

Review of some probability concepts: random variables

(A quick tour)

R. Bellio & N. Torelli

Spring 2018

University of Udine & 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.

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

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

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

Notable discrete random variables

Discrete r.v. often used in applications:

- Binomial distribution
- Poisson distribution
- Negative binomial distribution
- Geometric distribution
- Hypergeometric distribution

The first two deserve some further attention.

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

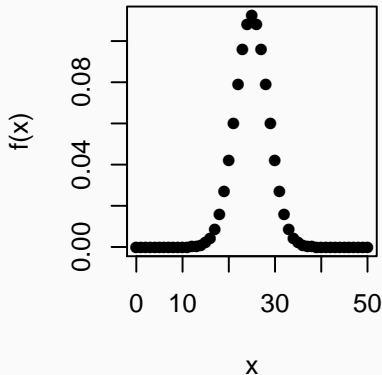
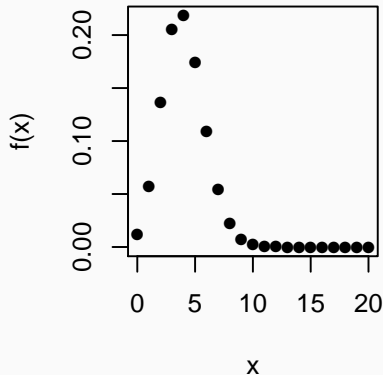
$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, \dots, n.$$

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

The case when $n = 1$ is known as **Bernoulli distribution**.

R lab: the binomial distribution

```
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 the binomial distribution with $n \rightarrow \infty$ and $p \rightarrow 0$, while their product is held constant at $\lambda = np$, yields the **Poisson distribution**.

The probability function is

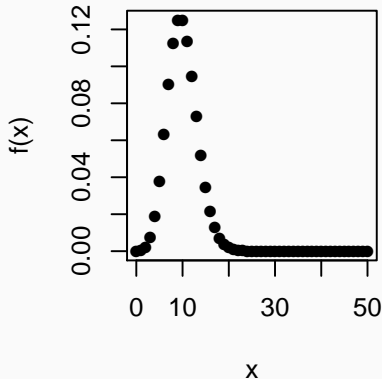
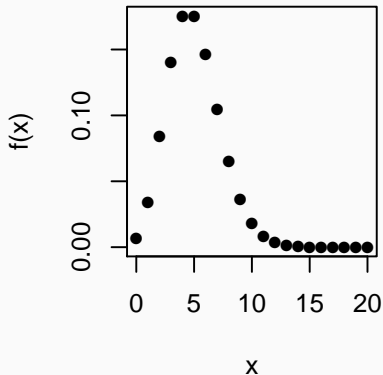
$$\Pr(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

with $\lambda > 0$.

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

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)")
```



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* $a + bX$, with a, b constants:

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

Notable continuous random variables

Important continuous distributions include:

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

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.

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

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

An important property is that for any constants a, b

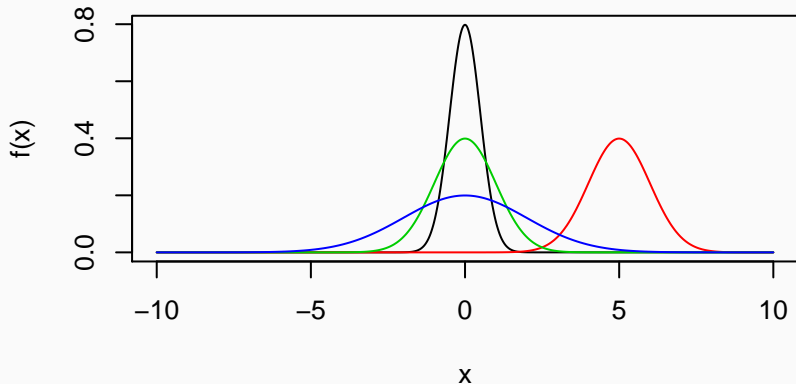
$$a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2),$$

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 χ^2 distribution

Let Z_1, \dots, 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**.

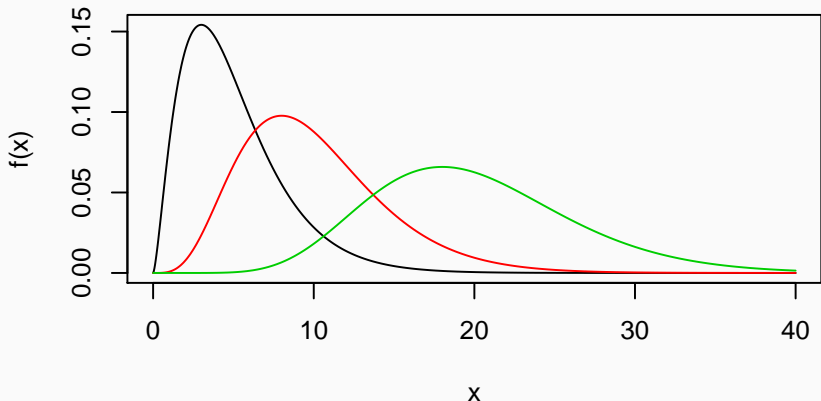
The notation is $X \sim \chi_k^2$, $E(X) = k$ and $\text{var}(X) = 2k$.

It is a special case of the Gamma distribution.

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 $X \sim \chi_n^2$ and $Y \sim \chi_m^2$, independent, then the r.v.

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

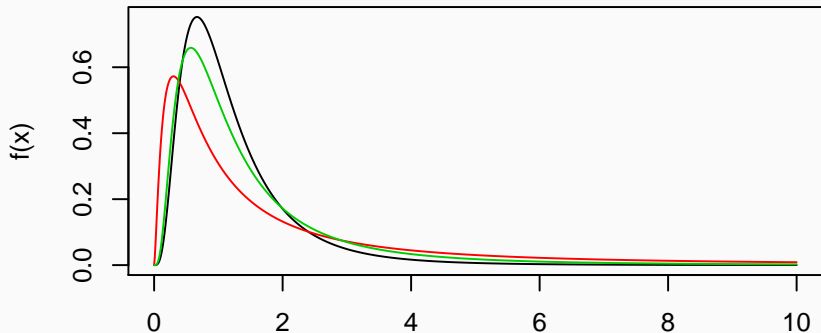
has an F **distribution with n and m degrees of freedom**.

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 $X \sim \chi_n^2$, 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 $\text{var}(T) = n/(n-2)$ provided that $n > 2$.

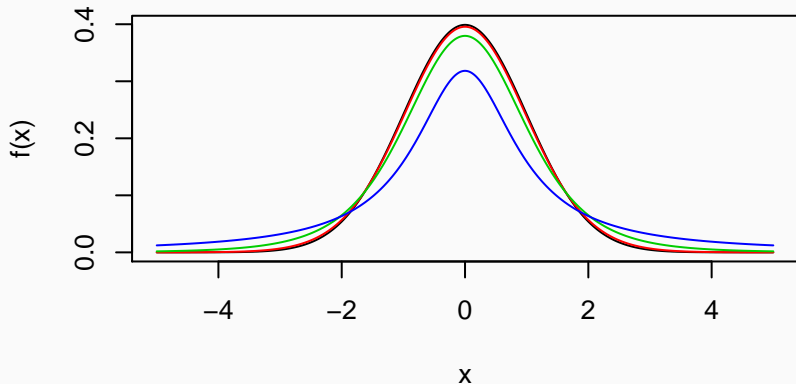
t_∞ is $\mathcal{N}(0, 1)$, while for n finite the distribution has heavier tails than the standard normal distribution.

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

$$F(x) = \Pr(X \leq x),$$

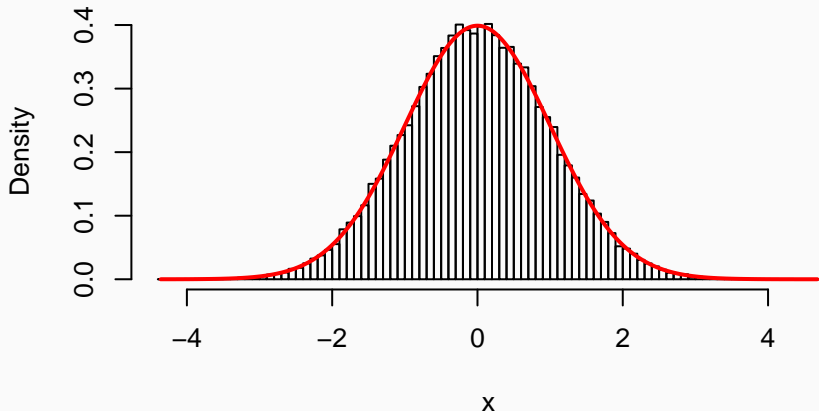
and it can be obtained from the probability function or the density function: the c.d.f. *identifies* the distribution.

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.

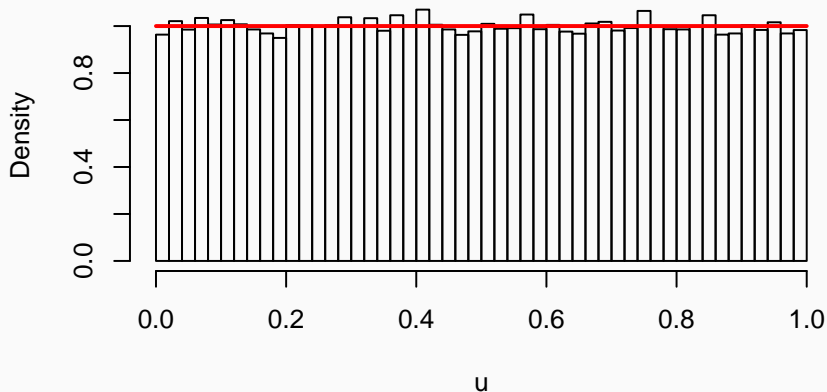
R lab: uniform transformation

```
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

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

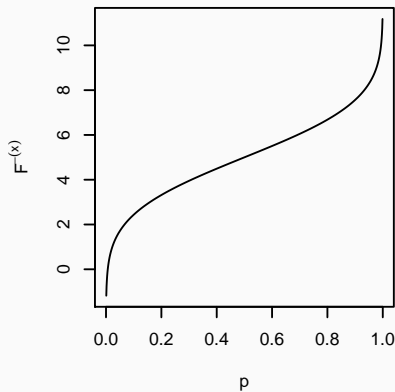
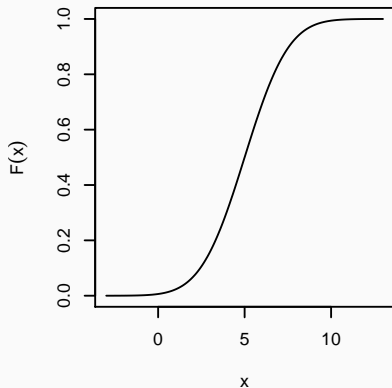
This is the usual inverse function of F when F is continuous.

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.

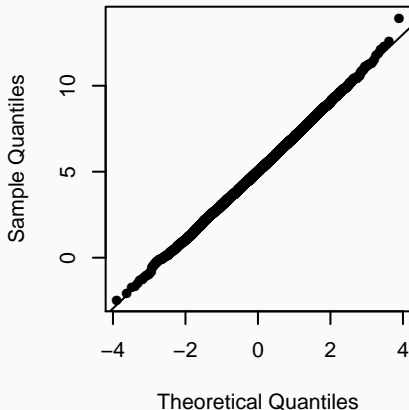
Example: normal cdf and quantile functions

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 $y = (y_1, \dots, y_n)$.

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**

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

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