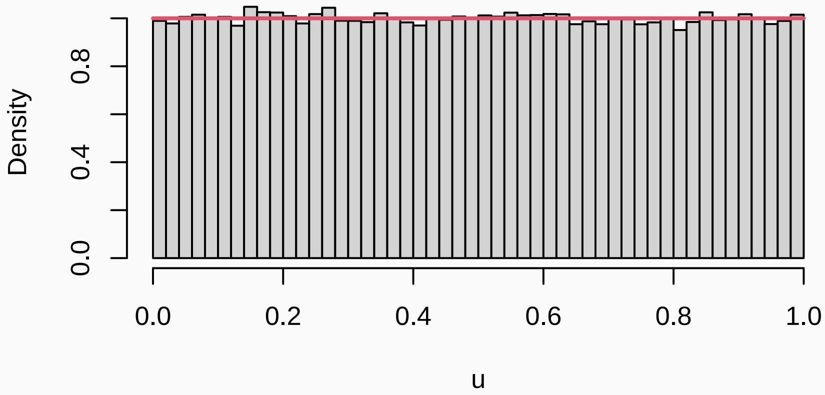
R lab: uniform transformation

```
x <- rnorm(10^5)    ### simulate values from N(0,1)
xx <- seq(min(x), max(x), l = 1000)
hist.scott(x, main = "") ### from MASS package
lines(xx, dnorm(xx), col = "red", lwd = 2)
```



R lab: uniform transformation (cont'd.)

```
u <- pnorm(x)    ### that's the uniform transformation
hist.scott(u, prob = TRUE, main="")
segments(0, 1, 1, 1, col = 2, lwd = 2)
```



The quantile function

The inverse of the c.d.f. is defined as

$$F^{-}(p) = \min\left(x | F(x) \geq p\right) , \qquad 0 \leq p \leq 1 \, .$$

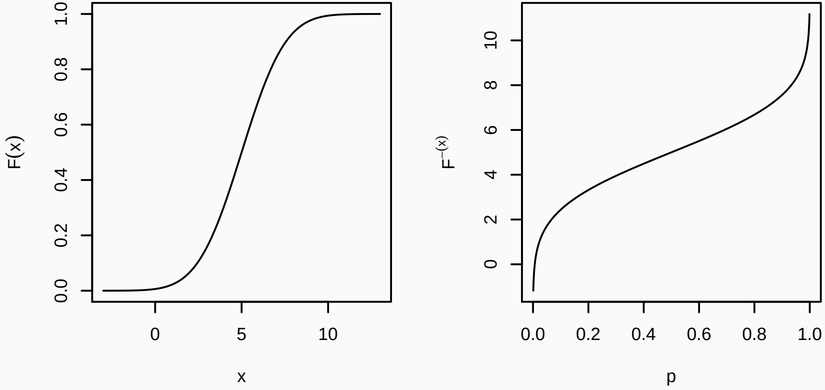
This is the usual inverse function of F when F is continuous.

Another useful property is that if $U \sim \mathcal{U}(0,1)$, namely it has a *uniform distribution* in [0,1], then the r.v. $X = F^{-}(U)$ has c.d.f. F.

This provides a simple method to generate random numbers from a distribution with known quantile function: it is the **inversion sampling method**, that only requires the ability to simulate from a uniform distribution.

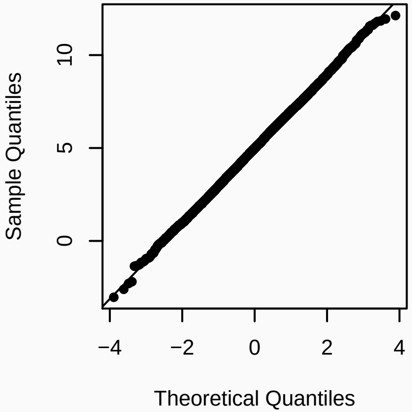
Example: normal cdf and quantile functions

Let us consider the case of $X \sim \mathcal{N}(5, 2^2)$, with c.d.f. and quantile functions given by `pnorm` and `qnorm`



R lab: inversion sampling

```
u <- runif(10^4); y <- qnorm(u, m = 5, s = 2)
par(pty = "s",  cex = 0.8)
qqnorm(y, pch = 16, main = "")
qqline(y)
```



Side note: quantile-quantile plot

The previous slide demonstrated the usage of the quantile function to build a tool for **model goodness-of-fit**.

The *quantile-quantile plot* visualizes the plausibility of a theoretical distribution for a set of observations $y = (y_1, \ldots, y_n)$.

This is done by comparing the quantile function of the assumed model with the sample quantiles, which are the points that lie on the inverse of the **empirical distribution function**

$$\widehat{F}_n(t) = \frac{\text{number of elements of } y \leq t}{n}.$$

If the agreement between the data and the theoretical distribution is good, the points on the plot would approximately lie on a line.