Contents

1 Probability (Recap)

Suggested textbooks

- S.N. Wood. Core Statistics. Cambridge University Press, 2015.
- B. Efron, T. Hastie. Computer Age Statistical Inference – Algorithms, Evidence, and Data Science. Cambridge University Press, 201.

1.1 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. Cumulative distribution functions
3. Quantile functions

1.1.1 Discrete distributions

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

The **probability (mass) function** (p.m.f.) 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.

1.1.1.1 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

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

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

1.1.1.2 Notable discrete random variables

Discrete r.v. often used in applications:

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

The fist two deserve some further attention.

1.1.1.3 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

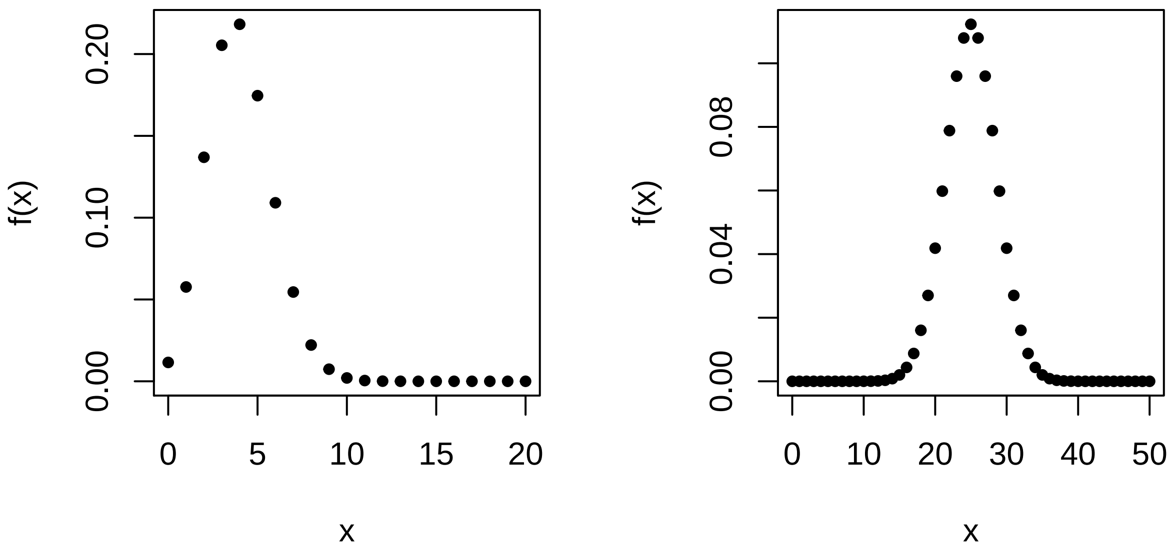
$$Pr(X = x) = \binom{n}{x} p^x (1p)^{nx}, \quad x = 0, \ldots, n.$$

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

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

R lab: the binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 20)
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)")
```



1.1.1.4 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

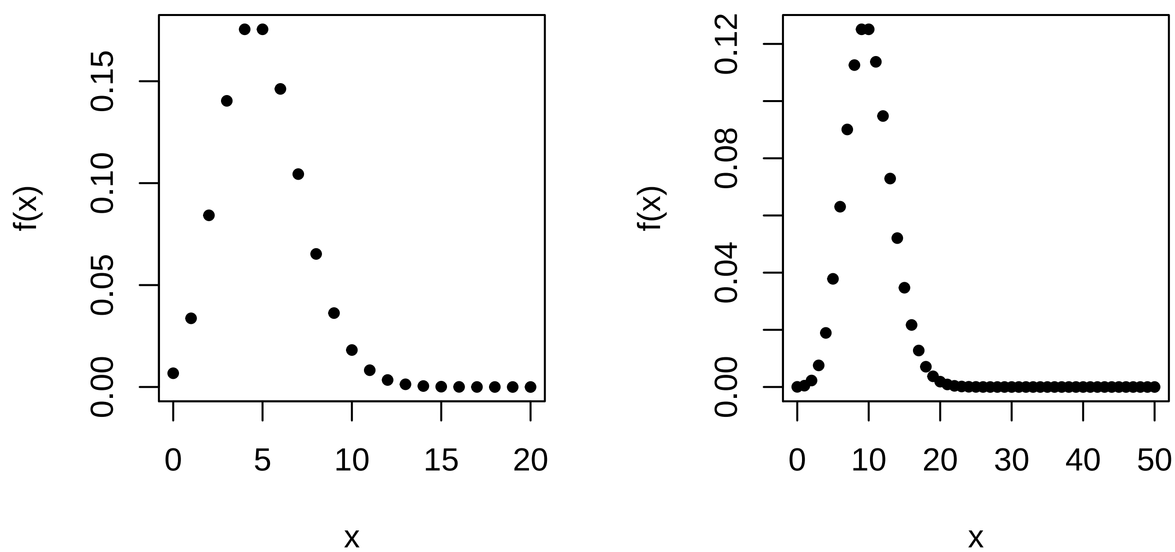
$$Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

with $\lambda > 0$.

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

R lab: the Poisson distribution

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



1.1.2 Continuous distributions

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.

1.1.2.1 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

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

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:

$$E(a+bX)=a+bE(X)$$

$$var(a+bX)=b^2var(X).$$

1.1.2.2 Notable continuous random variables

Important continuous distributions include:

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

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

1.1.2.3 The normal distribution

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

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

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

An important property is that for any constants a, b

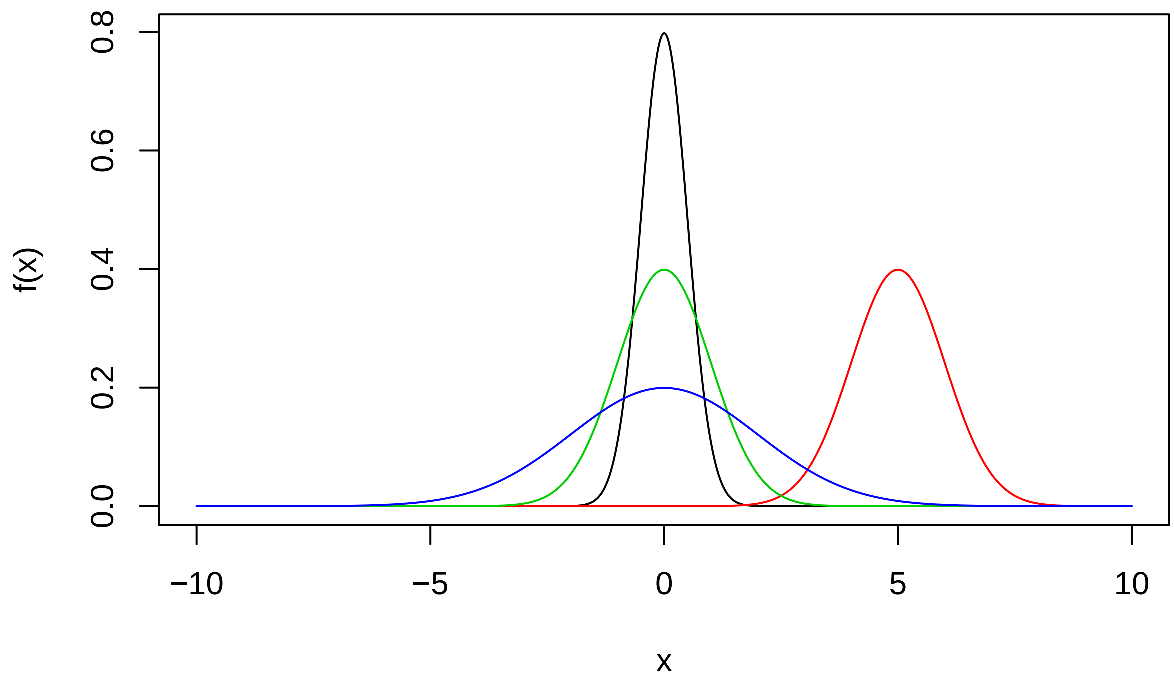
$$a+bX\sim \mathrm{N}(a+b\mu,b^2\sigma^2),$$

so that $Z=(X\mu)/\sigma\sim \mathrm{N}(0,1)$, 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)
```



1.1.3 C.d.f. and quantile functions

1.1.3.1 Cumulative distribution functions

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

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

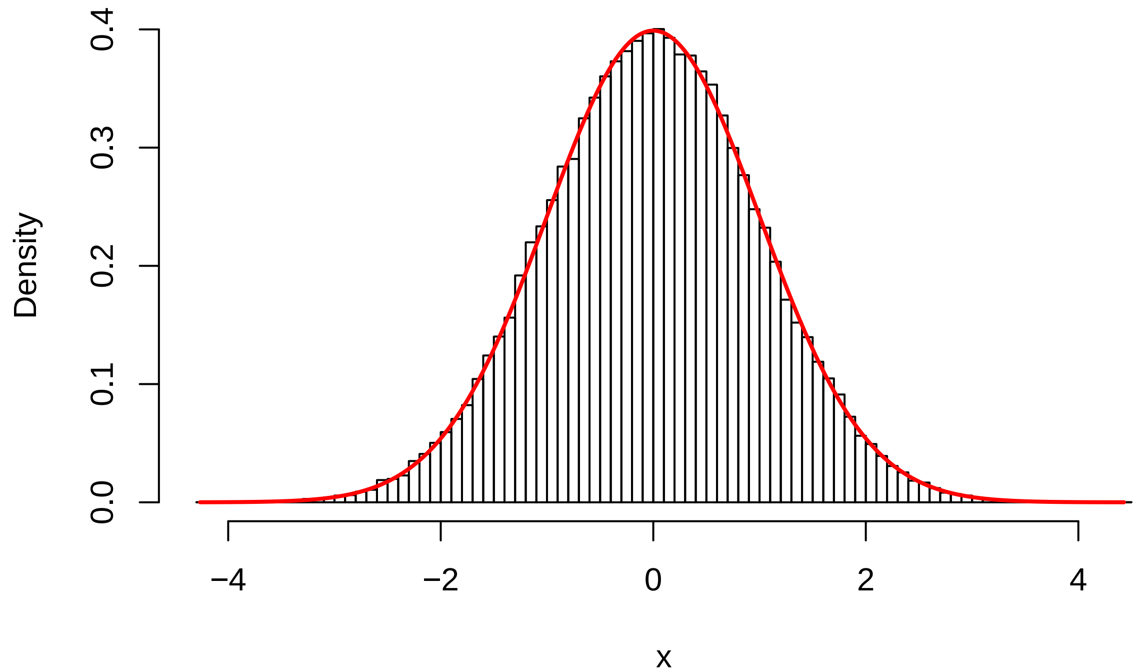
and it can be obtained from the probability function or the density function: the c.d.f. identifies the distribution.

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

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

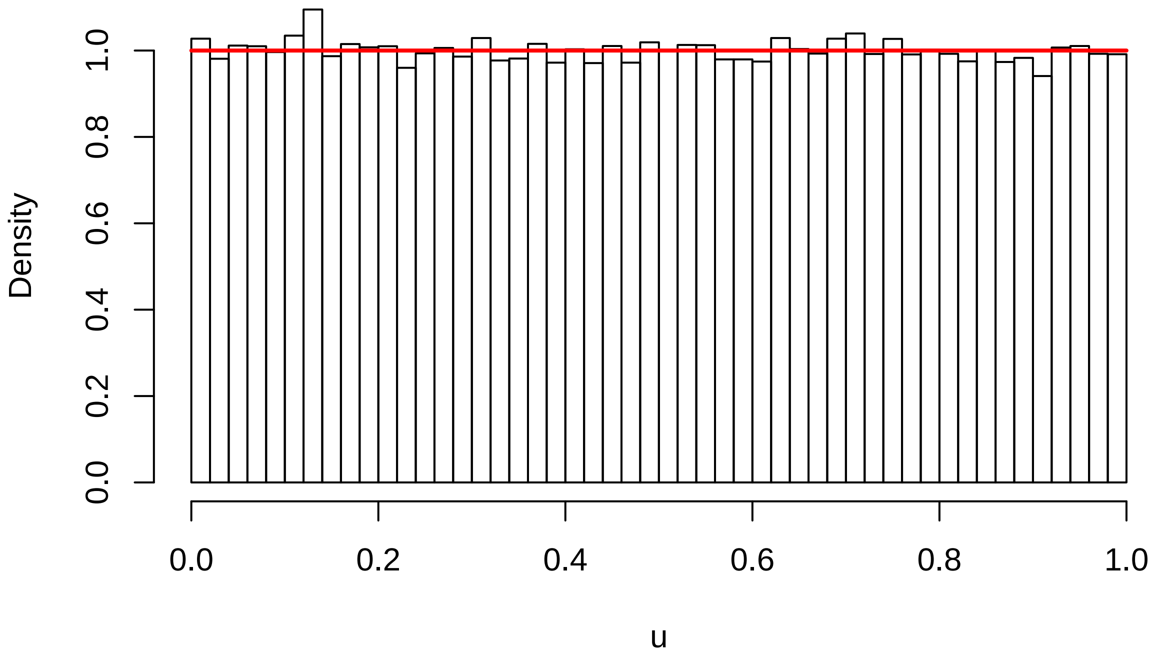
R lab: uniform transformation

```
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 cdf
hist.scott(u, main="")
segments(0, 1, 1, 1, col = 2, lwd = 2)
```



1.1.3.2 The quantile function

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

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

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

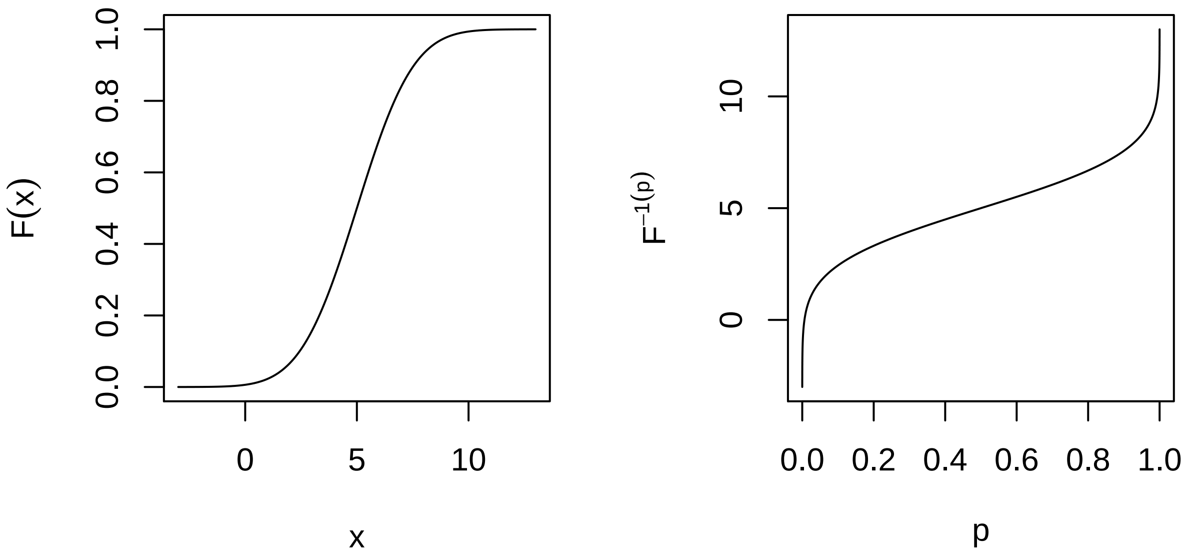
Another useful property is that if $U \sim U(0,1)$, namely it has a uniform distribution in $[0,1]$, then the r.v. $X = F^1(U)$ has c.d.f. F.

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

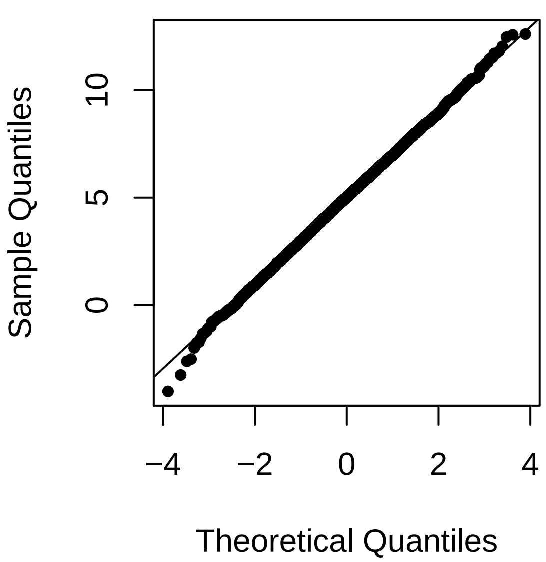
Let us consider the case of $X \sim N(5, 2^2)$, with c.d.f. and quantile functions given by pnorm and qnorm.

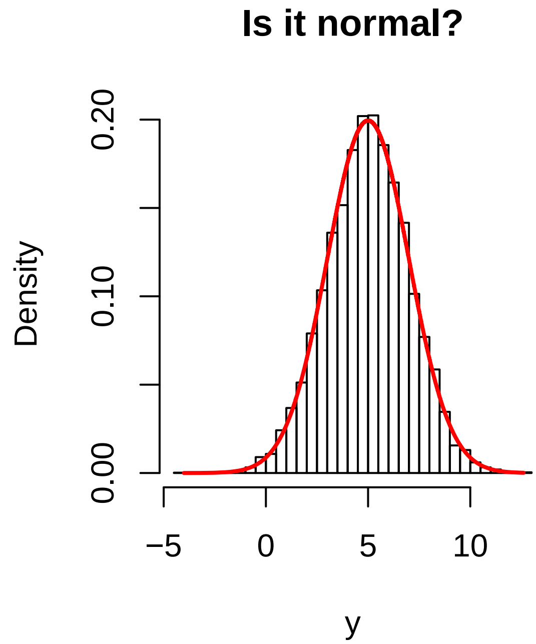
Make by exercise!



R lab: inversion sampling

```
u <- runif(10^4); y <- qnorm(u, m = 5, s = 2)
par(mfrow=c(1,2), pty = "s", pch=20)
qqnorm(y, main = "")
qqline(y)
## Now, trace the density function of y to check ...
```





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**

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

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