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 devices employed to build *models* of this variability.

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

TRASFORM AN EVENT → IN SOME PROBABILITY VALUE

WE DON'T KNOW THE VALUE THAT WILL OCCUR UNTIL IT HAPPEN

Distribution of a r.v.

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

1. Probability (mass) functions (pmf) FOR DESCRETE DISTRIBUTIONS
2. (Probability) density functions (pdf) FOR CONTINUES DISTRIBUTIONS
3. Cumulative distribution functions (cdf) FOR BOTH
4. Quantile functions FOR BOTH

Discrete distributions

1. Probability functions

PROBABILITY (MASS) FUNCTION = PROBABILITY

⚠ JUST FOR DISCRETE DISTRIBUTIONS

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.

SAMPLE SPACE OF DICE $S_X = \{1, \dots, 6\}$

$\Rightarrow f(x_i) = P(X = x) = \frac{1}{6} \quad \forall i \iff \sum_i f(x_i) = 1$

SUPPOSE WE HAVE $W = (X, Y)$ A JOINT PROBABILITY

THE $S_X = \{(1,1), (1,2), \ldots, (6,6)\}$; THE DOMAIN, OR SIZE, OF S_X IS $36 = 6 \cdot 6$

SUPPOSE WE HAVE $W = (X + Y)$

$S_X = \{2, 3, \ldots, 12\}$; DOMAIN OF S_X IS 11

HOW DO I COUNT IT?

LET'S DO IT WITH R : CODELECTURE04

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

=> THE VALUE WE EXPECT MORE

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

and the definition is extended to any function g of X

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

$g(x) = X^2 \Rightarrow E(X^2) \overset{i}{=} \sum_i x_i^2 \, f(x_i)^2$

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 SPREAD OF THE DISTRIBUTION

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

=> PUT THE VAR IN THE SAME SHAPE AS THE X_i AND NOT SQUARED

WRITE A PROBABILITY FUNCTION WHERE:

X		$f(x)$
-1 :	EARLYER	0.05
0 :	IN TIME	0.8
1 :	LATE	0.15

$$E(X) = -1 \cdot 0.05 + 0 \cdot 0.8 + 1 \cdot 0.15 \quad \text{(MEAN)}$$

$$E(X^2) = (-1)^2 \cdot 0.05 + 0 \cdot 0.8 + (1)^2 \cdot 0.15 \quad \text{(2° MOMENT)}$$

$$V(X) = E(X^2) - E(X)^2$$

RIGUARDARE PASSAGGIO

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 $E(X) = n\,p$, $\mathrm{var}(X) = n\,p\,(1-p)$.

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

SCALED BINOMIAL DISTRIBUTION => BINOMIAL/n

R lab: the binomial distribution

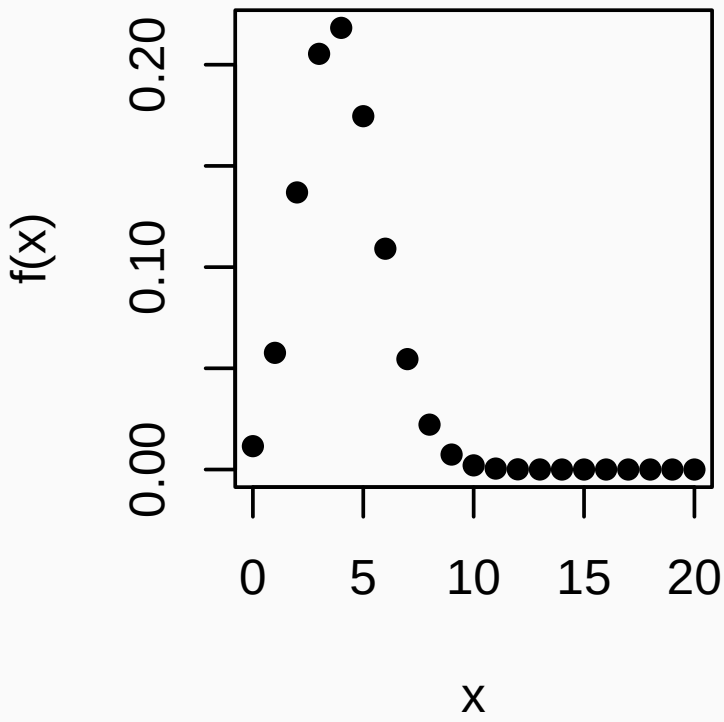
```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:20, dbinom(0:20, 20, 0.2), xlab = "x", ylab = "f(x)")
plot(0:50, dbinom(0:50, 50, 0.5), xlab ="x", ylab = "f(x)")
```

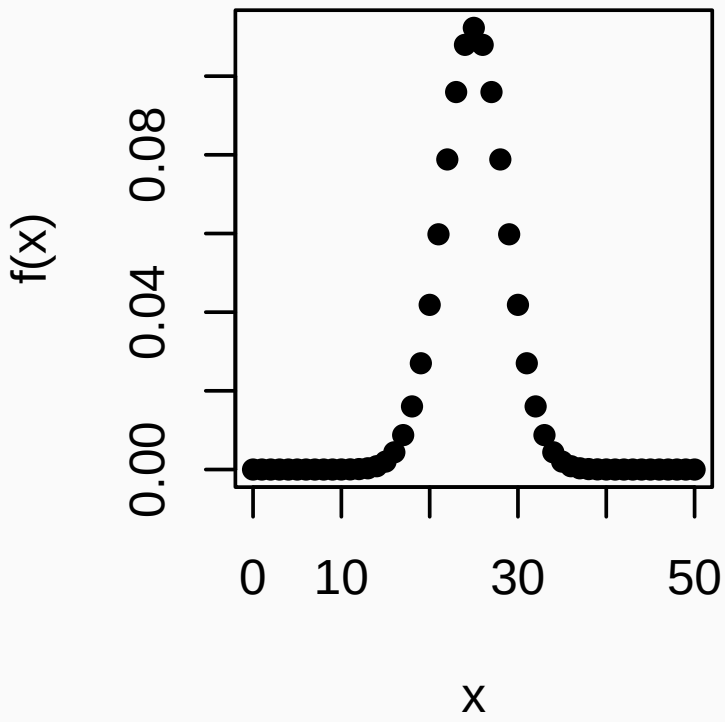
$E(X) = 20 \cdot (0.2)$

$E(X) = 50 \cdot (0.5)$

IF $n \to \infty$

THEN, IT CONVERGES TO A NORMAL DISTRIBUCTION





The Poisson distribution

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

Used for counts of events that occur randomly over time when: (1) counts of events in disjoint periods are independent, (2) it is essentially impossible to have two or more events simultaneously, (3) the rate of occurrence is constant.

The probability function is

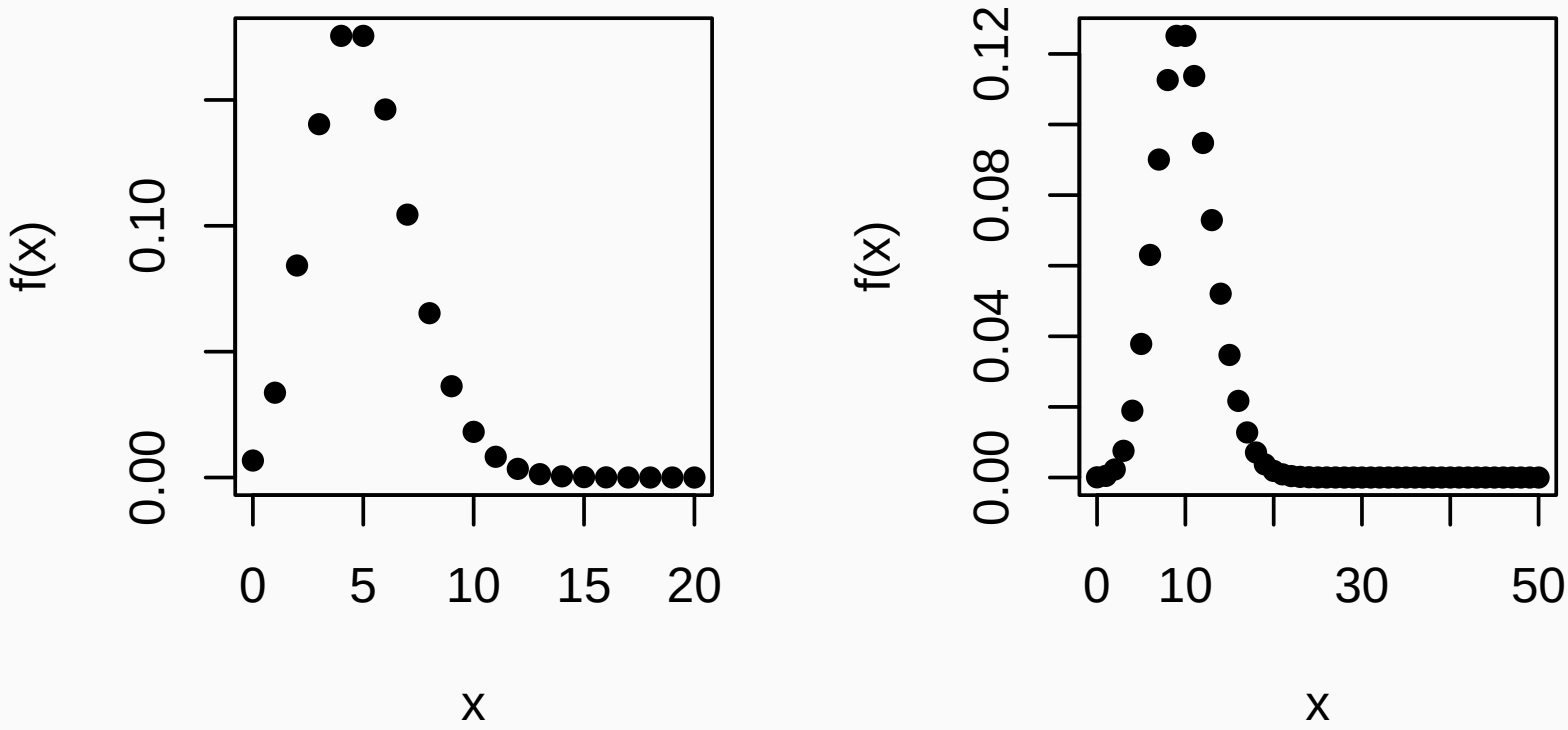
$$\Pr(X = x) = \frac{e^{-\lambda}\,\lambda^x}{x!}\,, \qquad x = 0, 1, 2, \ldots.$$

with $\lambda > 0$.

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

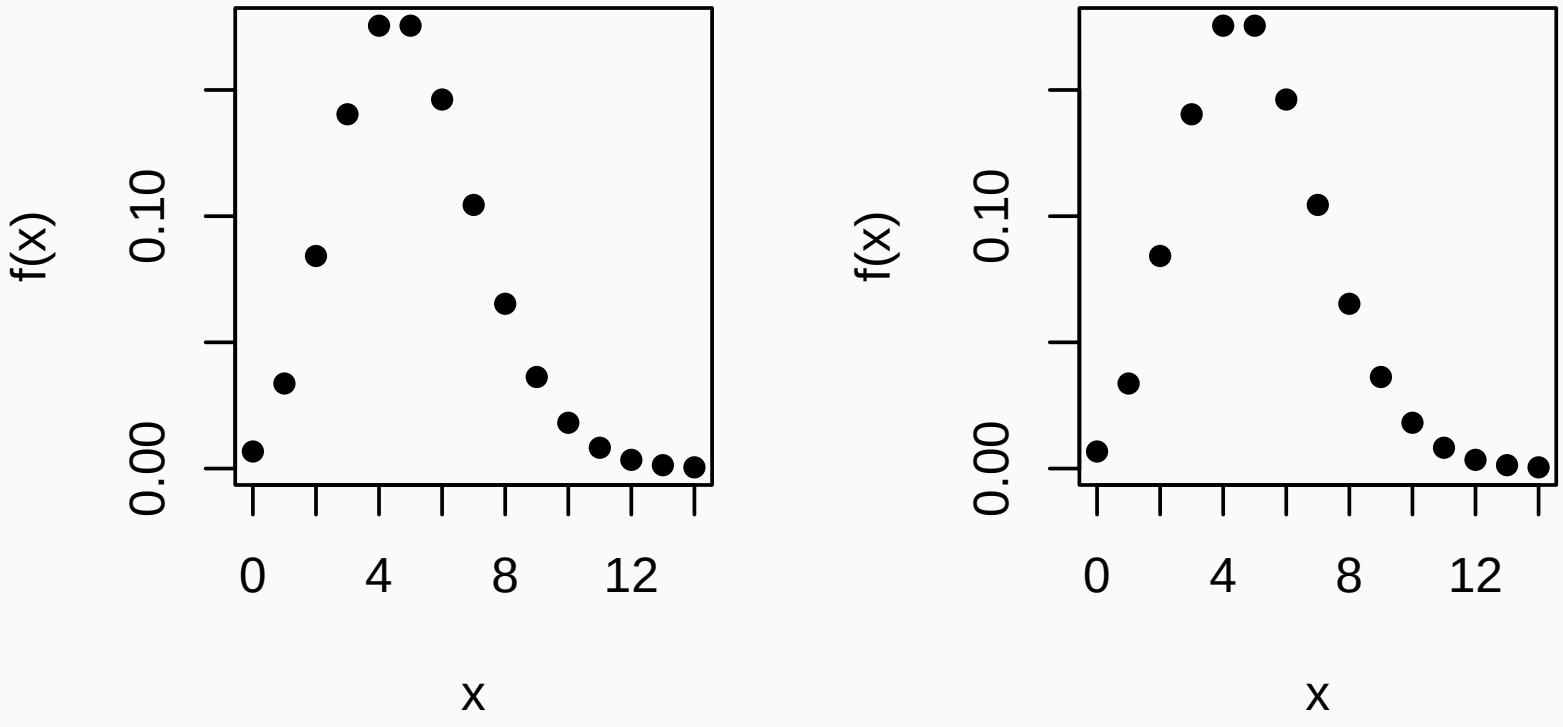
R lab: the Poisson distribution

```
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)")
```



R lab: Poisson distribution and Binomial distribution

```
par(mfrow=c(1,2), pty="s", pch = 16)
plot(0:14, dpois(0:14, 5), xlab = "x", ylab = "f(x)")
plot(0:14, dbinom(0:14, 50000000, 0.0000001),
     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 $E(X) = \frac{r}{p}$, $\operatorname{var}(X) = \frac{r(1-p)}{p^2}$.

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

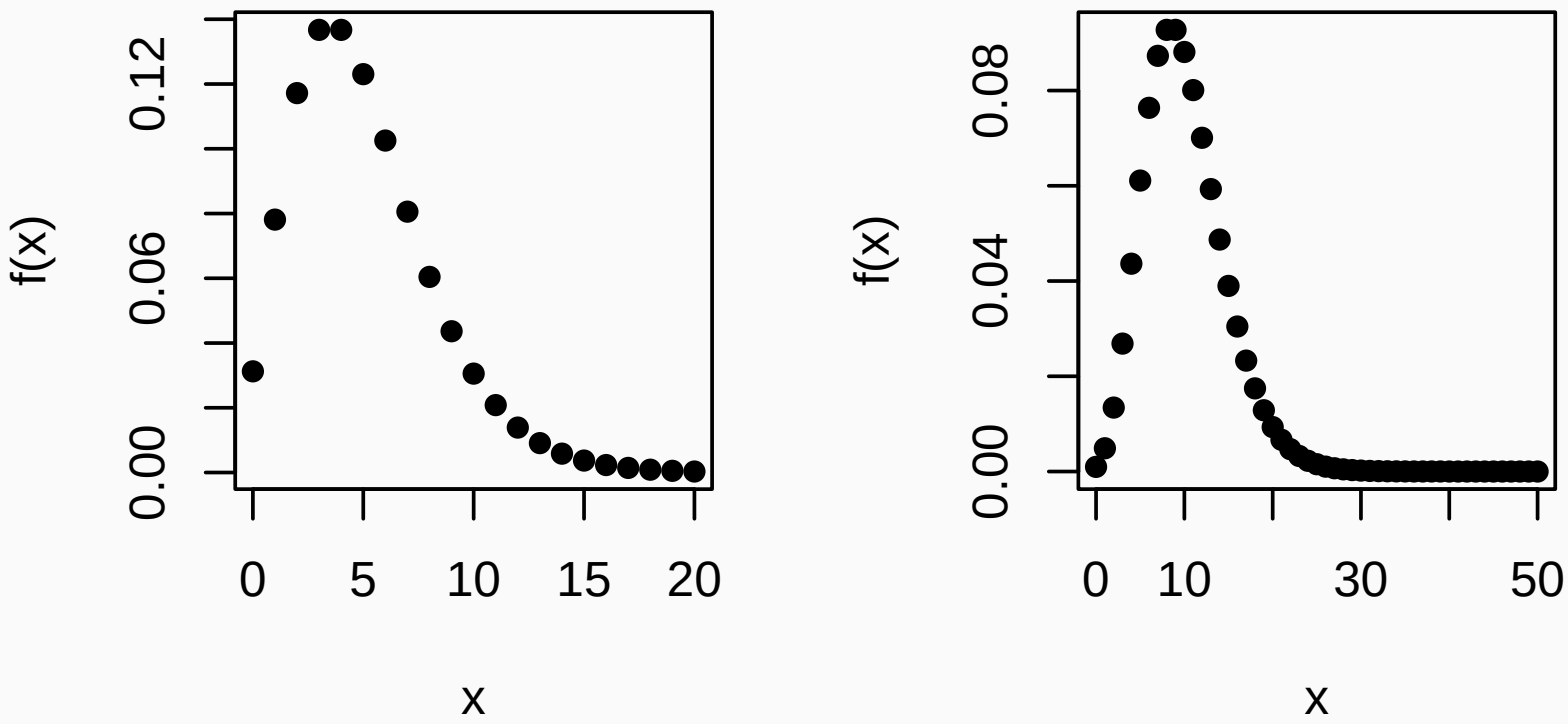
The case for $r = 1$ is known as the **Geometric** distribution.

IF $Y = X - r \Rightarrow X = Y + r$

$P(Y = y) = \binom{y+r-1}{r-1} p^r (1-p)^y$

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)")
```