



**AALBORG UNIVERSITET**

---

# Exam notes

4th semester: *Probability theory*

Kasper Rosenkrands

# Contents

<b>1</b>	<b>Basics of probability (incl. combinatorics, law of total probability and Bayes' formula).</b>	<b>1</b>
<b>2</b>	<b>Discrete stochastic variables and distributions (incl. means and variances).</b>	<b>2</b>
<b>3</b>	<b>Continuous stochastic variables and distributions (incl. means and variances).</b>	<b>4</b>
<b>4</b>	<b>Two random variables: select from topics such as joint distribution, conditional distribution, independence and convolution.</b>	<b>7</b>
4.1	Convolution . . . . .	8
<b>5</b>	<b>Two random variables: select from topics such as covariance and correlation, conditional expectation, conditional variance and the bivariate normal distribution.</b>	<b>9</b>
5.1	Conditional Expectation as a Random Variable . . . . .	9
5.2	Conditional Variance . . . . .	10
5.3	Covariance and Correlation . . . . .	10
5.4	The Correlation Coefficient . . . . .	11
<b>6</b>	<b>Generating functions (possibly with a focus on how probability generating functions relate to thinning of a Poisson process).</b>	<b>13</b>
6.1	The Probability Generating Function . . . . .	13
6.2	The Moment Generating Function . . . . .	14
6.3	The Poisson Process . . . . .	15
6.4	Thinning and Superposition . . . . .	16
<b>7</b>	<b>Limit theorems.</b>	<b>18</b>
7.1	The Law of Large Numbers . . . . .	18
7.2	Continuous Limits . . . . .	19
<b>8</b>	<b>Markov chains.</b>	<b>20</b>
8.1	Discrete-Time Markov Chains . . . . .	20
8.1.1	Time Dynamics of a Markov Chain . . . . .	20
8.1.2	Classification of States . . . . .	21

8.1.3 Stationary Distribution . . . . .	22
<b>9 Stochastic simulation.</b>	<b>23</b>

# 1 Basics of probability (incl. combinatorics, law of total probability and Bayes' formula).

**Theorem 1.1 (Law of Total Probability).** Let  $B_1, B_2, \dots$  be a sequence of events such that

- (a)  $P(B_k) > 0$  for  $k = 1, 2, \dots$
- (b)  $B_i$  og  $B_j$  are disjoint whenever  $i \neq j$
- (c)  $S = \bigcup_{k=1}^{\infty} B_k$

Then, for any event  $A$ , we have

$$P(A) = \sum_{k=1}^{\infty} P(A|B_k)P(B_k).$$

*Proof.* First note that

$$A = A \cap S = \bigcup_{k=1}^{\infty} (A \cap B_k),$$

by the distributive law for infinite unions. Since  $A \cap B_1, A \cap B_2, \dots$  are pairwise disjoint, we get

$$P(A) = P(A \cap S) = \sum_{k=1}^{\infty} P(A \cap B_k) = \sum_{k=1}^{\infty} P(A|B_k)P(B_k).$$

Which proves the theorem. Note that the result also holds for finite sequences. ■

**Corollary 1.6.** If  $0 < P(B) < 1$ , for  $B \subseteq S$ , then

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

**Proposition 1.11 (Bayes' Formula).** Under the same assumptions as in the law of total probability and if  $P(A) > 0$ , then for any event  $B_j$ , we have

$$\begin{aligned} P(B_j|A) &= \frac{P(A|B_j)P(B_j)}{\sum_{k=1}^{\infty} P(A|B_k)P(B_k)} \\ &= \frac{P(A|B_j)P(B_j)}{P(A)} \quad \text{according to the LTP.} \end{aligned}$$

## 2 Discrete stochastic variables and distributions (incl. means and variances).

**Definition 2.1.** A random variable is a real random variable that gets its values from a random experiment.

$$X : S \rightarrow \mathbb{R}.$$

**Definition 2.2.** If the range of  $X$  is countable, then  $X$  is called a *discrete random variable*.

**Definition 2.3.** Let  $X$  be a discrete random variable with range  $\{x_1, x_2, \dots\}$  (finite or countably infinite). The function

$$p(x_k) = P(X = x_k), \quad k = 1, 2, \dots$$

is called the *probability mass function* (pmf) of  $X$ .

**Proposition 2.1.** A function  $p$  is a possible pmf of a discrete random variable on the range  $\{1, 2, \dots\}$  if and only if

(a)  $p(x_k) \geq 0$  for  $k = 1, 2, \dots$

(b)  $\sum_{k=1}^{\infty} p(x_k) = 1$

**Definition 2.4.** Let  $X$  be any random variable. The function

$$F(x) = P(X \leq x), \quad \text{for } x \in \mathbb{R},$$

is called the (*cumulative*) *distribution function* (cdf) of  $X$ .

**Definition 2.8.** Let  $X$  be a discrete random variable with range  $\{x_1, x_2, \dots\}$  (finite or countably infinite) and probability mass function  $p$ . The *expected value* of  $X$  is defined as

$$E[X] = \sum_{k=1}^{\infty} x_k p(x_k).$$

**Proposition 2.9.** Let  $X$  be a discrete random variable with range  $\{0, 1, \dots\}$ . Then

$$E[X] = \sum_{n=0}^{\infty} P(X > n)$$

*Proof.* Note that  $k = \sum_{n=1}^k 1$ , and use the definition of expected value to obtain

$$\begin{aligned} E[X] &= \sum_{k=1}^{\infty} kP(X = k) = \sum_{k=1}^{\infty} \sum_{n=1}^k P(X = k) \\ &= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(X = k) = \sum_{n=1}^{\infty} P(X \geq n) \\ &= \sum_{n=0}^{\infty} P(X > n) \end{aligned}$$

■

**Proposition 2.11 (Linearity for the Expectation).** Let  $X$  be any random variable, and let  $a$  and  $b$  be real numbers. Then

$$E[aX + b] = aE[X] + b$$

*Proof.* We prove this in the discrete case, for  $a > 0$ . Let  $Y = aX + b$ , and note that  $Y$  is a discrete random variable and by definition 2.8 expected value

$$\begin{aligned} E[Y] &= \sum_{k=1}^{\infty} y_k p_Y(y_k) = \sum_{k=1}^{\infty} y_k p_X\left(\frac{y_k - b}{a}\right) \\ &= \sum_{k=1}^{\infty} (ax_k + b) p_X(x_k) \\ &= a \sum_{k=1}^{\infty} x_k p_X(x_k) + b \sum_{k=1}^{\infty} p_X(x_k) \\ &= aE[X] + b \end{aligned}$$

and we are done. **Måske skal man bruge 2.12 her (side 102).**

■

### 3 Continuous stochastic variables and distributions (incl. means and variances).

**Definition 2.5.** If the cdf  $F$  is a continuous and differentiable function, then  $X$  is said to be a *continuous random variable*.

**Proposition 2.3.** If  $F$  is the cdf of any random variable,  $F$  has the following properties:

- (a) It is nondecreasing
- (b) It is right-continuous
- (c) It has the limits  $F(-\infty) = 0$  and  $F(\infty) = 1$  (where the limits may or may not be attained at finite  $x$ ).

**Proposition 2.4.** Let  $X$  be any random variable with cdf  $F$ . Then

- (a)  $P(a < X \leq b) = F(b) - F(a), \quad a \leq b$
- (b)  $P(X > x) = 1 - F(x), \quad x \in \mathbb{R}$
- (c)  $F$  is continuous from the right with limit from the left and increasing

**Definition 2.6.** The function  $f(x) = F'(x)$  is called the *probability density function* (pdf) of  $X$ .

**Proposition 2.5.** Let  $X$  be a continuous random variable with pdf  $f$  and cdf  $F$ . Then

- (a)  $F(x) = \int_{-\infty}^x f(t)dt, \quad x \in \mathbb{R}$
- (b)  $f(x) = F'(x), \quad x \in \mathbb{R}$
- (c) For  $B \subseteq \mathbb{R}, \quad P(X \in B) = \int_B f(x)dx$

**Proposition 2.6.** A function  $f$  is a possible pdf of some continuous random variable if and only if

- (a)  $f(x) \geq 0, \quad x \in \mathbb{R}$
- (b)  $\int_{-\infty}^{\infty} f(x)dx = 1$

**Proposition 2.8.** Let  $X$  be a continuous random variable with pdf  $f_X$ , let  $g$  be a strictly increasing or strictly decreasing, differentiable function, and let  $Y = g(X)$ . Then  $Y$  has pdf

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y))$$

for  $y$  in range of  $Y$ .

**Definition 2.9.** Let  $X$  be a continuous random variable with pdf  $f$ . The *expected value* of  $X$  is defined as

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{\mathbb{R}} xf(x)dx,$$

notice that the last equality is not always satisfied, but for the purpose of this course it is.

**Proposition 2.10.** Let  $X$  be a continuous random variable with range  $[0, \infty)$ . Then

$$E[x] = \int_0^{\infty} P(X > x)dx$$

**Proposition 2.11 (Linearity for the Expectation).** Let  $X$  be any random variable, and let  $a$  and  $b$  be real numbers. Then

$$E[aX + b] = aE[X] + b$$

*Proof.* We prove this in the continuous case, for  $a > 0$ . Let  $Y = aX + b$ , and note that  $Y$  is a continuous random variable, which by proposition 2.8 has pdf

$$f_Y(y) = \frac{1}{a}f_X\left(\frac{y-b}{a}\right)$$

and by definition, the expected value of  $Y$  is

$$E[Y] = \int_{-\infty}^{\infty} yf_Y(y)dy = \frac{1}{a} \int_{-\infty}^{\infty} yf_X\left(\frac{y-b}{a}\right)dy$$

where the variable  $y = ax + b$  gives  $dy = a \, dx$  and hence

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} (ax + b)f_X(x)dx \\ &= a \int_{-\infty}^{\infty} xf_X(x)dx + b \int_{-\infty}^{\infty} f_X(x)dx \\ &= aE[X] + b \end{aligned}$$

and we are done. ■

**Definition 2.10.** Let  $X$  be a random variable with expected value  $\mu$ . The *variance* of  $X$  is defined as

$$\text{Var}[X] = E[(X - \mu)^2]$$

**Definition 2.11.** Let  $X$  be a random variable with variance  $\sigma^2 = \text{Var}[X]$ . The *standard deviation* of  $X$  is then defined as  $\sigma = \sqrt{\text{Var}[X]}$ .



**Corollary 2.2.**

$$\text{Var}[X] = E[X^2] - (E[X])^2$$

*Proof.* By proposition 2.12, we have

$$\begin{aligned} \text{Var}[X] &= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 dx \\ &= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2(E[X])^2 + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

■

**Proposition 2.14 (Chebyshev's Inequality).** Let  $X$  be any random variable with mean  $\mu$  and variance  $\sigma^2$ . For any constant  $c > 0$ , we have

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

*Proof.* Let us prove the continuous case. Fix  $c$  and let  $B$  be the set  $\{x \in \mathbb{R} : |x - \mu| \geq c\sigma\}$ . We get

$$\begin{aligned} \sigma^2 &= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx \\ &\geq \int_B (x - \mu)^2 f(x) dx \geq c^2 \sigma^2 \int_B f(x) dx = c^2 \sigma^2 P(X \in B). \end{aligned}$$

Which gives the desired inequality.

■

## 4 Two random variables: select from topics such as joint distribution, conditional distribution, independence and convolution.

**Definition 3.1.** Let  $X$  and  $Y$  be random variables. The pair  $(X, Y)$  is then called a (two-dimensional) *random vector*.

**Definition 3.2.** The *joint distribution function* (joint cdf) of  $(X, Y)$  is defined as

$$F(x, y) = P(X \leq x, Y \leq y)$$

for  $x, y \in \mathbb{R}$ .

**Proposition 3.1 (Marginal cdf).** If  $(X, Y)$  has joint cdf  $F$ , then  $X$  and  $Y$  have cdfs

$$F_X(x) = F(x, \infty) \quad \text{and} \quad F_Y(y) = F(\infty, y)$$

for  $x, y \in \mathbb{R}$ . Notice that there is a slight abuse of notation here,  $F(x, \infty)$  refers to  $\lim_{y \rightarrow \infty} F(x, y)$ .

**Definition 3.5.** If there exists a function  $f$  such that

$$P((X, Y) \in B) = \int \int_B f(x, y) dx dy$$

for all subsets  $B \subseteq \mathbb{R}^2$ , then  $X$  and  $Y$  are said to be *jointly continuous*. The function  $f$  is called the *joint pdf*.

**Proposition 3.3.** If  $X$  and  $Y$  are jointly continuous with joint cdf  $F$  and joint pdf  $f$ , then

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y), \quad x, y \in \mathbb{R}.$$

**Proposition 3.4.** A function  $f$  is a possible joint pdf for the random variables  $X$  and  $Y$  if and only if

(a)  $f(x, y) \geq 0$  for all  $x, y \in \mathbb{R}$

(b)  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

**Proposition 3.5.** Suppose that  $X$  and  $Y$  are jointly continuous with joint pdf  $f$ . Then  $X$  and  $Y$  are continuous random variables with marginal pdfs

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad x \in \mathbb{R}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx, \quad y \in \mathbb{R}.$$

**Definition 3.7.** Let  $(X, Y)$  be jointly continuous with joint pdf  $f$ . The *conditional pdf of  $Y$  given  $X = x$*  is defined as

$$f_Y(y|x) = \frac{f(x, y)}{f_X(x)}, \quad y \in \mathbb{R}.$$

The following proposition is a continuous version of the law of total probability.

**Proposition 3.6.** Let  $X$  and  $Y$  be jointly continuous. Then

(a)

$$f_Y(y) = \int_{-\infty}^{\infty} f_Y(y|x) f_X(x) dx, \quad y \in \mathbb{R}$$

(b)

$$P(Y \in B) = \int_{-\infty}^{\infty} P(Y \in B|X = x) f_X(x) dx, \quad B \subseteq \mathbb{R}.$$

*Proof.* For (a), just combine Proposition 3.5 with the definition of conditional pdf for (b), part (a) gives

$$\begin{aligned} P(Y \in B) &= \int_B f_Y(y) dy = \int_B \int_{-\infty}^{\infty} f_Y(y|x) f_X(x) dx \, dy \\ &= \int_{-\infty}^{\infty} \int_B f_Y(y|x) f_X(x) dy \, dx \\ &= \int_{-\infty}^{\infty} P(Y \in B|X = x) f_X(x) dx \end{aligned}$$

as desired. ■

**Proposition 3.10.** Suppose that  $X$  and  $Y$  are jointly continuous with joint pdf  $f$ . Then  $X$  and  $Y$  are independent if and only if

$$f(x, y) = f_X(x) f_Y(y)$$

for all  $x, y \in \mathbb{R}$ .

**Corollary 3.1.** *The random variables  $X$  and  $Y$  are independent if and only if “the joint is the product of the marginals.”*

## 4.1 Convolution

**Proposition 3.34.** Let  $X$  and  $Y$  be independent continuous random variables with pdfs  $f_X$  and  $f_Y$ , respectively. The pdf of the sum  $X + Y$  is then

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_Y(x - u) f_X(u) du, \quad x \in \mathbb{R}.$$

## 5 Two random variables: select from topics such as covariance and correlation, conditional expectation, conditional variance and the bivariate normal distribution.

**Definition 3.10.** Let  $y$  be random variable and  $B$  an event with  $P(B) > 0$ . The *conditional expectation* of  $Y$  given  $B$  is defined as

$$E[Y|B] = \begin{cases} \sum_{k=1}^{\infty} y_k P(Y = y_k|B) & \text{if } Y \text{ is discrete with range } \{y_1, y_2, \dots\} \\ \int_{-\infty}^{\infty} y f_Y(y|B) dy & \text{if } Y \text{ is continuous} \end{cases}$$

**Definition 3.11.** Suppose that  $X$  and  $Y$  are discrete. We define

$$E[Y|X = x_j] = \sum_{k=1}^{\infty} y_k p_Y(y_k|x_j).$$

**Definition 3.12.** Suppose that  $X$  and  $Y$  are jointly continuous. We define

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_Y(y|x) dx.$$

Following the usual intuitive interpretation, this is the expected value for  $Y$  if we know that  $X = x$ . The law of total expectation now takes the following form.

**S.** suppose that  $X$  and  $Y$  are jointly continuous. Then

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x] f_X(x) dx.$$

### 5.1 Conditional Expectation as a Random Variable

**Definition 3.13.** The *conditional expectation* of  $Y$  given  $X$ ,  $E[Y|X]$ , is a random variable that equals  $E[Y|X = x]$  whenever  $X = x$ .

**Corollary 3.5.**

$$E[Y] = E[E[Y|X]]$$

**Definition 3.14.** Let  $g(X)$  be a predictor of  $Y$ . The *mean square error* is defined as

$$E[(Y - g(X))^2].$$

**Proposition 3.18.** Among all predictors  $g(X)$  of  $Y$ , the mean square error is minimized by  $E[Y|X]$ .

We omit the proof and instead refer to an intuitive argument. Suppose that we want to predict  $Y$  as much as possible by a constant value  $c$ . Then, we want to minimize  $E[(Y - c)^2]$ , and with  $\mu = E[Y]$  we get

$$\begin{aligned} E[(Y - c)^2] &= E[(Y - \mu + \mu - c)^2] \\ &= E[(Y - \mu)^2] + 2(\mu - c)E[(Y - \mu)] + (\mu - c)^2 \\ &= \text{Var}[Y] + (\mu - c)^2 \end{aligned}$$

since  $E[Y - \mu] = 0$ . But the last expression is minimized when  $\mu = c$  and hence  $\mu$  is the best predictor of  $Y$  among all constants. This is not too surprising; if we do not know anything about  $Y$ , the best guess should be the expected value  $E[Y]$ . Now, if we observe another random variable  $X$ , the same ought to be true:  $Y$  is best predicted by its expected value given the random variable  $X$ , that is,  $E[Y|X]$ .

## 5.2 Conditional Variance

**Definition 3.15.** The *conditional variance* of  $Y$  given  $X$  is defined as

$$\text{Var}[Y|X] = E[(Y - E[Y|X])^2|X].$$

Note that the conditional variance is also a random variable and we think of it as the variance of  $Y$  given the value  $X$ . In particular, if we have the observed  $X = x$ , then we can denote and define

$$\text{Var}[Y|X = x] = E[(Y - E[Y|X = x])^2|X = x].$$

also note that if  $X$  and  $Y$  are independent,  $E[Y|X] = E[Y]$ , and the definition boils down to the regular variance. There is an analog of Corollary 2.2, which we leave to the reader to prove.

**Corollary 3.7.**

$$\text{Var}[Y|X] = E[Y^2|X] - (E[Y|X])^2$$

There is also a “law of total variance”, which looks slightly more complicated than that of total expectation.

**Proposition 3.19.**

$$\text{Var}[Y] = \text{Var}[E[Y|X]] + E[\text{Var}[Y|X]]$$

## 5.3 Covariance and Correlation

**Definition 3.16.** The *covariance* of  $X$  and  $Y$  is defined as

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])].$$

**Proposition 3.20.**

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$$

**Corollary 3.8.** If  $X$  and  $Y$  are independent, then  $\text{Cov}[X, Y] = 0$ .

**Proposition 3.21.**

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y].$$

*Proof.* By the definition of variance and covariance and repeated use of properties of expected values, we get

$$\begin{aligned} \text{Var}[X + Y] &= E[(X + Y - E[X + Y])^2] \\ &= E[(X - E[X] + Y - E[Y])^2] \\ &= E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y] \end{aligned}$$

and we are done. ■

**Proposition 3.22.** Let  $X, Y$ , and  $Z$  be random variables and let  $a$  and  $b$  be real numbers. Then

- (a)  $\text{Cov}[X, X] = \text{Var}[X]$
- (b)  $\text{Cov}[aX, bX] = ab\text{Cov}[X, X]$
- (c)  $\text{Cov}[X + Y, Z] = \text{Cov}[X, Z] + \text{Cov}[Y, Z]$

together these properties indicate that covariance is *bilinear*.

## 5.4 The Correlation Coefficient

**Definition 3.17.** The *correlation coefficient* of  $X$  and  $Y$  is defined as

$$\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}.$$

The correlation coefficient is dimensionless. To demonstrate this, take  $a, b > 0$  and note that

$$\begin{aligned} \rho(aX, bY) &= \frac{\text{Cov}[aX, bY]}{\sqrt{\text{Var}[aX]\text{Var}[bY]}} \\ &= \frac{ab\text{Cov}[X, Y]}{\sqrt{a^2\text{Var}[X]b^2\text{Var}[Y]}} = \rho(X, Y). \end{aligned}$$

We also call  $\rho(X, Y)$  simply the correlation between  $X$  and  $Y$ . Here are some good properties of the correlation coefficient.

**Proposition 3.25.** The correlation coefficient of any pair of random variables  $X$  and  $Y$  satisfies

- (a)  $-1 \leq \rho(X, Y) \leq 1$
- (b) If  $X$  and  $Y$  are independent, then  $\rho(X, Y) = 0$
- (c)  $\rho(X, Y) = 1$  if and only if  $Y = aX + b$ , where  $a > 0$
- (d)  $\rho(X, Y) = -1$  if and only if  $Y = aX + b$ , where  $a < 0$

*Proof.* Let  $\text{Var}[X] = \sigma_1^2$  and  $\text{Var}[Y] = \sigma_2^2$ . For (a), first apply Proposition 3.21 to the random variables  $X/\sigma_1$  and  $Y/\sigma_2$  and use the properties of the variance and covariance to obtain

$$0 \leq \text{Var}\left[\frac{X}{\sigma_1} + \frac{Y}{\sigma_2}\right] = \frac{\text{Var}[X]}{\sigma_1^2} + \frac{\text{Var}[Y]}{\sigma_2^2} + \frac{2\text{Cov}[X, Y]}{\sigma_1\sigma_2} = 2 + 2\rho$$

which gives  $\rho \geq -1$ . To show that  $\rho \leq 1$ , instead use  $X/\sigma_1$  and  $-Y/\sigma_2$ . Part (b) follows from Corollary 3.8, and parts (c) and (d) follows from Proposition 2.16, applied to the random variables  $X/\sigma_1 - Y/\sigma_2$  and  $X/\sigma_1 + Y/\sigma_2$ , respectively. Note that this also gives  $a$  and  $b$  expressed in terms of the means, variances, and correlation coefficient (see problem 90). ■

**Her mangle der bivariate normal!**

## 6 Generating functions (possibly with a focus on how probability generating functions relate to thinning of a Poisson process).

*Generating functions*, or *transforms*, are very useful in probability theory as in other fields of mathematics. Several different generating functions are used, depending on the type of random variable. We will discuss two, one that is useful for discrete random variables and the other for continuous random variables.

### 6.1 The Probability Generating Function

When we study nonnegative, integer-valued random variables, the following function proves to be a useful tool.

**Definition 3.23.** Let  $X$  be nonnegative and integer valued. The function

$$G_X(s) = E[s^X] = \sum_{k=0}^{\infty} s^k P(X = k) \quad 0 \leq s \leq 1$$

is called the *probability generating function* (pgf) of  $X$ .

**Corollary 3.12.** Let  $X$  be a nonnegative and integer valued with pgf  $G_X$ . then

$$G_X(0) = p_X(0) \quad \text{and} \quad G_X(1) = 1$$

**Proposition 3.36.** Let  $X$  be nonnegative and integer valued with pgf  $G_X$ . then

$$p_X(k) = \frac{G_X^{(k)}(0)}{k!}, \quad k = 0, 1, \dots$$

where  $G_X^{(k)}(0)$  denotes the  $k$ th derivative of  $G_X$ .

**Proposition 3.37.** If  $X$  has pgf  $G_X$ , Then

$$E[X] = G_X'(1) \quad \text{and} \quad \text{Var}[X] = G_X''(1) + G_X'(1) - G_X'(1)^2$$

**Proposition 3.38.** Let  $X_1, X_2, \dots, X_n$  be independent random variables with pgfs  $G_1, G_2, \dots, G_n$  respectively and let  $S_n = X_1 + X_2 + \dots + X_n$ . Then  $S_n$  has pgf

$$G_{S_n}(s) = G_1(s)G_2(s) \cdots G_n(s), \quad 0 \leq s \leq 1$$

*Proof.* Since  $X_1, \dots, X_n$  are independent, the random variables  $s^{X_1}, \dots, s^{X_n}$  are also independent for each  $s$  in  $[0,1]$ , and we get

$$\begin{aligned} G_{S_n}(s) &= E[s^{X_1+X_2+\dots+X_n}] \\ &= E[s^{X_1}]E[s^{X_2}] \cdots E[s^{X_n}] \\ &= G_1(s)G_2(s) \cdots G_n(s) \end{aligned}$$

and we are done. ■



**Proposition 3.39.** Let  $X_1, X_2, \dots$  be i.i.d. nonnegative and integer valued with common pgf  $G_X$ , and let  $N$  be nonnegative and integer valued, and independent of the  $X_k$ , with pgf  $G_N$ . Then  $S_N = X_1 + \dots + X_N$  has pgf

$$G_{S_N}(s) = G_N(G_X(s))$$

the composition of  $G_N$  and  $G_X$ .

*Proof.* We condition on  $N$  to obtain

$$\begin{aligned} G_{S_N}(s) &= E[E[s^{S_N} | N = n]] \\ &= \sum_{n=0}^{\infty} E[s^{S_N} | N = n] P(N = n) \\ &= \sum_{n=0}^{\infty} E[s^{S_N}] P(N = n) \end{aligned}$$

since  $N$  and  $S_n$  are independent. Now note that

$$E[s^{S_N}] = G_X(s)^n$$

by Proposition 3.38 and we get

$$G_{S_N}(s) = \sum_{n=0}^{\infty} G_X(s)^n P(N = n) = G_N(G_X(s))$$

the pgf of  $N$  evaluated at the point  $G_X(s)$ . ■

**Corollary 3.13.** Under the assumptions of Proposition 3.39, it holds that

$$\begin{aligned} E[S_N] &= E[N]\mu \\ \text{Var}[S_N] &= E[N]\sigma^2 + \text{Var}[N]\mu^2 \end{aligned}$$

where  $\mu = E[X_k]$  and  $\sigma^2 = \text{Var}[X_k]$ .

## 6.2 The Moment Generating Function

The probability generating function is an excellent tool for nonnegative and integer-valued random variables. For other random variables, we can instead use the following more general generating function.

**Definition 3.24.** Let  $X$  be a random variable. The function

$$M_X(t) = E[e^{tX}], \quad t \in \mathbb{R}$$

is called the *moment generating function* (mgf) of  $X$ .

If  $X$  is continuous with pdf  $f_X$ , we compute the mgf by

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

and for discrete  $X$ , we get a sum instead. Note that if  $X$  is nonnegative integer valued with pgf  $G_X$ , then

$$M_X(t) = G_X(e^t)$$

which immediately give the mgf for the distributions for which we computed the pgf above.

**Corollary 3.15.** If  $X$  has mgf  $M_X(t)$ , then

$$E[X] = M'_X(0) \quad \text{and} \quad \text{Var}[X] = M''_X(0) - (M'_X(0))^2$$

*Proof.* By differentiating  $M_X(t)$  with respect to  $t$ , we get

$$M'_X(t) = \frac{d}{dt} E[e^{tx}] = E[Xe^{tx}]$$

where we assumed that we can interchange differentiation and expectation. Note how  $t$  is the variable of differentiation and we view  $X$  as fixed. In the case of a discrete  $X$ , this amounts to differentiating a sum termwise, and in the case of a continuous  $X$ , it means that we can differentiate under the integral sign. This is by no means obvious but can be verified. We will not address this issue further. Differentiating once more gives

$$M''_X(t) = \frac{d}{dt} E[Xe^{tx}] = E[X^2 e^{tx}]$$

which gives

$$\text{Var}[X] = E[X^2] - (E[X])^2 = M''_X(0) - (M'_X(0))^2$$

as desired. ■

Note we get the general result

$$E[X^n] = M_X^{(n)}(0), \quad n = 1, 2, \dots$$

where  $M_X^{(n)}$  is the  $n$ th derivative of  $M_X$ . The number  $E[X^n]$  is called the  $n$ th *moment* of  $X$ , hence the term moment generating function. Compare it with the probability generating function that generates probabilities by differentiating and setting  $s = 0$ . The moment generating function also turns sum into products, according to the following proposition, which you may prove as an exercise.

**Proposition 3.40.** Let  $X_1, X_2, \dots, X_n$  be independent random variables with mgfs  $M_1, M_2, \dots, M_n$ , respectively, and let  $S_n = X_1 + \dots + X_n$ . Then  $S_n$  has mgfs

$$M_{S_n}(t) = M_1(t)M_2(t) \cdots M_n(t), \quad t \in \mathbb{R}$$

## 6.3 The Poisson Process

**Definition 3.25.** A point process where times between consecutive points are i.i.d. random variables that are  $\text{exp}(\lambda)$  is called a *Poisson process* with rate  $\lambda$ .

**Proposition 3.41.** Consider a Poisson process with rate  $\lambda$ , where  $X(t)$  is the number of points in an interval of length  $t$ . Then

$$X(t) \sim \text{Poi}(\lambda t)$$

Recall that the parameter in the Poisson distribution is also the expected value. Hence, we have

$$E[X(t)] = \lambda t$$

which makes sense since  $\lambda$  is the mean number of points per time unit and  $t$  is the length of the time interval. In practical applications, we need to be careful to use the same time units for  $\lambda$  and  $t$ .

**Proposition 3.42.** In a Poisson process with rate  $\lambda$

- (a)  $T_1, T_2, \dots$  are independent and  $\exp(\lambda)$
- (b)  $X(T_1), X(T_2), \dots$  are independent and  $X(t_j) \sim \text{Poi}(\lambda t_j)$ ,  $j = 1, 2, \dots$

**Proposition 3.43.** Consider a Poisson process with rate  $\lambda$ . If there are  $n$  points in the time interval  $[0, t]$ , their joint distribution is the same as that of  $n$  i.i.d. random variables that are  $\text{unif}[0, t]$

The property in the proposition is called the *order statistic property* of the Poisson process since the points are distributed as  $n$  order statistics from a uniform distribution on  $[0, t]$ . The proof is similar to what we did previously in the case of  $n = 1$ , invoking Proposition 3.32.

## 6.4 Thinning and Superposition

Suppose now that we have a Poisson process with rate  $\lambda$ , where we do not observe every point, either by accident or on purpose. For example, phone calls arrive as a Poisson process, but calls are lost when the line is busy. Hurricanes are formed according to a Poisson process, but we are interested only in those that make landfall. These examples of *thinning* of a Poisson process.

We assume that each point is observed with probability  $p$  and that different points are observed independent of each other. Then, it turns out that the thinned process is also Poisson process.

**Proposition 3.44.** The thinned process is a Poisson process with rate  $\lambda p$ .

*Proof.* We work with characterization (b) in Proposition 3.42. Clearly, the numbers of observed points in disjoint intervals are independent. To show the Poisson distribution, we use probability generating functions. Consider an interval of length  $t$ , letting  $X(t)$  be the total number of points and  $X_p(t)$  be the number of observed points in this interval. Then,

$$X_p(t) = \sum_{k=1}^{X(t)} I_k$$

where  $I_k$  is 1 if the  $k$ th point was observed and 0 otherwise. By Proposition 3.39,  $X_p(t)$  has pgf

$$G_{X_p}(s) = G_{X(t)}(G_I(s))$$

where

$$G_{X(t)}(s) = e^{\lambda t(s-1)}$$

and

$$G_I(s) = 1 - p + ps$$

and we get

$$G_{X_p}(s) = e^{\lambda t(1-p+ps-1)} = e^{\lambda pt(s-1)}$$

which we recognize as the pgf of a Poisson distribution with parameter  $\lambda pt$ . ■

**Proposition 3.45.** The processes of observed and unobserved points are independent.

*Proof.* Fix an interval of length  $t$ , let  $X(t)$  be the total number of points, and  $X_p(t)$  and  $X_{1-p}(t)$  the number of observed and unobserved points, respectively. Hence,  $X(t) = X_p(t) + X_{1-p}(t)$  and by Proposition 3.44, we obtain

$$X_p(t) \sim \text{Poi}(\lambda pt) \quad \text{and} \quad X_{1-p}(t) \sim \text{Poi}(\lambda(1-p)t).$$

Also given that  $X(t) = n$ , the number of observed points has a binomial distribution with parameters  $n$  and  $p$ . We get

$$\begin{aligned} P(X_p(t) = j, X_{1-p}(t) = k) &= P(X_p(t) = j, X(t) = k + j) \\ &= P(X_p(t) = j | X(t) = k + j) P(X(t) = k + j) \\ &= \binom{k+j}{j} p^j (1-p)^k e^{-\lambda t} \frac{(\lambda t)^{k+j}}{(k+j)!} \\ &= \frac{(k+j)!}{k!j!} p^j (1-p)^k e^{-\lambda t} \frac{(\lambda t)^{k+j}}{(k+j)!} \\ &= e^{-\lambda pt} \frac{(\lambda pt)^j}{j!} e^{-\lambda(1-p)t} \frac{(\lambda(1-p)t)^k}{k!} \\ &= P(X_p(t) = j) P(X_{1-p}(t) = k) \end{aligned}$$
■

**Proposition 3.46.** Consider two independent Poisson processes with rate  $\lambda_1$  and  $\lambda_2$ , respectively. The superposition of the two processes is a Poisson process with rate  $\lambda_1 + \lambda_2$ .

*Proof.* For this we use characterization (a) in Proposition 3.42. Fix a time where there is a point in either of the two processes. Denote the time until the next point in the first process by  $S$  and the corresponding time in the second by  $T$ . Since the two processes are independent,  $S$  and  $T$  are independent random variables. Also, by the memoryless property, the distributions are  $S \sim \exp(\lambda_1)$  and  $T \sim \exp(\lambda_2)$ , regardless of which of the two processes the last point came from. Hence, the time until the next point in the superposition process is the minimum of  $S$  and  $T$ , and by example 3.50, this is  $\exp(\lambda_1 + \lambda_2)$ . Thus, the superposition process is a Poisson process with rate  $\lambda_1 + \lambda_2$  and we are done. ■

## 7 Limit theorems.

When we introduced expected values, we argued that these could be considered averages of a large number of observations. Thus, if we have observations  $X_1, X_2, \dots, X_n$  and we do not know the mean  $\mu$ , a reasonable approximation ought to be the *sample mean*

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$$

in other words, the average of  $X_1, \dots, X_n$ . Suppose now that the  $X_k$  are i.i.d. with mean  $\mu$  and variance  $\sigma^2$ . By the formulas for the mean and variance of sums of independent variables, we get

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{k=1}^n X_k\right] = \sum_{k=1}^n \frac{1}{n} E[X_k] = \mu$$

and

$$Var[\bar{X}] = Var\left[\frac{1}{n} \sum_{k=1}^n X_k\right] = \sum_{k=1}^n \frac{1}{n^2} Var[X_k] = \frac{\sigma^2}{n}$$

that is,  $\bar{X}$  has the same expected value as each individual  $X_k$  and a variance that becomes smaller the larger the value of  $n$ .

### 7.1 The Law of Large Numbers

Although we can never guarantee that  $|\bar{X} - \mu|$  is smaller than a given  $\varepsilon$  we can say that is very likely that  $|\bar{X} - \mu|$  is small if  $n$  is large. That is the idea behind the following result.

**Theorem 4.1. (The Law of Large Numbers).** Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables with mean  $\mu$ , and let  $\bar{X}$  be their sample mean. Then, for every  $\varepsilon > 0$

$$P(|\bar{X} - \mu| > \varepsilon) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

*Proof.* Assume that the  $X_k$  have finite variance,  $\sigma^2 < \infty$ . Apply Chebyshev's inequality to  $\bar{X}$  and let  $c = \varepsilon\sqrt{n}/\sigma$ . Since  $E[\bar{X}] = \mu$  and  $Var[\bar{X}] = \sigma^2/n$ , we get

$$P(|\bar{X} - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

The assumptions of finite variance is necessary for this proof to work. However, the law of large numbers is true also if the variance is infinite, but the proof in that case is more involved and we will not give it. ■

We say that  $\bar{X}$  *converges in probability* to  $\mu$  and write

$$\bar{X} \xrightarrow{P} \mu \quad \text{as } n \rightarrow \infty$$

**Corollary 4.1.** Consider an experiment where the event  $A$  occurs with probability  $p$ . Repeat the experiment independently, let  $S_n$  be the number of times we get the event  $A$  in  $n$  trials, and let  $f_n = S_n/n$ , the relative frequency. Then

$$f_n \xrightarrow{P} p \quad \text{as } n \rightarrow \infty$$

*Proof.* Define the indicators

$$I_k = \begin{cases} 1 & \text{if we get } A \text{ in the } k\text{th trial} \\ 0 & \text{otherwise} \end{cases}$$

for  $k = 1, 2, \dots, n$ . Then the  $I_k$  are i.i.d. and we know from Section 2.5.1 that they have mean  $\mu = p$ . Since  $f_n$  is the sample mean of the  $I_k$ , the law of large numbers gives  $f_n \xrightarrow{P} p$  as  $n \rightarrow \infty$ . ■

**Theorem 4.2 (The Central Limit Theorem).** Let  $X_1, X_2, \dots$  be i.i.d. random variables with mean  $\mu$  and variance  $\sigma^2 < \infty$  and let  $S_n = \sum_{k=1}^n X_k$ . Then, for each  $x \in \mathbb{R}$ , we have

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n \rightarrow \infty]{d} N(0, 1)$$

as  $n \rightarrow \infty$ , where  $\Phi$  is the cdf of the standard normal distribution.

**Definition 4.1.** Let  $X_1, X_2, \dots$  be a sequence of discrete random variables such that  $X_n$  has pmf  $p_{X_n}$ . If  $X$  is a discrete random variable with pmf  $p_X$  and

$$p_{X_n}(x) \rightarrow p_X(x) \quad \text{as } n \rightarrow \infty \quad \text{for all } x$$

then we say that  $X_n$  converges in distribution to  $X$ , written  $X_n \xrightarrow{d} X$ .

**Proposition 4.2.** Let  $X_1, X_2, \dots$  be a sequence of random variables such that  $X_n \sim \text{bin}(n, p_n)$ , where  $np_n \rightarrow \lambda > 0$  as  $n \rightarrow \infty$ , and let  $X \sim \text{Poi}(\lambda)$ . Then  $X_n \xrightarrow{d} X$ .

## 7.2 Continuous Limits

Let us next consider the case when the limiting random variable is continuous. As we already know from the de Moivre-Laplace theorem, the limit can be continuous even if the random variables themselves are not.

**Definition 4.2.** Let  $X_1, X_2, \dots$  be a sequence of random variables such that  $X_n$  has cdf  $F_n$ . If  $X$  is a continuous random variable with cdf  $F$  and

$$F_n(x) \rightarrow F(x) \quad \text{as } n \rightarrow \infty \quad \text{for all } x \in \mathbb{R}$$

we say that  $X_n$  converges in distribution to  $X$ , written  $X_n \xrightarrow{d} X$ .

The most important result of this type is the central limit theorem. Another class of important results regarding convergence in distribution deals with the so-called *extreme values*, for example, the minimum and maximum in a sequence of random variables.

## 8 Markov chains.

Many real-world applications of probability theory have one particular feature that data are collected sequentially in time. A few examples are weather data, stock market indices, air-pollution data, demographic data, and political tracking polls. These also have another feature in common that successive observations are typically not independent. We refer to any such collection of observation as a *stochastic process*. Formally, a stochastic process is a collection of random variables that take values in a set  $S$ , the *state space*. The collection is indexed by another set  $T$  the *index set*. The two most common sets are the natural numbers  $T = \{0, 1, 2, \dots\}$  and the non negative real numbers  $T = [0, \infty)$ , which usually represent discrete time and continuous time, respectively. The first index set thus gives a sequence of random variables  $\{X_0, X_1, X_2, \dots\}$  and the second, a collection of random variables  $\{X(t), t \geq 0\}$ , one random variable for each time  $t$ . In general, the index set does not have to describe time and is also commonly used to describe spatial location. The state space can be finite, countably infinite, or uncountable, depending on the application.

### 8.1 Discrete-Time Markov Chains

**Definition 8.1..** Let  $X_0, X_1, X_2, \dots$  be a sequence of discrete random variables, taking values in some set  $S$  and that are such that

$$P(X_{n+1} = j | X_0 = i_0, \dots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j | X_n = i) \quad (8.1)$$

for all  $i, j, i_0, \dots, i_{n-1}$  in  $S$  and all  $n$ . The sequence  $\{X_n\}$  is then called a *Markov chain*. Furthermore (8.1) is called the *Markov property*.

We often think of the index  $n$  as discrete time and say that  $X_n$  is the *state* of the chain at time  $n$ , where the state space  $S$  may be finite or countably infinite.

In general, the probability  $P(X_{n+1} = j | X_n = i)$  depends in  $i, j$ , and  $n$ . It is, however, often the case that there is no dependence on  $n$ . We call such chains *time-homogenous* and restrict our attention to these chains. Since the conditional probability in the definition thus depends only on  $i$  and  $j$ , we use the notation

$$p_{ij} = P(X_{n+1} = j | X_n = i), \quad i, j \in S$$

and call these the *transition probabilities* of the Markov chain. Thus, if the chain is in state  $i$ , the probabilities  $p_{ij}$  describe how the chain chooses which state to jump to next. Obviously, the transition probabilities have to satisfy the following two criteria:

$$(a) \ p_{ij} \geq 0 \quad \text{for all } i, j \in S, \quad (b) \ \sum_{j \in S} p_{ij} = 1 \quad \text{for all } i \in S$$

#### 8.1.1 Time Dynamics of a Markov Chain

The most fundamental aspect of a Markov chain in which we are interested is how it develops over time. The transition matrix provides us with a description of the stepwise behaviour, but suppose that we want to compute the distribution of the chain two steps ahead. Let

$$p_{ij}^{(2)} = P(X_2 = j | X_0 = i)$$

and condition on the intermediate step  $X_1$ . The law of total probability gives

$$\begin{aligned} p_{ij}^{(2)} &= \sum_{k \in S} P(X_2 = j | X_0 = i, X_1 = k) P(X_1 = k | X_0 = i) \\ &= \sum_{k \in S} P(X_2 = j | X_1 = k) P(X_1 = k | X_0 = i) = \sum_{k \in S} p_{ik} p_{kj} \end{aligned}$$

where we used the Markov property for the second-to-last equality. We switched the order between the factors in the sum to get the intuitively appealing last expression; in order to go from  $i$  to  $j$  in two steps, we need to visit intermediate step  $k$  and jump from there to  $j$ . Now, recall how matrix multiplication works to help us realize from the expression above that  $p_{ij}^{(2)}$  is the  $(i, j)$ th entry in the matrix  $P^2$ . Thus, in order to get the two-step transition probabilities, we square the transition matrix. Generally, define the  $n$ -step transition probabilities as

$$p_{ij}^{(n)} = P(X_n = j | X_0 = i)$$

and let  $P^{(n)}$  be the  $n$ -step transition matrix. Repeating the argument above gives  $P^{(n)} = P^n$ , the  $n$ th power of the one-step transition matrix. In particular, this gives the relation

$$P^{(n+m)} = P^{(n)} P^{(m)}$$

for all  $m, n$ , commonly referred to as the *Chapman-Kolmogorov equations*. Spelled out coordinatewise, they become

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{jk}^{(m)}$$

for all  $m, n$  and all  $i, j \in S$ . In words, to go from  $i$  to  $j$  in  $n + m$  steps, we need to visit some intermediate step  $k$  after  $n$  steps. We let  $P^{(0)} = I$ , the identity matrix.

### 8.1.2 Classification of States

**Definition 8.2.** If  $p_{ij}^{(n)} > 0$  for some  $n$ , we say that state  $j$  is *accessible* from state  $i$ , written  $i \rightarrow j$ . If  $j \rightarrow i$ , we say that  $i$  and  $j$  *communicate* and write this  $i \leftrightarrow j$ .

In general, if we fix a state  $i$  in the state space of a Markov chain, we can find all states that communicate with  $i$  and form a *communicating class* containing  $i$ . It is easy to realize that not only does  $i$  communicate with all states in this class but they all communicate with each other. By convention, every state communicates with itself (it can reach itself in 0 steps) so every state belongs to a class. If you wish to be more mathematical, the relation “ $\leftrightarrow$ ” is an equivalence relation and thus divides the state space into equivalence classes that are precisely the communicating classes.

**Definition 8.3.** If all states in  $S$  communicate with each other, the Markov chain is said to be *irreducible*.

**Definition 8.4.** Consider a state  $i \in S$  and the  $\tau_i$  be the number of steps it takes for the chain to first visit  $i$ . Thus

$$\tau_i = \min\{n \geq 1 : X_n = i\}$$

where  $\tau_i = \infty$  if  $i$  is never visited. If  $P_i(\tau_i < \infty) = 1$ , the state  $i$  is said to be *recurrent* and if



$P_i(\tau_i < \infty) < 1$ , it is said to be *transient*.

**Corollary 8.1.** In an irreducible Markov chain, either all states are transient or all states are recurrent.

**Corollary 8.2.** Suppose that  $S$  is finite. A state  $i$  is transient if and only if there is another state  $j$  such that  $i \rightarrow j$  but  $j \nrightarrow i$ .

**Corollary 8.3.** If a Markov chain has finite state space, there is at least one recurrent state.

**Proposition 8.1.** State  $i$  is

$$\begin{aligned} \text{transient if } \sum_{n=1}^{\infty} p_{ii}^{(n)} &< \infty \\ \text{recurrent if } \sum_{n=1}^{\infty} p_{ii}^{(n)} &= \infty \end{aligned}$$

### 8.1.3 Stationary Distribution

**Definition 8.5.** Let  $P$  be the transition matrix of a Markov chain with state space  $S$ . A probability distribution  $\pi = (\pi_1, \pi_2, \dots)$  on  $S$  satisfying

$$\pi P = \pi$$

is called a *stationary distribution* of the chain.

**Proposition 8.2.** Consider an irreducible Markov chain. If a stationary distribution exists, it is unique.

**Proposition 8.3.** If  $S$  is finite and the Markov chain is irreducible, a unique stationary distribution  $\pi$  exists.

**Definition 8.6.** Let  $i$  be a recurrent state. If  $E_i[\tau_i] < \infty$ , then  $i$  is said to be *positively recurrent*. If  $E_i[\tau_i] = \infty$ ,  $i$  is said to be *null recurrent*.

**Corollary 8.4.** For an irreducible Markov chain, there are three possibilities: **(a)** all states are positive recurrent, **(b)** all states are null recurrent, and **(c)** all states are transient.

## 9 Stochastic simulation.