Probability and Statistics

Kristel Van Steen, PhD^2

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 2: RANDOM VARIABLES AND ASSOCIATED FUNCTIONS

3 Two or more random variables

3.1 Joint probability distribution function

3.2 The discrete case: Joint probability mass function

A two-dimensional random walk

3.3 The continuous case: Joint probability density function

Meeting times

4 Conditional distribution and independence

5 Expectations and moments

5.1 Mean, median and mode

A one-dimensional random walk

5.2 Central moments, variance and standard deviation

5.3 Moment generating functions

6 Functions of random variables

6.1 Functions of one random variable

6.2 Functions of two or more random variables

6.3 Two or more random variables: multivariate moments

7 Inequalities

7.1 Jensen inequality

7.2 Markov’s inequality

7.3 Chebyshev’s inequality

7.4 Cantelli’s inequality

7.5 The law of large numbers

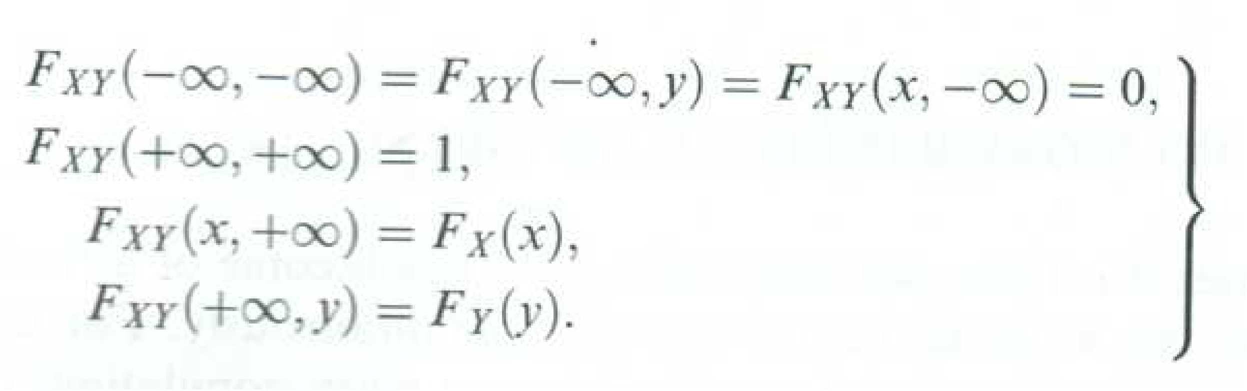
3.1 Joint probability distribution functions

- The joint probability distribution function of random variables X and Y, denoted by $F_{XY}(x,y)$, is defined by

$$F_{XY}(x,y) = P(X \leq x \cap Y \leq y),$$

for all x, y

- As before, some obvious properties follow from this definition of joint **cumulative distribution function**:



$$\left.\begin{aligned} &F_{XY}(-\infty,-\infty) = F_{XY}(-\infty,y) = F_{XY}(x,-\infty) = 0, \\ &F_{XY}(+\infty,+\infty) = 1, \\ &\quad F_{XY}(x,+\infty) = F_X(x), \\ &\quad F_{XY}(+\infty,y) = F_Y(y). \end{aligned}\right\}$$

- $F_X(x)$ and $F_Y(y)$ are called **marginal distribution functions** of X and Y, resp.

Copulas

- Consider a random vector (X_1, X_2) and suppose that its margins F_1 and F_2 are continuous. By applying the **probability integral transformation** to each component, the random vector

$$(U_1, U_2) = (F_1(X_1), F_2(X_2))$$

has uniform margins. The **copula** of (X_1, X_2) is defined as the joint cumulative distribution function of (U_1, U_2):

$$C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2)$$

3.2 The discrete case: joint probability mass functions

- Let X and Y be two discrete random variables that assume at most a countable infinite number of value pairs (x_i, y_j), i,j = 1,2, …, with nonzero probabilities. Then the **joint probability mass function** of X and Y is defined by

$$P_{XY}(x,y) = P(X = x \cap Y = y),$$

for all x and y. It is zero everywhere except at the points (x_i, y_j), i,j = 1,2, …,

where it takes values equal to the joint probability $P(X = x_i \cap Y = y_j)$.

(Example of a simplified random walk)

3.3 The continuous case: joint probability density functions

- The joint probability density function $f_{XY}(x,y)$ of 2 continuous random variables X and Y is defined by the partial derivative

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

- Since $F_{XY}(x,y)$ is monotone non-decreasing in both x and y, the associated joint probability density function is nonnegative for all x and y.
- As a direct consequence:

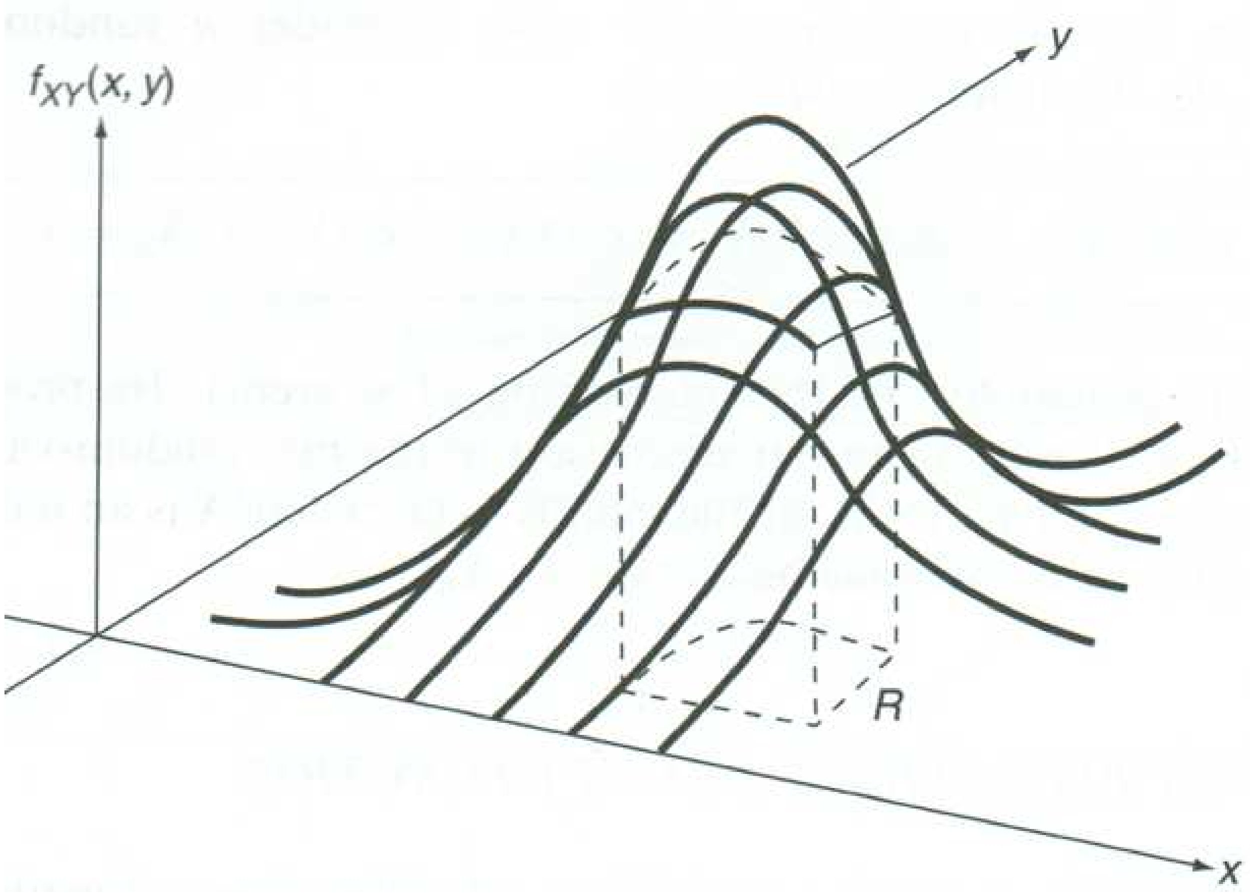
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x\mathrm{d}y = 1,$$

$$\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y = f_X(x),$$

$$\int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}x = f_Y(y).$$

where $f_X(x) \,\mathrm{and}\, f_Y(y)$ are now called the marginal density functions of X and Y respectively

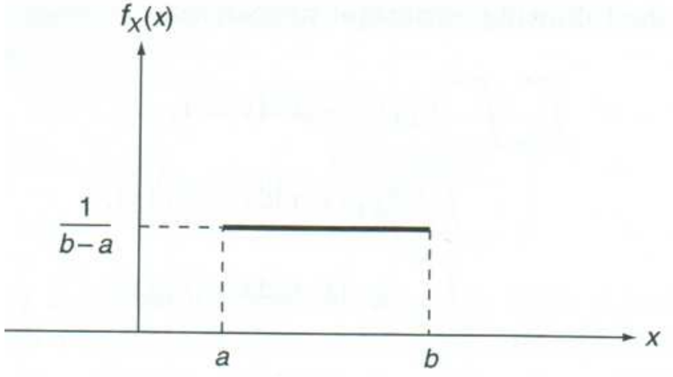
- Also, $F_{XY}(x,y) = P(X \leq x \cap Y \leq y) = \int_{<\infty}^{y} \int_{-\infty}^{x} f_{XY}(u,v) du dv$ and $P(x_1 < X \leq x_2 \cap y_1 < Y \leq y2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x,y) dx dy$ for $x_1 < x_2 \, \text{and} \, y_1 < y_2$

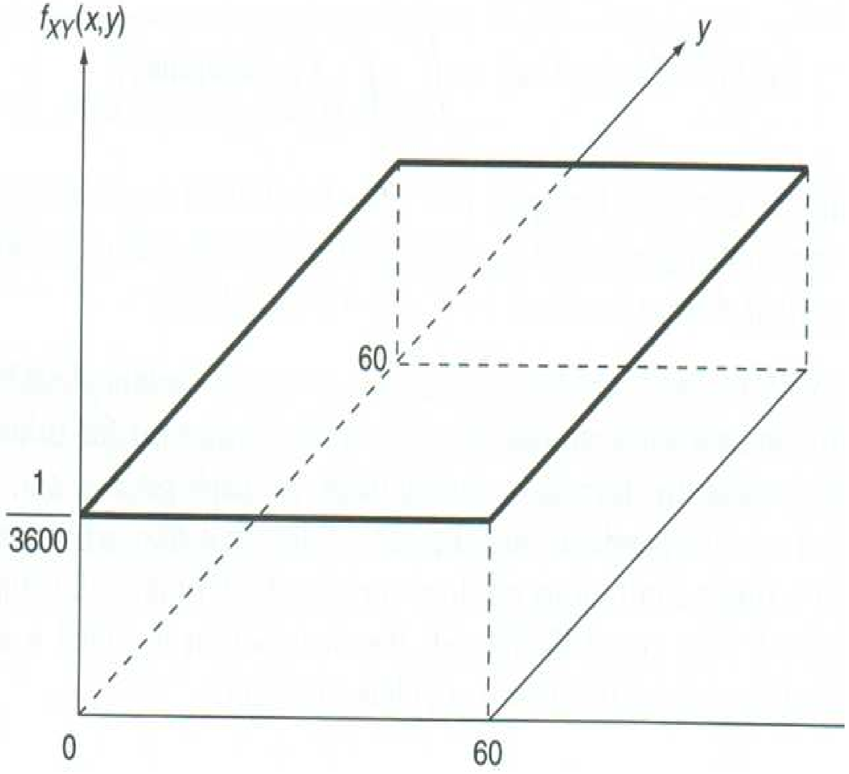


Meeting times

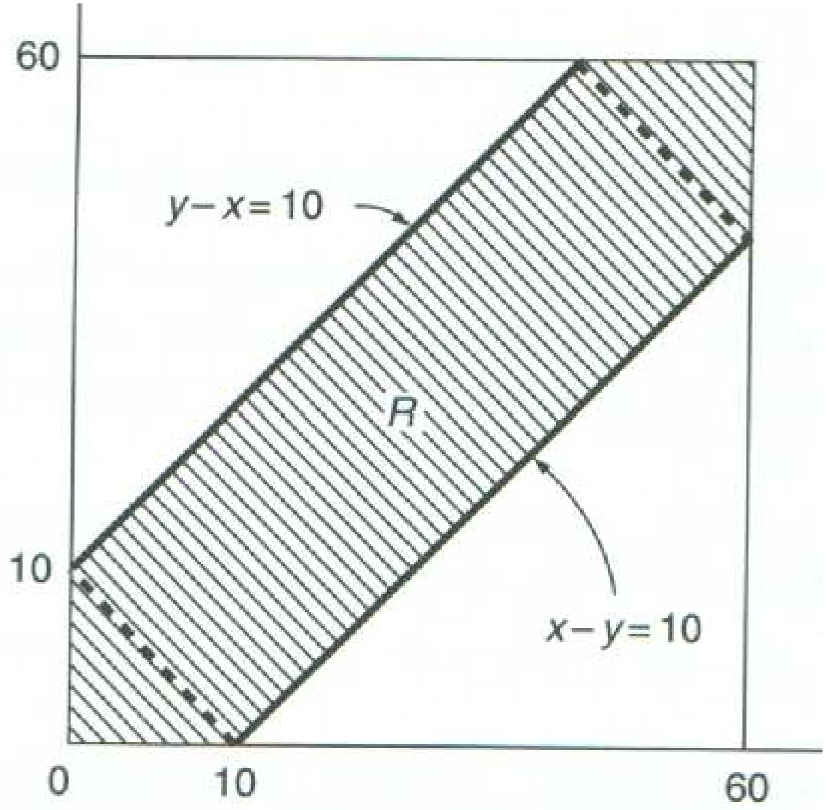
- A boy and a girl plan to meet at a certain place between 9am and 10am, each not wanting to wait more than 10 minutes for the other. If all times of arrival within the hour are equally likely for each person, and if their times of arrival are independent, find the probability that they will meet.
- Answer: for a single continuous random variable X that takes all values over an interval a to b with equal likelihood, the distribution is called a uniform distribution and its density function has the form

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{, for } a \leq x \leq b \\ 0 & \text{, otherwise} \end{cases}$$





The joint density function of two independent uniformly distributed random variables is a flat surface within prescribed bounds. The volume under the surface is unity.



$$
\begin{aligned}
P(\text{they will meet}) &= P(|X - Y| \leq 10) \\
&= [2(5)(10) + 10\sqrt{2}(50\sqrt{2})]/3600 = \frac{11}{36}
\end{aligned}
$$

- We can derive from the joint probability, the joint probability distribution function, as usual

$$F_{XY}(x, y) = \begin{cases} 0, & \text{for } (x, y) < (0, 0); \\ 1, & \text{for } (x, y) > (60, 60). \end{cases}$$

$$F_{XY}(x, y) = \int_0^y \int_0^x \left(\frac{1}{3600}\right) \mathrm{d}x\mathrm{d}y = \frac{xy}{3600}.$$

- From this we can again derive the marginal probability density functions, which clearly satisfy the earlier definition for 2 random variables that are uniformly distributed over the interval [0,60]

4 Conditional distribution and independence

- The concepts of conditional probability and independence introduced before also play an important role in the context of random variables
- The **conditional distribution** of a random variable X, given that another random variable Y has taken a value y, is defined by

$$F_{XY}(x|y) = P(X \leq x|Y = y)$$

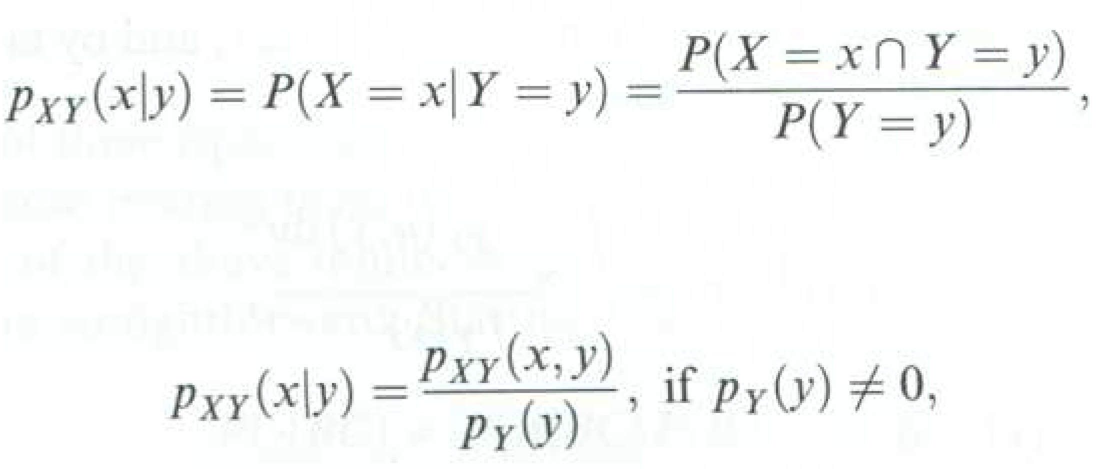
- When a random variable X is discrete, the definition of **conditional mass function** of X given Y=y is

$$p_{XY}(x|y) = P(X = x|Y = y)$$

- For a continuous random variable X, the **conditional density function** of X given Y=y is

$$f_{XY}(x|y) = \frac{dF_{XY}(x|y)}{dx}$$

- In the discrete case, using the definition of conditional probability, we have



$$p_{XY}(x|y) = P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)},$$

$$p_{XY}(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}, \text{ if } p_Y(y) \neq 0,$$

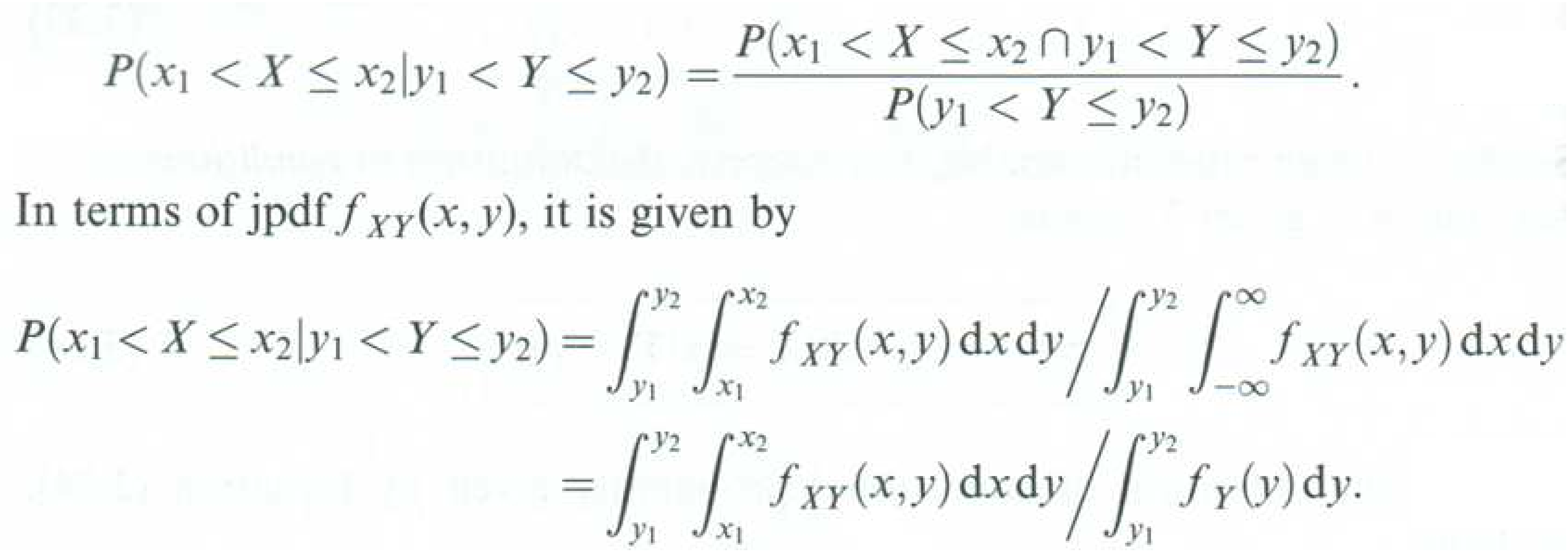
an expression which is very useful in practice when wishing to derive joint probability mass functions …

- Using the definition of independent events in probability theory, when the random variables X and Y are assumed to be independent,

$$\boxed{P_{XY}(x|y) = P_X(x)}$$

so that $P_{XY}(x, y) = P_X(x)P_Y(y)$

- The definition of a conditional density function for a random continuous variable X, given Y=y, entirely agrees with intuition …:



$$P(x_1 < X \leq x_2 | y_1 < Y \leq y_2) = \frac{P(x_1 < X \leq x_2 \cap y_1 < Y \leq y_2)}{P(y_1 < Y \leq y_2)}.$$

In terms of jpdf $f_{XY}(x,y)$, it is given by

$$\begin{aligned} P(x_1 < X \leq x_2 | y_1 < Y \leq y_2) &= \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y \Big/ \int_{y_1}^{y_2} \int_{-\infty}^{\infty} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x,y)\,\mathrm{d}x\,\mathrm{d}y \Big/ \int_{y_1}^{y_2} f_Y(y)\,\mathrm{d}y. \end{aligned}$$

By setting $x_1 = -\infty, x_2 = x, y_1 = y, y_2 = y + \Delta y$ and by taking the limit $\Delta y \to 0$, this reduced to $F_{XY}(x|y) = \dfrac{\int_{-\infty}^{x} f_{XY}(u,y)\,\mathrm{d}u}{f_Y(y)},$ provided $f_Y(y) \neq 0$

From

$$f_{XY}(x|y) = \frac{dF_{XY}(x|y)}{dx}$$

and

$$F_{XY}(x|y) = \frac{\int_{-\infty}^{x} f_{XY}(u, y)\,\mathrm{d}u}{f_Y(y)},$$

we can derive that

$$f_{XY}(x|y) = \frac{\mathrm{d}F_{XY}(x|y)}{\mathrm{d}x} = \frac{f_{XY}(x, y)}{f_Y(y)}, \quad f_Y(y) \neq 0,$$

a form that is identical to the discrete case. But note that

$$F_{XY}(x|y) \neq \frac{F_{XY}(x, y)}{F_Y(y)}.$$

- When random variables X and Y are independent, however, $F_{XY}(x|y) = F_X(x)$ (using the definition for $F_{XY}(x|y)$) and (using the expression

$$f_{XY}(x|y) = \frac{\mathrm{d}F_{XY}(x|y)}{\mathrm{d}x} = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad f_Y(y) \neq 0,$$

it follows that

$$f_{XY}(x|y) = f_X(x),$$

$$\boxed{f_{XY}(x,y) = f_X(x) f_Y(y),}$$

- Finally, when random variables X and Y are discrete,

$$F_{XY}(x|y) = \sum_{i=1}^{i:x_i \leq x} p_{XY}(x_i|y),$$

and in the case of a continuous random variable,

$$F_{XY}(x|y) = \int_{-\infty}^{x} f_{XY}(u|y)\mathrm{d}u.$$

Note that these are very similar to those relating the distribution and density functions in the univariate case.

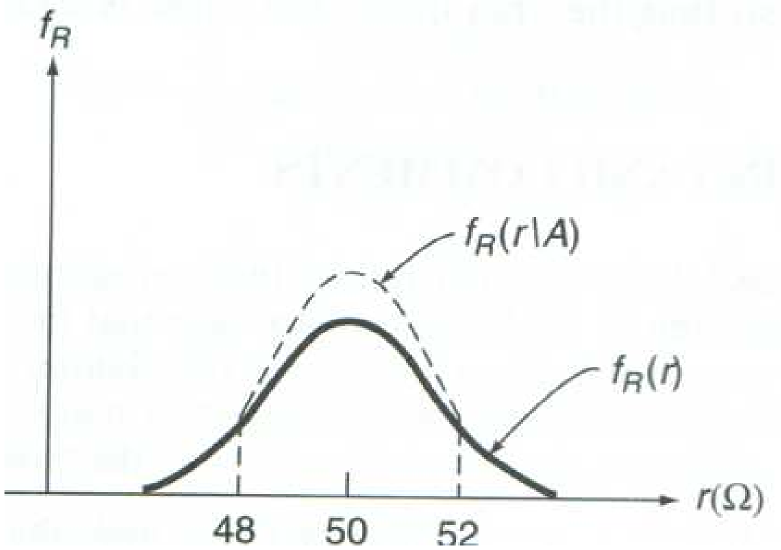
- Generalization to more than two variables should now be straightforward, starting from the probability expression

$$P(ABC) = P(A|BC)P(B|C)P(C)$$

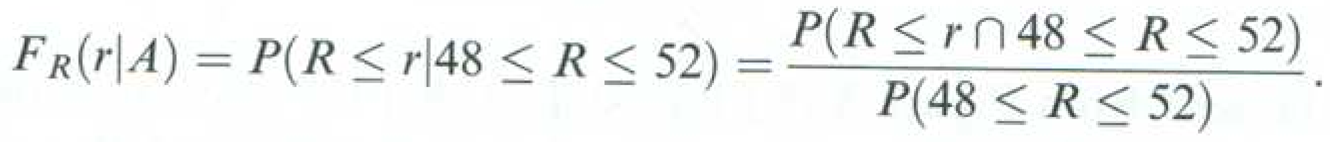
Resistor problem

- Resistors are designed to have a resistance of R of $50 \pm 2\Omega$. Owing to some imprecision in the manufacturing process, the actual density function of R has the form shown (right), by the solid curve.
- Determine the density function of R after screening (that is: after all the resistors with resistances beyond the 48-52 Ω range are rejected.

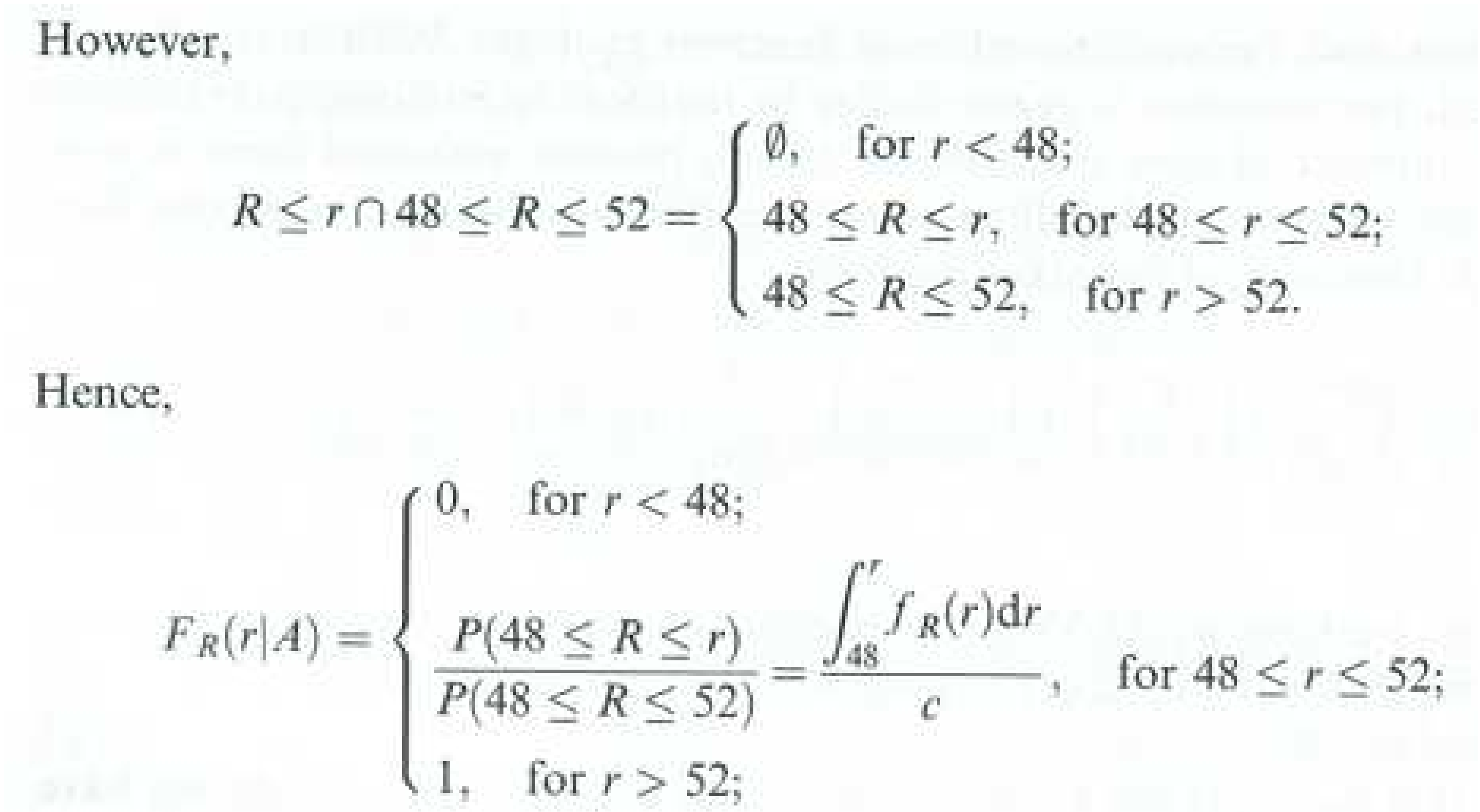
- Answer: we are interested in the conditional density function $f_R(r|A)$ where A is the event $\{48 \leq R \leq 52\}$



We start by considering



$$F_R(r|A) = P(R \le r|48 \le R \le 52) = \frac{P(R \le r \cap 48 \le R \le 52)}{P(48 \le R \le 52)}.$$



However,

$$R \le r \cap 48 \le R \le 52 = \begin{cases} \emptyset, & \text{for } r < 48; \\ 48 \le R \le r, & \text{for } 48 \le r \le 52; \\ 48 \le R \le 52, & \text{for } r > 52. \end{cases}$$

Hence,

$$F_R(r|A) = \begin{cases} 0, & \text{for } r < 48; \\ \dfrac{P(48 \le R \le r)}{P(48 \le R \le 52)} = \dfrac{\int_{48}^{r} f_R(r)\mathrm{d}r}{c}, & \text{for } 48 \le r \le 52; \\ 1, & \text{for } r > 52; \end{cases}$$

Where

$$c = \int_{48}^{52} f_R(r)dr$$

is a constant.

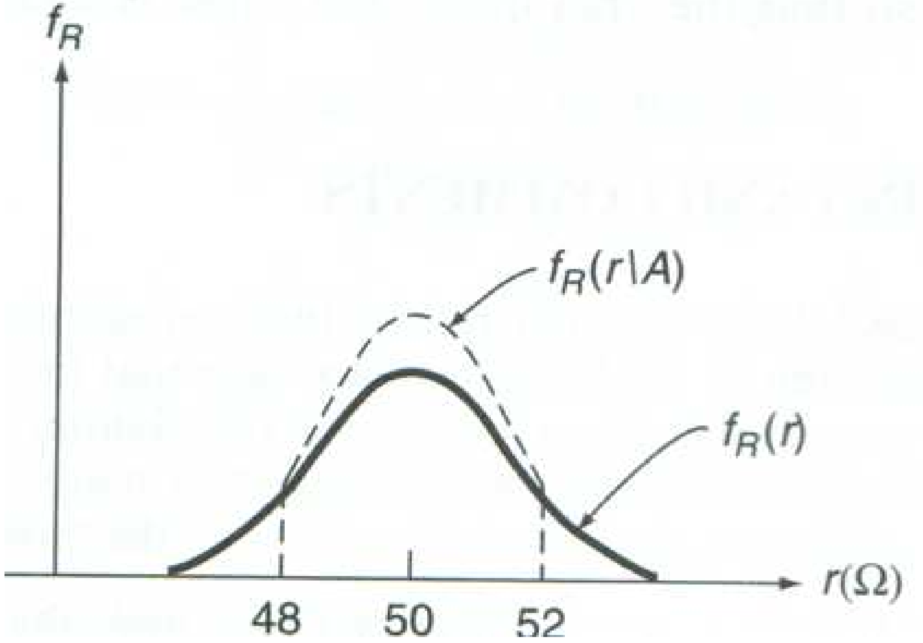
The desired function is then obtained by differentiation. We thus obtain

$$f_R(r|A) = \frac{dF_R(r|A)}{dr} = \begin{cases} \frac{f_R(r)}{c} & \text{for } 48 \leq r \leq 52 \\ 0 & \text{otherwise} \end{cases}$$

Now, look again at a graphical representation of this function. What do you observe?

Answer:

The effect of screening is essentially a truncation of the tails of the distribution beyond the allowable limits. This is accompanied by an adjustment within the limits by a multiplicative factor 1/c so that the area under the curve is again equal to 1.



5 Expectations and moments

5.1 Mean, median and mode

Expectations

- Let g(X) be a real-valued function of a random variable X. The mathematical expectation or simply expectation of g(X) is denoted by E(g(X)) and defined as

$$E(g(X)) = \sum_i g(x_i)P_X(x_i)$$

if X is discrete where $x_1, x_2, \ldots$are possible values assumed by X.

- When the range of i extends from 1 to infinity, the sum above exists if it converges absolutely; that is, $\sum_{i=1}^{\infty} |g(x_i)|P_X(x_i) < \infty$

- If the random variable X is continuous, then

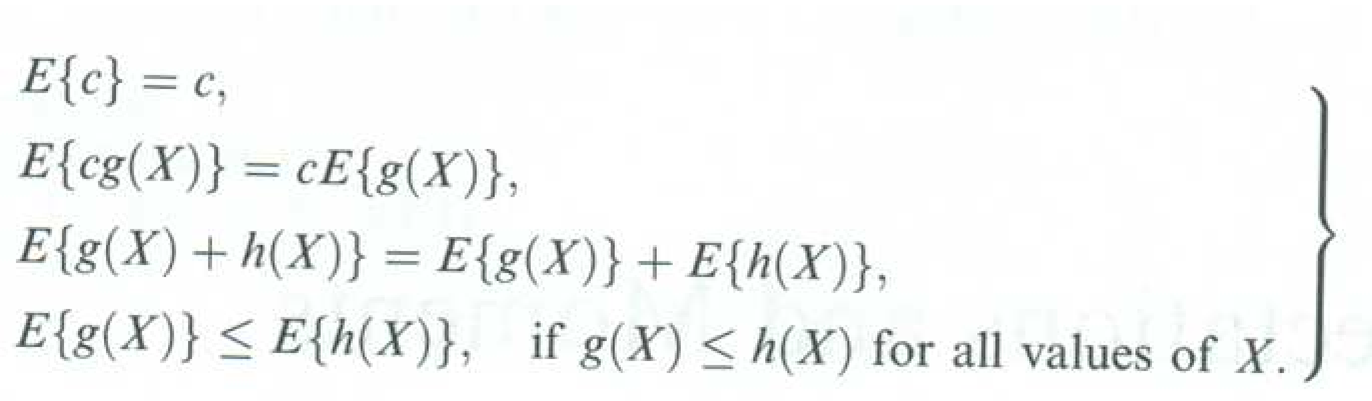
$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx,$$

if the improper integral is absolutely convergent, that is,

$$\int_{-\infty}^{+\infty} |g(x)| f_X(x) dx < \infty$$

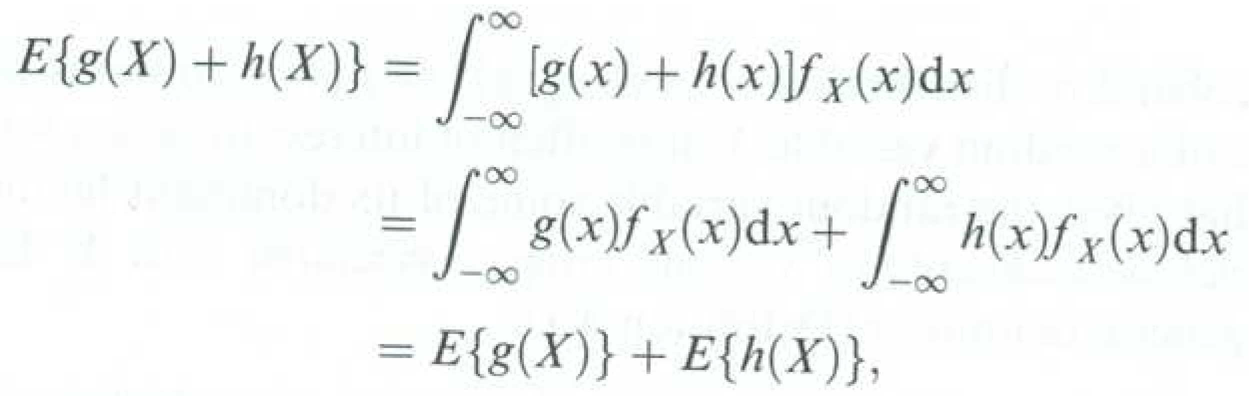
then this number will exist.

- Basic properties of the expectation operator E(.), for any constant c and any functions g(X) and h(X) for which expectations exist include:



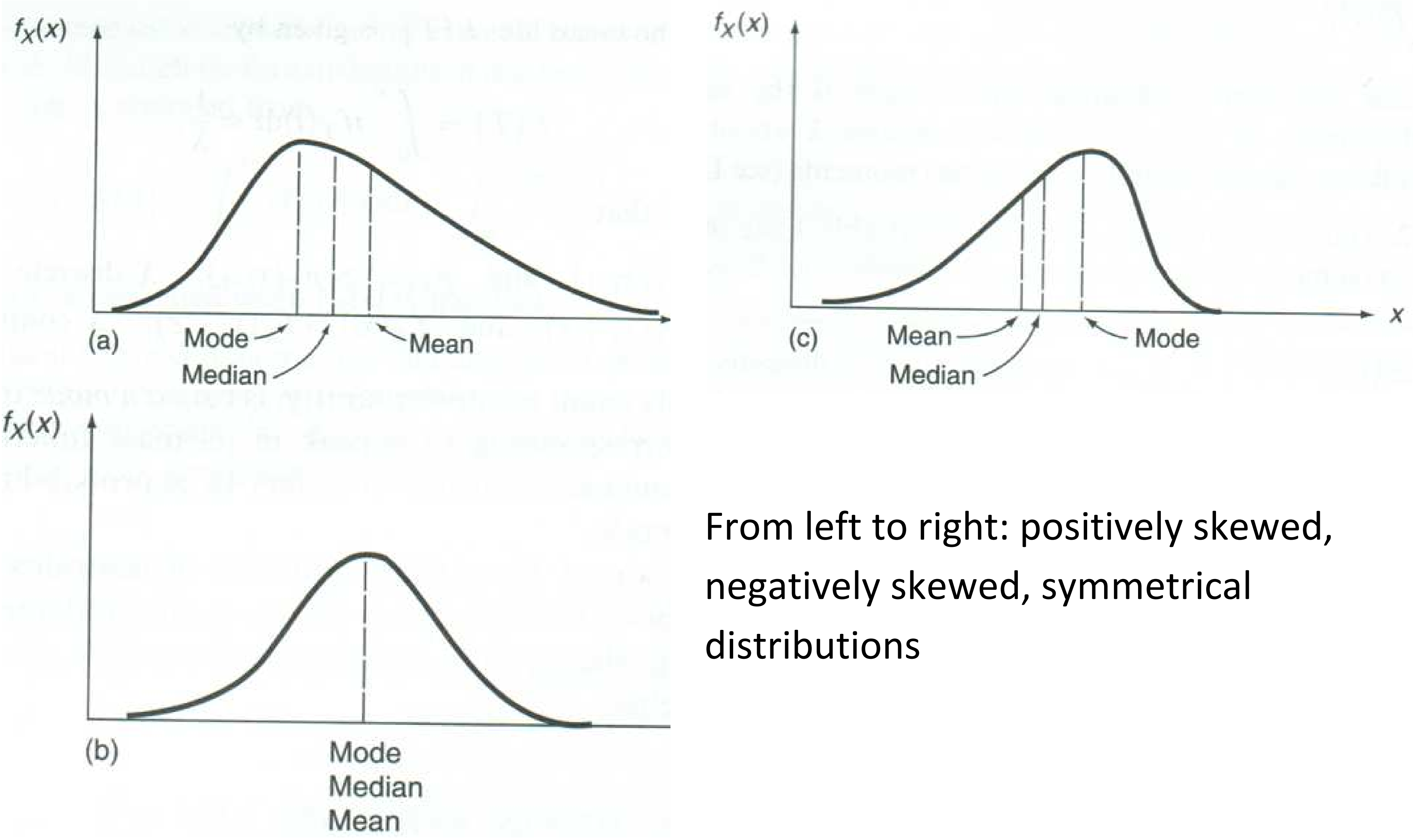
$$\left.\begin{aligned}
&E\{c\} = c,\\
&E\{cg(X)\} = cE\{g(X)\},\\
&E\{g(X) + h(X)\} = E\{g(X)\} + E\{h(X)\},\\
&E\{g(X)\} \leq E\{h(X)\}, \quad \text{if } g(X) \leq h(X) \text{ for all values of } X.
\end{aligned}\right\}$$

Proofs are easy. For example, in the 3rd scenario and continuous case :



$$\begin{aligned}
E\{g(X) + h(X)\} &= \int_{-\infty}^{\infty} [g(x) + h(x)] f_X(x) \mathrm{d}x\\
&= \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x + \int_{-\infty}^{\infty} h(x) f_X(x) \mathrm{d}x\\
&= E\{g(X)\} + E\{h(X)\},
\end{aligned}$$

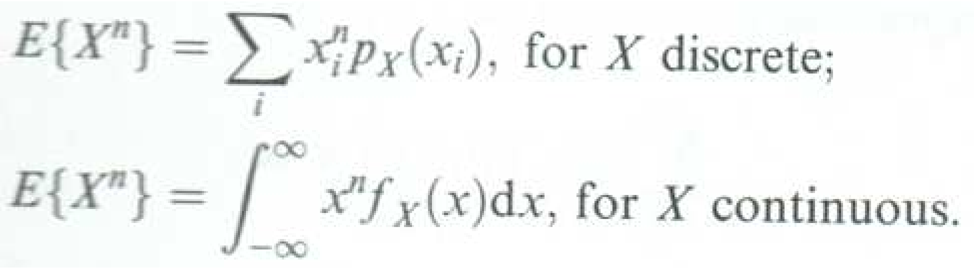
- Two other *measures of centrality* of a random variable:
    - A **median** of X is any point that divides the mass of its distribution into two equal parts → think about our quantile discussion
    - A **mode** is any value of X corresponding to a peak in its mass function or density function



From left to right: positively skewed, negatively skewed, symmetrical distributions

Moments of a single random variable

- Let $g(X) = X^n, n = 1, 2, ...$; the expectation $E(X^n))$, when it exists, is called the **nth moment** of X and denoted by μ'_n:



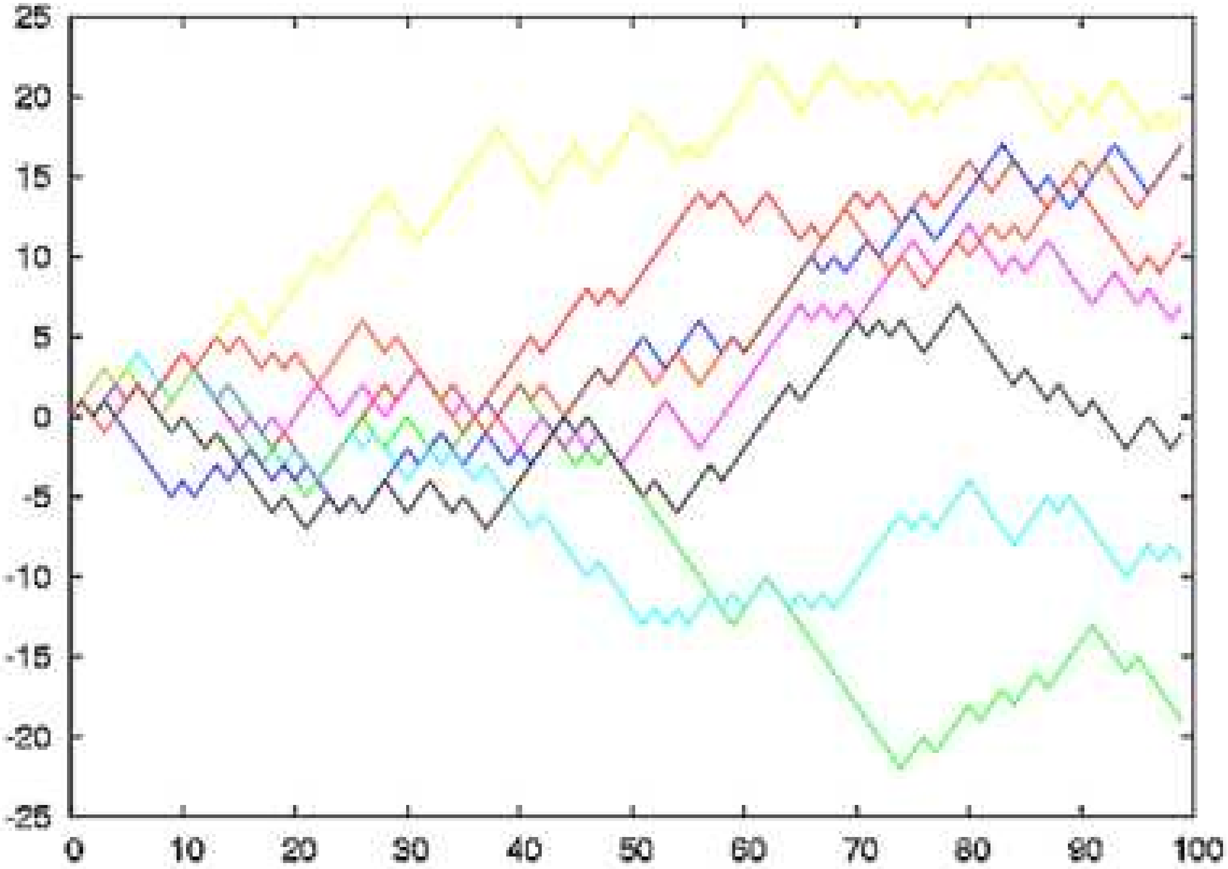
$$E\{X^n\} = \sum_i x_i^n p_X(x_i), \text{ for } X \text{ discrete;}$$

$$E\{X^n\} = \int_{-\infty}^{\infty} x^n f_X(x) \mathrm{d}x, \text{ for } X \text{ continuous.}$$

- The **first moment of X** is also called the **mean**, expectation, average value of X and is a measure of centrality

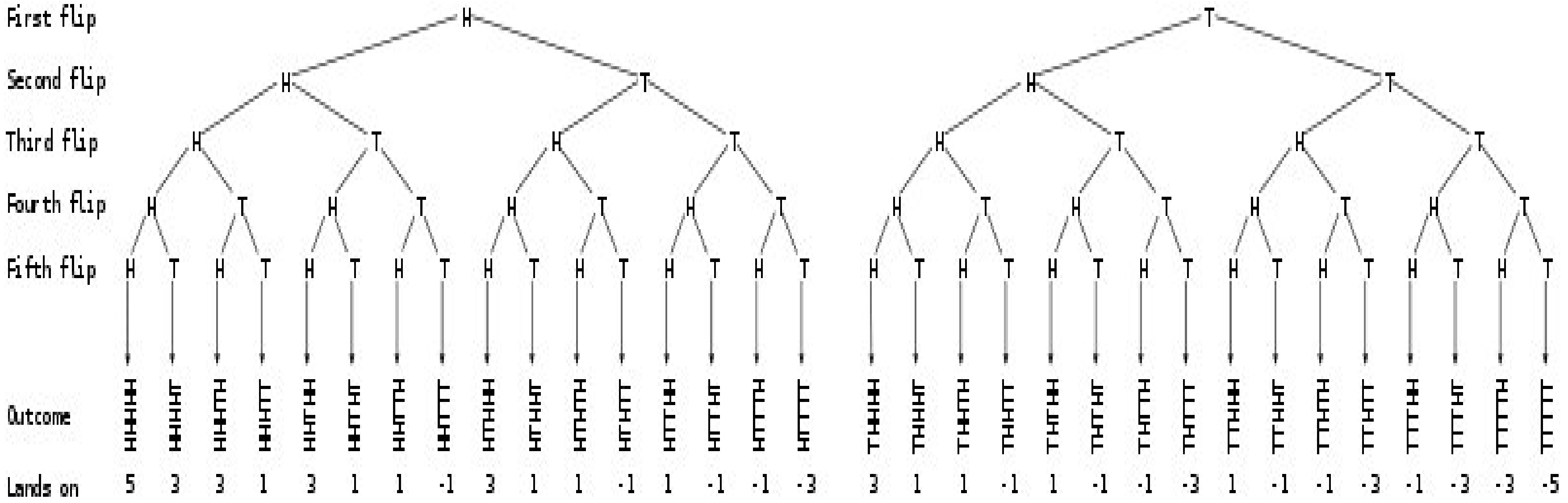
A one-dimensional random walk – read at home

- An elementary example of a random walk is the random walk on the integer number line, which starts at 0 and at each step moves +1 or −1 with equal probability.
- This walk can be illustrated as follows: A marker is placed at zero on the number line and a fair coin is flipped. If it lands on heads, the marker is moved one unit to the right. If it lands on tails, the marker is moved one unit to the left. After five flips, it is possible to have landed on 1, −1, 3, −3, 5, or −5. With five flips, three heads and two tails, in any order, will land on 1. There are 10 ways of landing on 1 or −1 (by flipping three tails and two heads), 5 ways of landing on 3 (by flipping four heads and one tail), 5 ways of landing on −3 (by flipping four tails and one head), 1 way of landing on 5 (by flipping five heads), and 1 way of landing on −5 (by flipping five tails).



- Example of eight random walks in one dimension starting at 0. The plot shows the current position on the line (vertical axis) versus the time steps (horizontal axis).

- See the figure below for an illustration of the possible outcomes of 5 flips.


First flip
Second flip
Third flip
Fourth flip
Fifth flip
Outcome
Lands on
H
T
HHHHH HHHHT HHHTH HHHTT HHTHH HHTHT HHTTH HHTTT HTHHH HTHHT HTHTH HTHTT HTTHH HTTHT HTTTH HTTTT
THHHH THHHT THHTH THHTT THTHH THTHT THTTH THTTT TTHHH TTHHT TTHTH TTHTT TTTHH TTTHT TTTTH TTTTT
5 3 3 1 3 1 1 -1 3 1 1 -1 1 -1 -1 -3
3 1 1 -1 1 -1 -1 -3 1 -1 -1 -3 -1 -3 -3 -5

- To define this walk formerly, take independent random variables $Z_1, Z_2, \ldots$ where each variable is either 1 or -1 with a 50% probability for either value, and set $S_0 = 0$ and $S_n = \sum_{j=1}^{n} Z_j$. The series is called the simple random walk on $\mathbb{Z}$. This series of 1’s and -1’s gives the distance walked, if each part of the walk is of length 1.

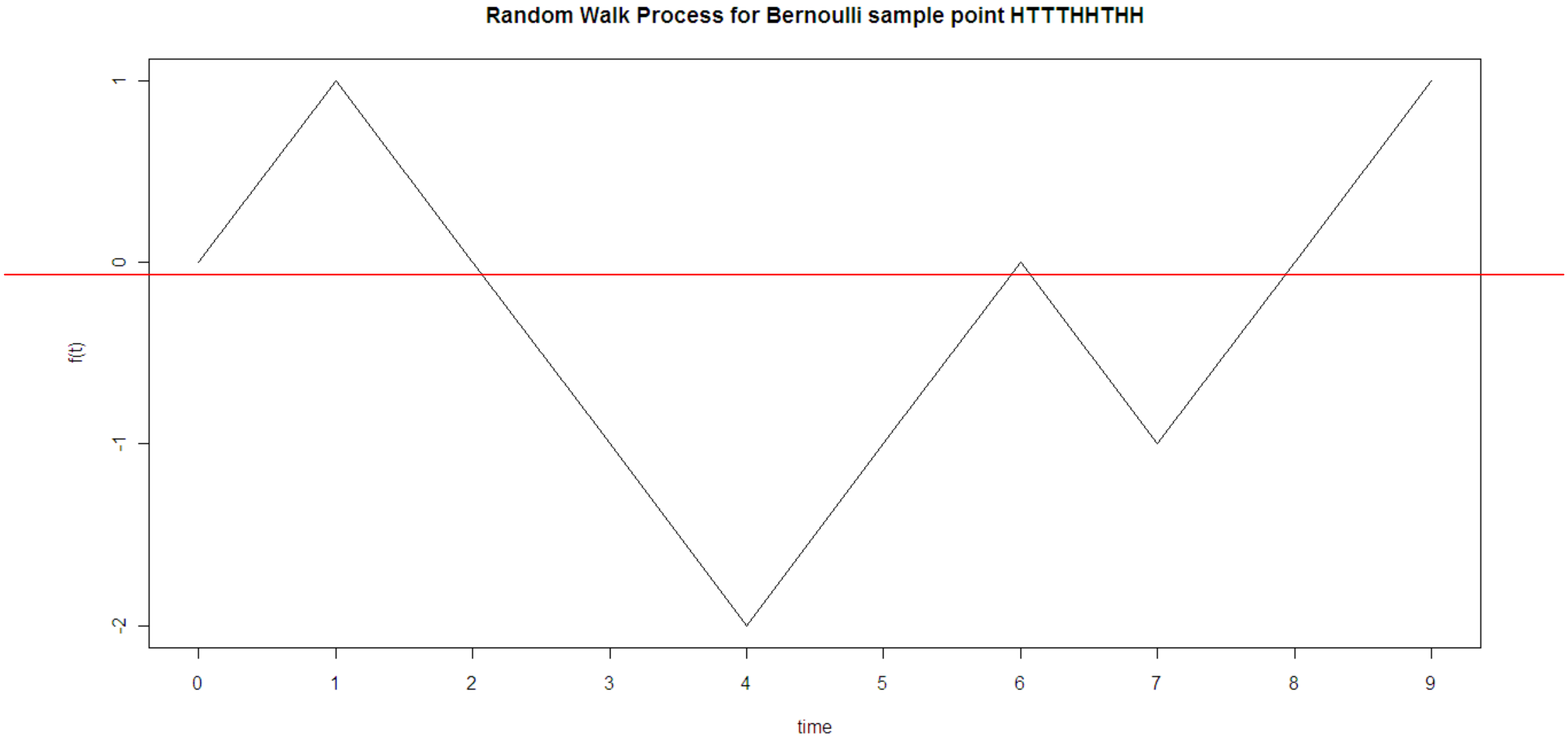
- The expectation $E(S_n)$ of S_n is 0. That is, the mean of all coin flips approaches zero as the number of flips increase. This also follows by the finite additivity property of expectations:

$$E(S_n) = \sum_{j=1}^{n} E(Z_j) = 0.$$

- A similar calculation, using independence of random variables and the fact that $E(Z_n^2) = 1,$ shows that

$$E(S_n^2) = \sum_{j=1}^{n} E(Z_j^2) = n.$$

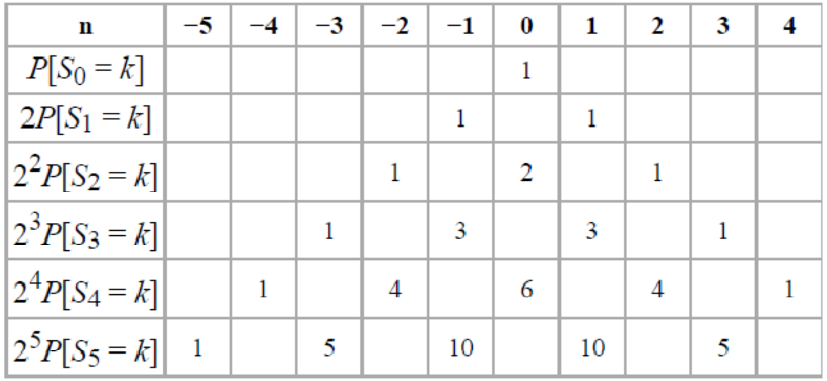
- This hints that $E(|S_n|)$, the expected translation distance after n steps, should be of the order of $\sqrt{n}$.



- Suppose we draw a line some distance from the origin of the walk. How many times will the random walk cross the line?

- The following, perhaps surprising, theorem is the answer: for any random walk in one dimension, every point in the domain will almost surely be crossed an infinite number of times. [In two dimensions, this is equivalent to the statement that any line will be crossed an infinite number of times.] This problem has many names: the *level-crossing problem*, the *recurrence* problem or the *gambler's ruin* problem.
- The source of the last name is as follows: if you are a gambler with a finite amount of money playing *a fair game* against a bank with an infinite amount of money, you will surely lose. The amount of money you have will perform a random walk, and it will almost surely, at some time, reach 0 and the game will be over.

- At zero flips, the only possibility will be to remain at zero. At one turn, you can move either to the left or the right of zero: there is one chance of landing on -1 or one chance of landing on 1. At two turns, you examine the turns from before. If you had been at 1, you could move to 2 or back to zero. If you had been at -1, you could move to -2 or back to zero. So, f.i. there are two chances of landing on zero, and one chance of landing on 2. If you continue the analysis of probabilities, you can see *Pascal's triangle*



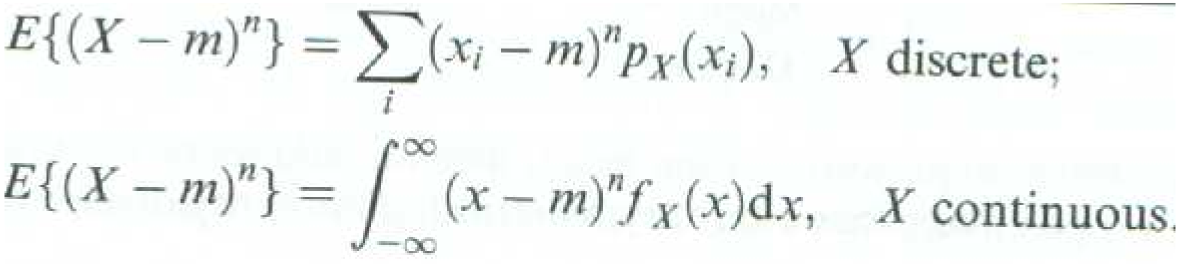
n	−5	−4	−3	−2	−1	0	1	2	3	4
$P[S_0 = k]$						1				
$2P[S_1 = k]$					1		1			
$2^2P[S_2 = k]$				1		2		1		
$2^3P[S_3 = k]$			1		3		3		1	
$2^4P[S_4 = k]$		1		4		6		4		1
$2^5P[S_5 = k]$	1		5		10		10		5	

$$P[S_n = k] = \frac{1}{2^n}\binom{n}{(n+k)/2}$$

5.2 Variance and standard deviation

Central moments

- The central moments of a random variable X are the moments of X with respect to its mean. So the **nt central moment** of X, denoted as μ_n, is defined as

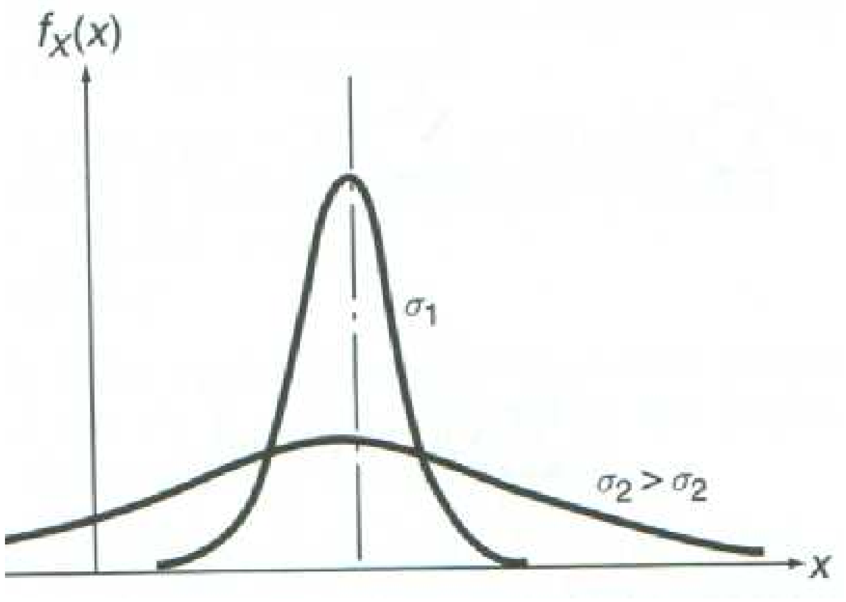


$$E\{(X-m)^n\} = \sum_i (x_i - m)^n p_X(x_i), \quad X \text{ discrete;}$$

$$E\{(X-m)^n\} = \int_{-\infty}^{\infty} (x-m)^n f_X(x)\mathrm{d}x, \quad X \text{ continuous.}$$

- The **variance** of X is the second central moment and usually denoted as σ_X^2 or Var(X). It is the most common *measure of dispersion* of a distribution about its mean, and by definition always nonnegative.
- Important properties of the variance of a random variable X include: $Var(X+c) = Var(X)$ and $Var(cX) = c^2Var(X)$

- The **standard deviation** of X, another such measure of dispersion, is the square root of Var(X) and often denoted by σ_X.
- One of the advantages of using σ_X instead of σ_X^2 is that the standard deviation of X has the same unit as the mean. It can therefore be compared with the mean on the same scale to gain some measure of the degree of spread of the distribution.
- A dimensionless number that characterizes dispersion relative to the mean which also facilitates comparison among random variables of different units is the **coefficient of variation**, defined by $\dfrac{\sigma_X}{\mu_1'(X)}$



Relation between variance and simple moments

$$\sigma^2 = \mu_2' - (\mu_1')^2$$

Indeed, with $m = \mu_1'$:

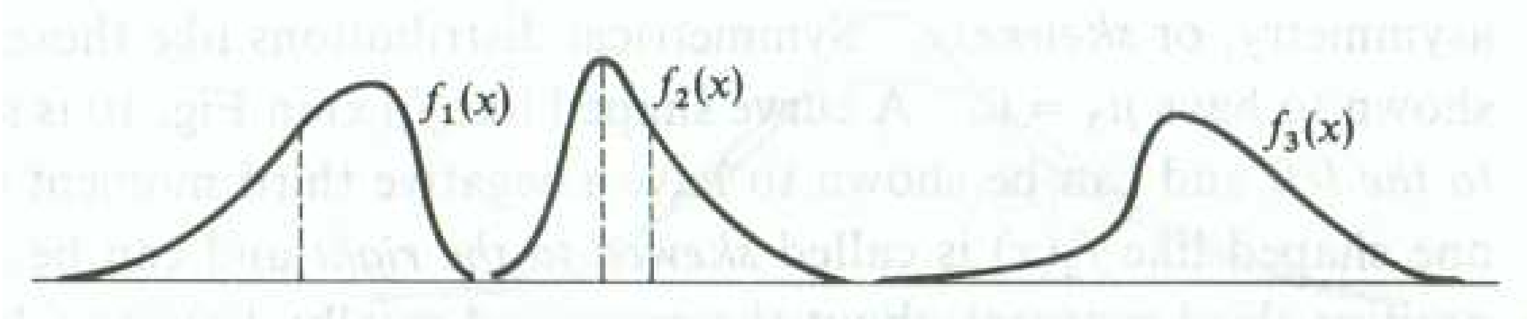
$\sigma^2 = E((X - m)^2) = E(X^2 - 2mX + x^2) = E(X^2) - 2mE(X) + m^2 = \mu_2' - 2m^2 + m^2 = \mu_2' - m^2$

- Hence, there are two ways of computing variances …,
    - using the original definition, or
    - using the relation to the first and second simple moments

Third central moment

- The third moment about the mean is sometimes called a measure of asymmetry, or **skewness**
- The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero.
- Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.
- By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail.
- Some measurements have a lower bound and are thus skewed right. For example, in reliability studies, failure times cannot be negative.

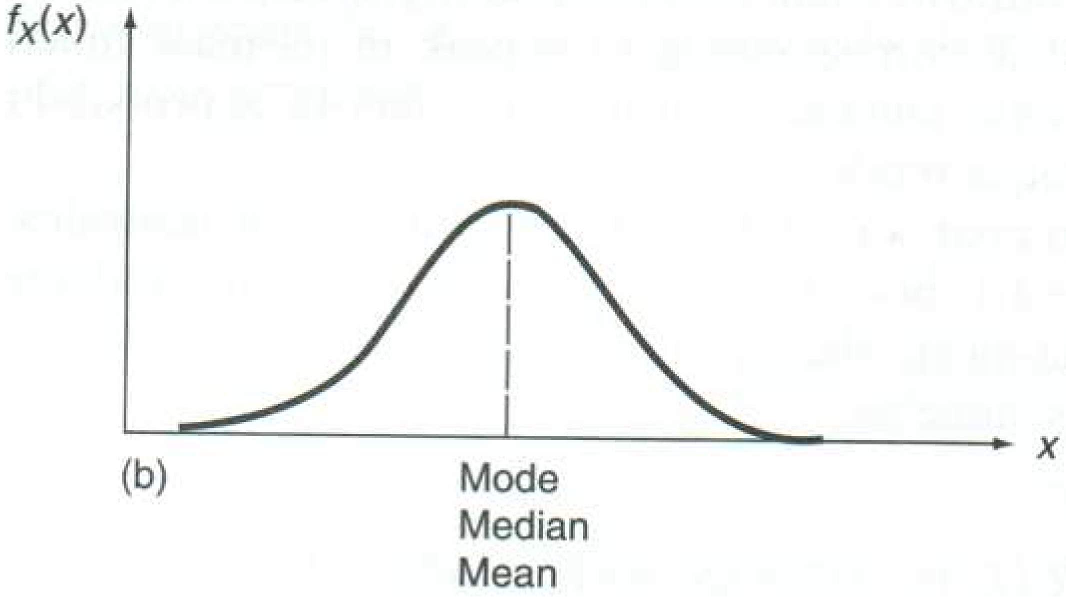
- Knowledge of the third moment hardly gives a clue about the shape of the distribution … (e.g., $f_3(x)$ is far from symmetrical but its third moment is zero)



- The ratio

$$\mu_3/\sigma^3$$

  is called the **coefficient of skewness** and is unitless.



The quantity $\mathscr{s} = (\text{mean} - \text{median})/(\text{standard deviation})$ provides an alternative measure of skewness. It can be proved that $-1 \le \mathscr{s} \le 1$.

Fourth central moment

- The fourth moment about the mean is sometimes called a measure of excess, or **kurtosis**
- It refers to the degree of flatness of a density near its center, and usually the **coefficient of excess kurtosis** is considered (-3 ensures that the excess is zero for normal distributions):

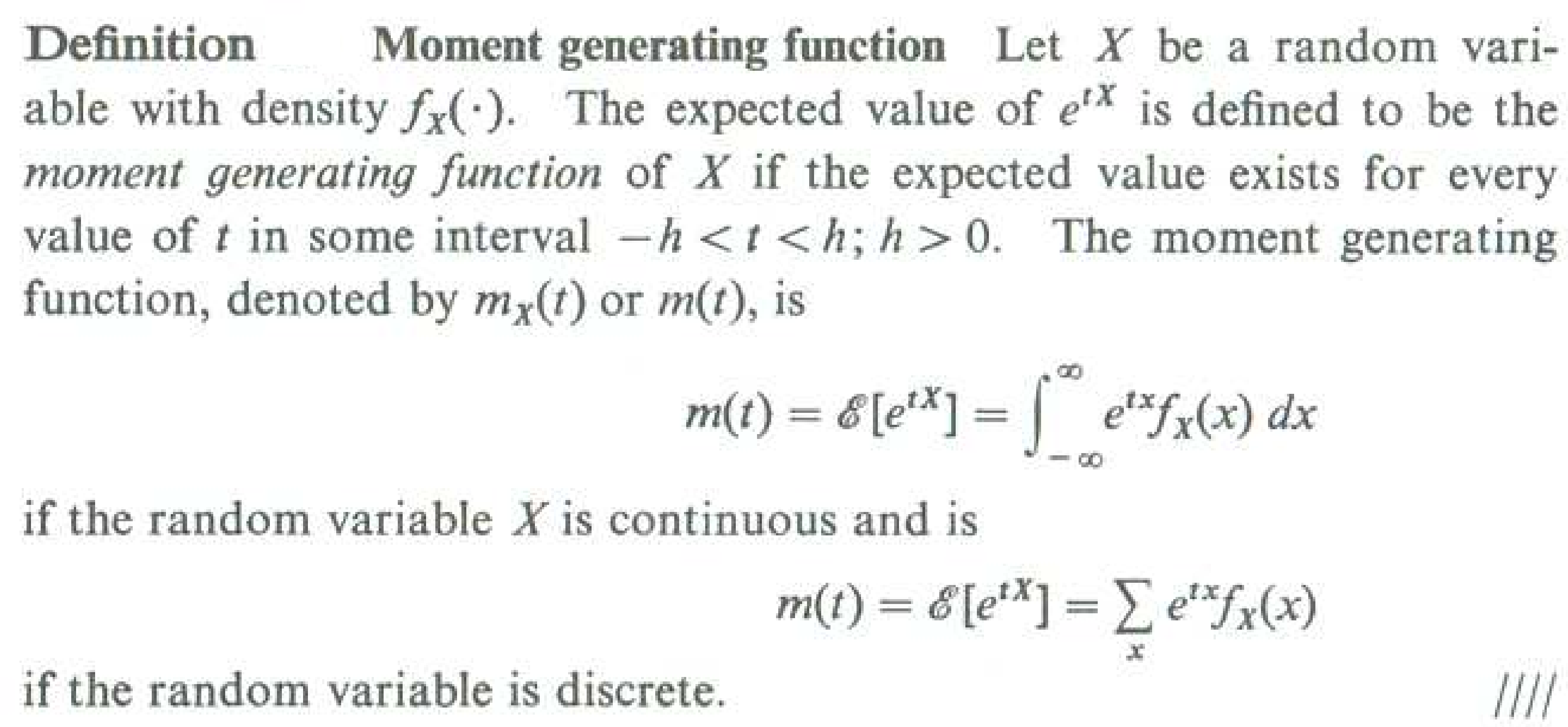
$$\mu_4/\sigma^4 - 3.$$

- A distribution with negative excess kurtosis is called **platykurtic.** A distribution with positive excess kurtosis is called **leptokurtic.** Distributions with zero excess kurtosis are called **mesokurtic**
- This measure however suffers from the same failing as does the measure of skewness: It does not always measure what it is supposed to.

The importance of moments … or not?

- In applied statistics, the first two moments are obviously of great importance. It is usually necessary to know at least the location of the distribution and to have some idea about its dispersion or spread
- These characteristics can be estimated by examining a sample drawn from a set of objects known to have the distribution in question (see future chapters)
- In some cases, if the moments are known, then the density can be determined (e.g., cfr Normal Distribution).
- It would be useful if a function could be found that would give a representation of all the moments. Such a function is called a **moment generating function.**

5.3 Moment generating functions



Definition **Moment generating function** Let X be a random variable with density $f_X(\cdot)$. The expected value of e^{tX} is defined to be the *moment generating function* of X if the expected value exists for every value of t in some interval $-h < t < h$; $h > 0$. The moment generating function, denoted by $m_X(t)$ or $m(t)$, is

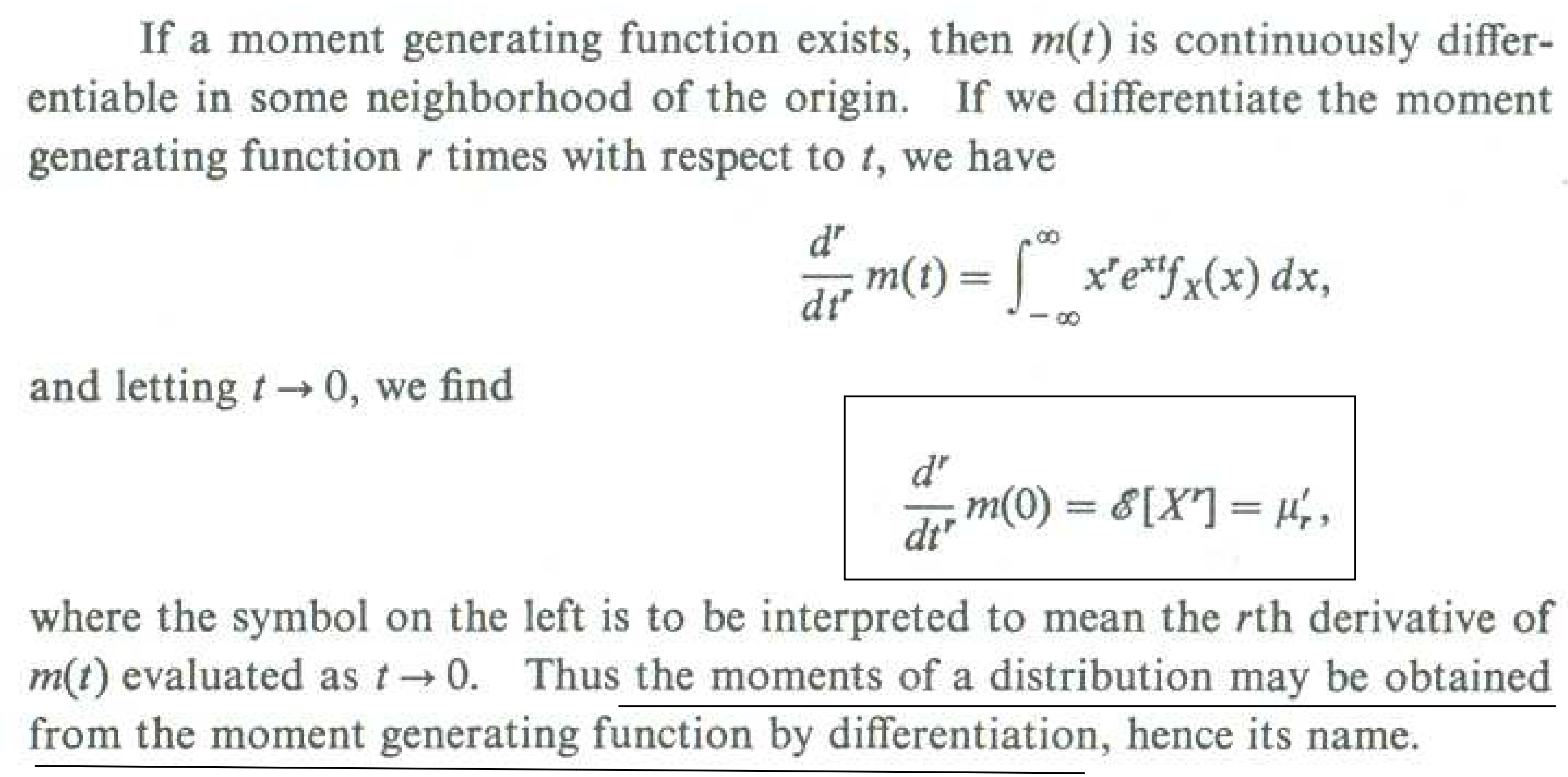
$$m(t) = \mathscr{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$$

if the random variable X is continuous and is

$$m(t) = \mathscr{E}[e^{tX}] = \sum_x e^{tx} f_X(x)$$

if the random variable is discrete. ////

Origin of the name “moment generating function”



If a moment generating function exists, then $m(t)$ is continuously differentiable in some neighborhood of the origin. If we differentiate the moment generating function r times with respect to t, we have

$$\frac{d^r}{dt^r} m(t) = \int_{-\infty}^{\infty} x^r e^{xt} f_X(x)\, dx,$$

and letting $t \to 0$, we find

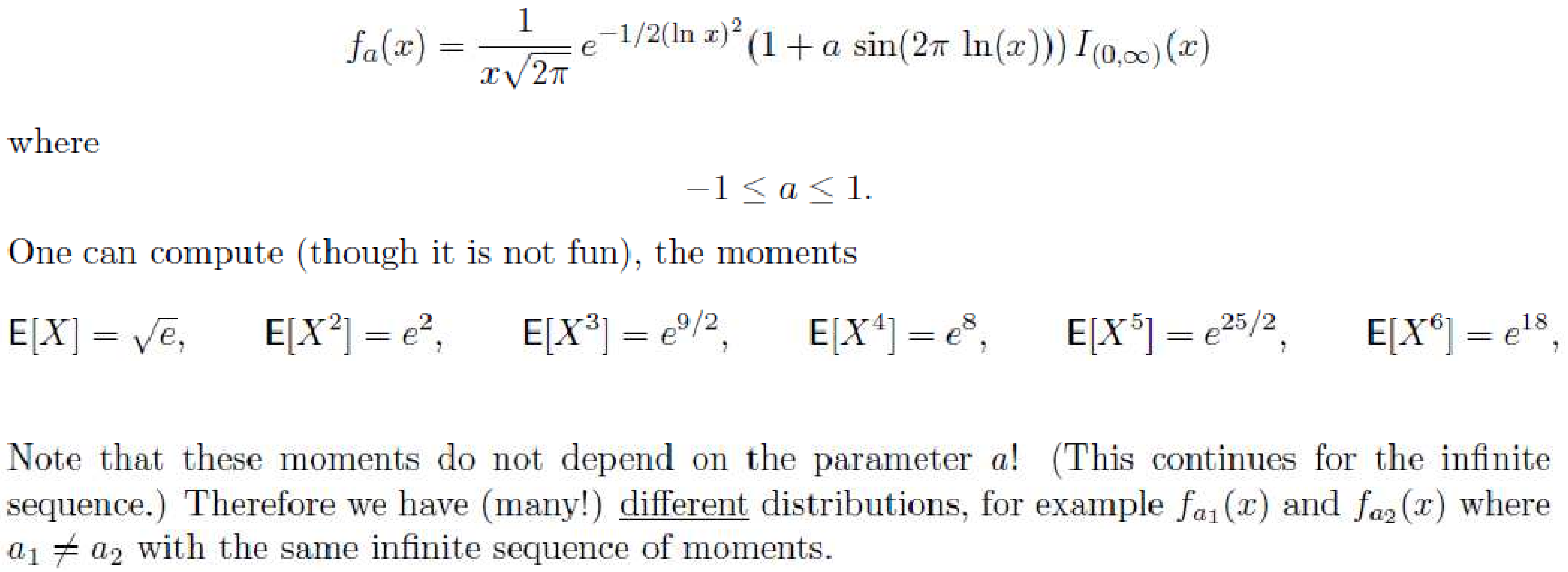
$$\boxed{\frac{d^r}{dt^r} m(0) = \mathscr{E}[X^r] = \mu'_r ,}$$

where the symbol on the left is to be interpreted to mean the rth derivative of $m(t)$ evaluated as $t \to 0$. Thus the moments of a distribution may be obtained from the moment generating function by differentiation, hence its name.

Importance of moment generating functions

- In principle ***it is possible that there exists a sequence of moments for which there is a large collection of different distributions functions having these same moments*** → so a sequence of moments does not determine uniquely the corresponding distribution function …

For example:



$$f_a(x) = \frac{1}{x\sqrt{2\pi}} e^{-1/2(\ln x)^2}(1 + a \sin(2\pi \ln(x)))\, I_{(0,\infty)}(x)$$

where

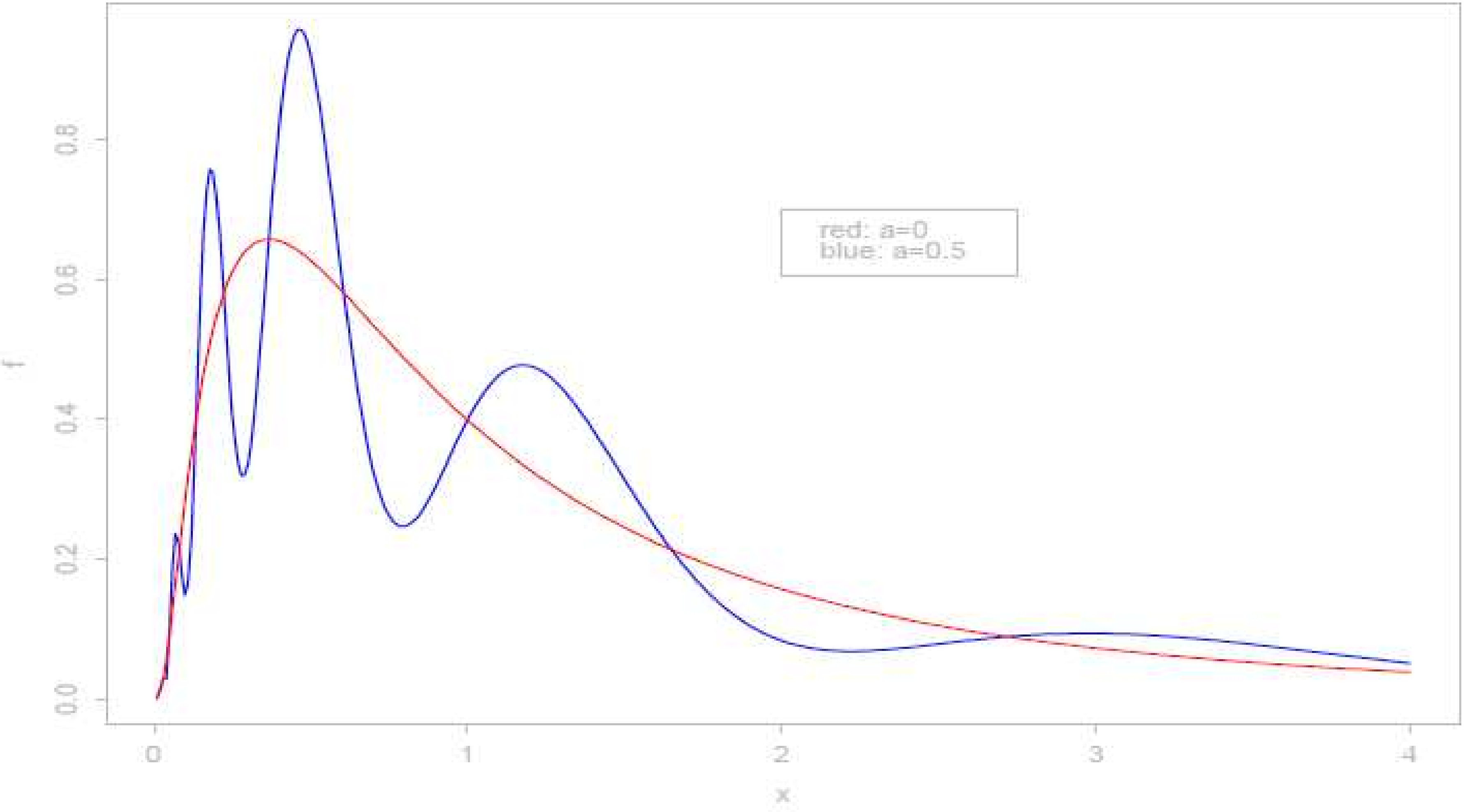
$$-1 \le a \le 1.$$

One can compute (though it is not fun), the moments

$$\mathsf{E}[X] = \sqrt{e}, \qquad \mathsf{E}[X^2] = e^2, \qquad \mathsf{E}[X^3] = e^{9/2}, \qquad \mathsf{E}[X^4] = e^8, \qquad \mathsf{E}[X^5] = e^{25/2}, \qquad \mathsf{E}[X^6] = e^{18},$$

Note that these moments do not depend on the parameter a! (This continues for the infinite sequence.) Therefore we have (many!) different distributions, for example $f_{a_1}(x)$ and $f_{a_2}(x)$ where $a_1 \neq a_2$ with the same infinite sequence of moments.

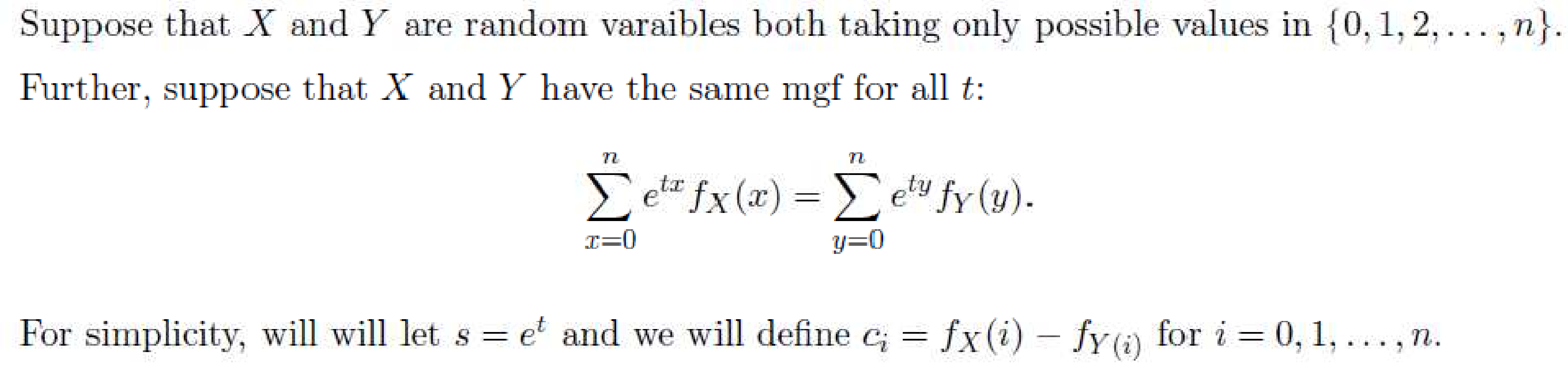
- Densities for two distributions with the SAME infinite series of moments



- Is there any moment criterion for identifying distributions that would ensure that two distributions are identical?

  Yes:
  If random variables X and Y both have moment generating functions $M_X(t)$ and $M_Y(t)$ that exist in some neighborhood of zero and if they are equal for all t in this neighborhood, then X and Y have the same distributions!

  <u>“Simple proof” of a special case:</u>



Suppose that X and Y are random varaibles both taking only possible values in $\{0, 1, 2, \dots, n\}$.

Further, suppose that X and Y have the same mgf for all t:

$$\sum_{x=0}^{n} e^{tx} f_X(x) = \sum_{y=0}^{n} e^{ty} f_Y(y).$$

For simplicity, will will let $s = e^t$ and we will define $c_i = f_X(i) - f_{Y(i)}$ for $i = 0, 1, \dots, n$.

Now

$$\sum_{x=0}^{n} e^{tx} f_X(x) - \sum_{y=0}^{n} e^{ty} f_Y(y) = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} f_X(x) - \sum_{y=0}^{n} s^{y} f_Y(y) = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} f_X(x) - \sum_{x=0}^{n} s^{x} f_Y(x) = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} \left[f_X(x) - f_Y(x) \right] = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} c_x = 0 \qquad \forall\ s > 0$$

The above is simply a polynomial in s with coefficients $c_0, c_1, \ldots, c_n$. The only way it can be zero for all values of s is if $c_0 = c_1 = \cdots = c_n = 0$.

So, we have that

$$0 = c_i = f_X(i) - f_Y(i) \qquad \text{for } i = 0, 1, \ldots, n.$$

Therefore

$$f_X(i) = f_Y(i) \qquad \text{for } i = 0, 1, \ldots, n.$$

In other words the density functions for X and Y are exactly the same. In other *other* words, X and Y have the same distributions!

6 Functions of random variables

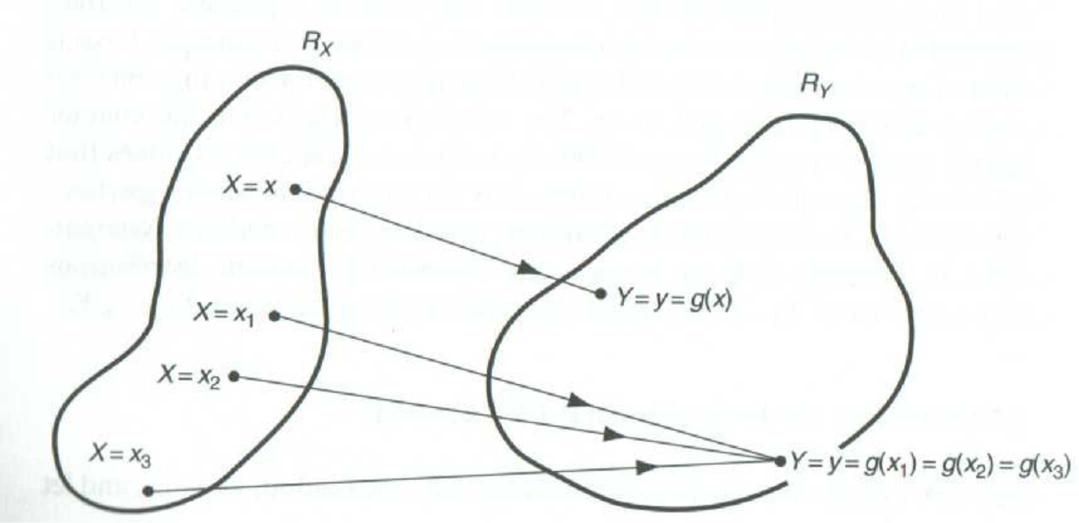
6.1 Functions of one random variable

- Real-life examples often present themselves with far more complex density functions that the one described so far.
- In many cases the random variable of interest is a function of one that we know better, or for which we are better able to describe its density or distributional properties
- For this reason, we devote an entire part on densities of "functions of random variables". We first assume a random variable X and Y=g(X), with g(X) a continuous function of X.
    - How does the corresponding distribution for Y look like?
    - What are its moment properties?

Discrete random variables

- Suppose that the possible values taken by X can be enumerated as $x_1, x_2, \ldots$. Then the corresponding possible values of Y can be enumerated as $y_1 = g(x_1), y_2 = g(x_2), \ldots$.
- Let the probability mass function of X be given by $P_X(x_i) = p_i, i = 1, 2, \ldots$ then the probability mass function of Y is determined as

$$P_Y(y_i) = P_Y(g(x_i)) = p_i, i = 1, 2, \ldots$$



Continuous random variables

- To carry out similar mapping steps as outlined for discrete random variables, care must be exercised in choosing appropriate corresponding regions in ranges spaces $R_X \text{ and } R_Y$
- For strictly monotone increasing functions of x $(g(x_2) > g(x_1), \text{ whenever } x_2 > x_1)$:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

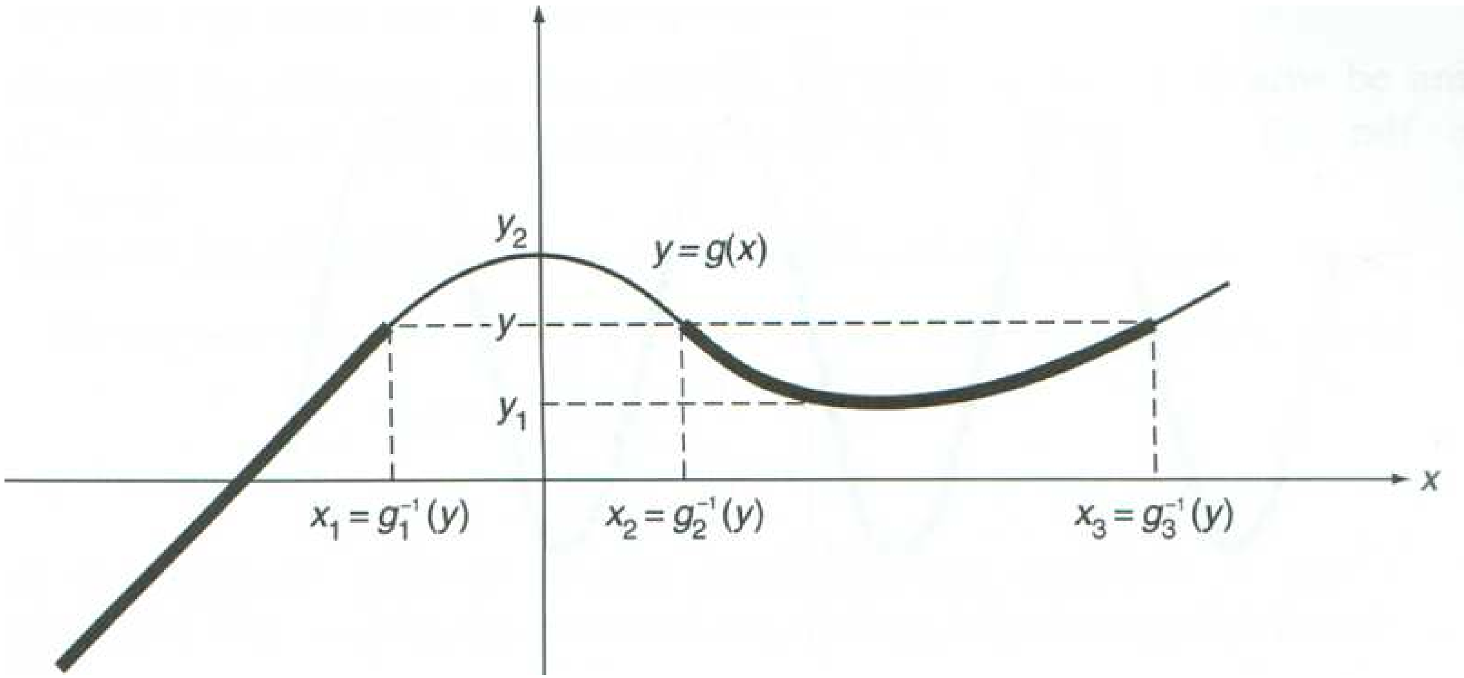
By differentiating both sides:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\{F_X(g^{-1}(y))\} = f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy}$$

- In general, for X a continuous random variable and Y=g(X), with g(X) continuous in X and strictly monotone, $f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|$

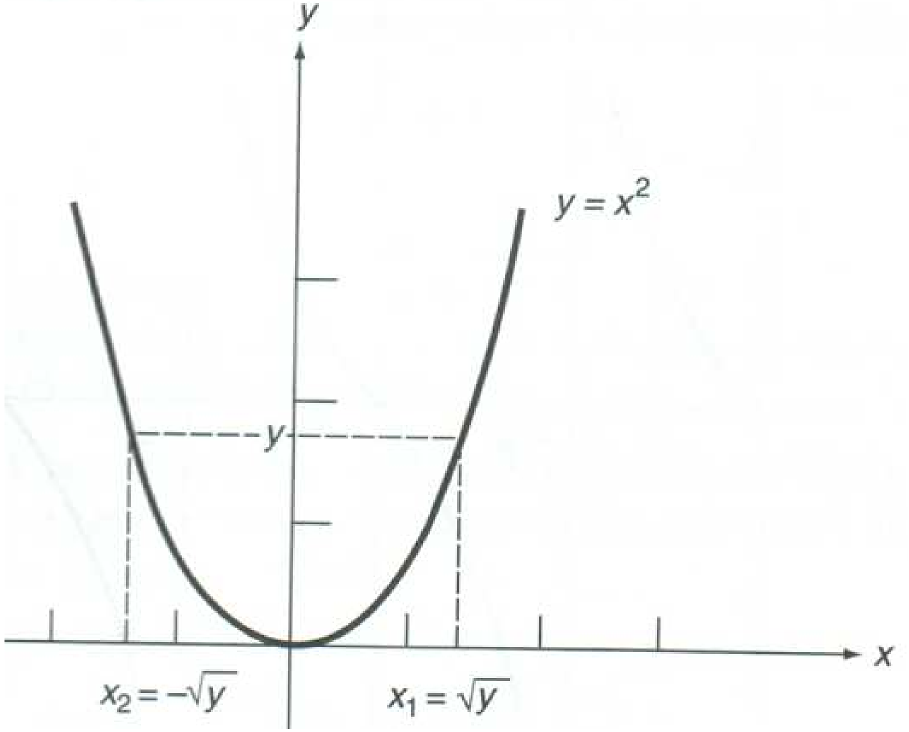
- Let X be a continuous random variable and Y=g(X), where g(X) is continuous in X, and y=g(x) admits at most a countable number (r) of roots $x_1 = g_1^{-1}(y), x_2 = g_2^{-1}(y), ...$ then

$$f_Y(y) = \sum_{j=1}^{r} f_X(g_j^{-1}(y))|\frac{dg_j^{-1}(y)}{dy}|$$



Example of Y=X^2 and X normally distributed, leading to $\chi^2(1)$-distribution

$$f_X(x) = \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}, \quad -\infty < x < \infty.$$



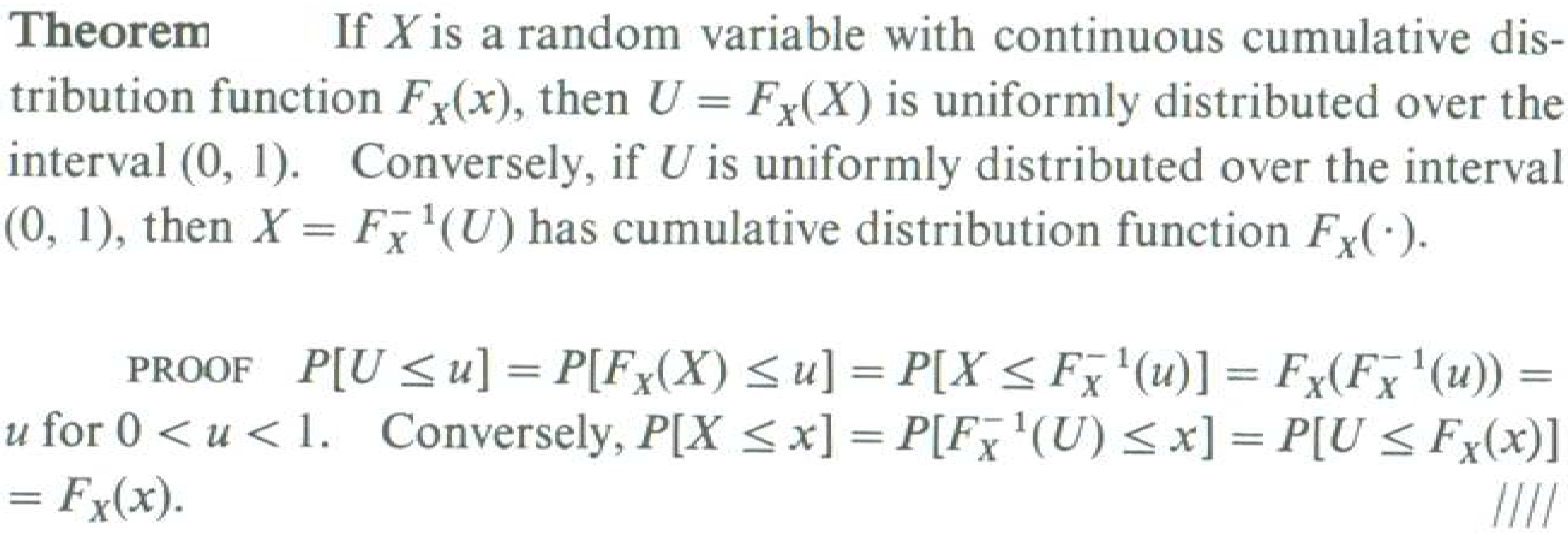
$$f_Y(y) = \sum_{j=1}^{2} f_X[g_j^{-1}(y)] \left| \frac{dg_j^{-1}(y)}{dy} \right|$$

$$= \frac{f_X(-y^{1/2})}{2y^{1/2}} + \frac{f_X(y^{1/2})}{2y^{1/2}}$$

$$= \frac{1}{(2\pi y)^{1/2}} e^{-y/2},$$

$$f_Y(y) = \begin{cases} \dfrac{1}{(2\pi y)^{1/2}} e^{-y/2}, & \text{for } y \geq 0; \\ 0, & \text{elsewhere.} \end{cases}$$

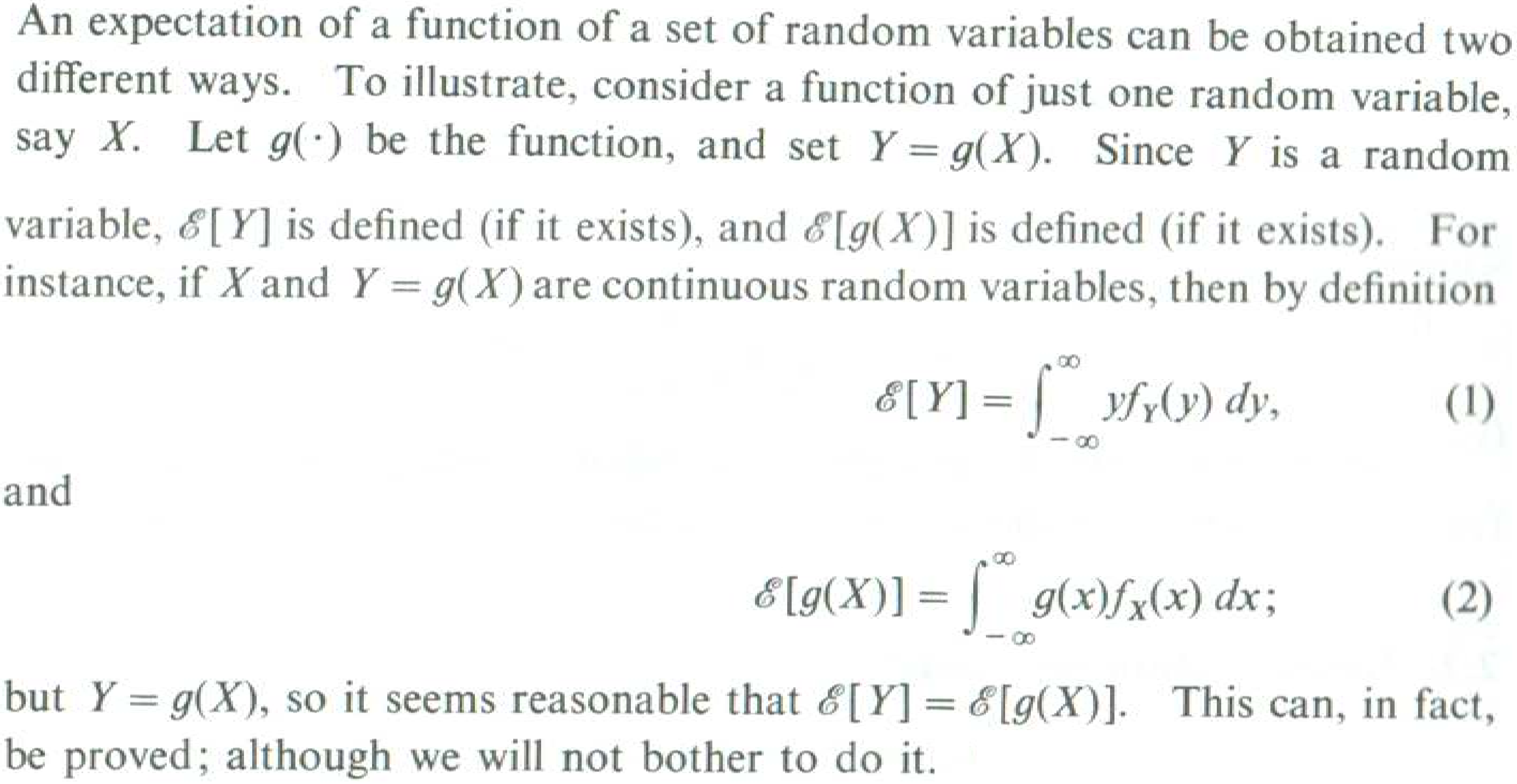
Recall: the probability integral transform



Theorem If X is a random variable with continuous cumulative distribution function $F_X(x)$, then $U = F_X(X)$ is uniformly distributed over the interval (0, 1). Conversely, if U is uniformly distributed over the interval (0, 1), then $X = F_X^{-1}(U)$ has cumulative distribution function $F_X(\cdot)$.

PROOF $P[U \leq u] = P[F_X(X) \leq u] = P[X \leq F_X^{-1}(u)] = F_X(F_X^{-1}(u)) = u$ for $0 < u < 1$. Conversely, $P[X \leq x] = P[F_X^{-1}(U) \leq x] = P[U \leq F_X(x)] = F_X(x)$. ////

Two ways to compute expectations



An expectation of a function of a set of random variables can be obtained two different ways. To illustrate, consider a function of just one random variable, say X. Let $g(\cdot)$ be the function, and set $Y = g(X)$. Since Y is a random variable, $\mathscr{E}[Y]$ is defined (if it exists), and $\mathscr{E}[g(X)]$ is defined (if it exists). For instance, if X and $Y = g(X)$ are continuous random variables, then by definition

$$\mathscr{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy, \qquad (1)$$

and

$$\mathscr{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx; \qquad (2)$$

but $Y = g(X)$, so it seems reasonable that $\mathscr{E}[Y] = \mathscr{E}[g(X)]$. This can, in fact, be proved; although we will not bother to do it.

Two ways to compute moments

- Expressing the nth moment of Y as $E(Y^n) = E(g^n(X))$,

$$E\{Y^n\} = E\{g^n(X)\} = \sum_i g^n(x_i) p_X(x_i), \quad X \text{ discrete;}$$

$$E\{Y^n\} = E\{g^n(X)\} = \int_{-\infty}^{\infty} g^n(x) f_X(x) \mathrm{d}x, \quad X \text{ continuous.}$$

- Alternatively, using characteristic functions (here: j is the imaginary unit):

$$\left.\begin{aligned} \phi_Y(t) &= E\{\mathrm{e}^{\mathrm{j}tY}\} = E\{\mathrm{e}^{\mathrm{j}tg(X)}\} = \sum_i \mathrm{e}^{\mathrm{j}tg(x_i)} p_X(x_i), \quad X \text{ discrete;} \\ \phi_Y(t) &= E\{\mathrm{e}^{\mathrm{j}tY}\} = E\{\mathrm{e}^{\mathrm{j}tg(X)}\} = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{j}tg(x)} f_X(x) \mathrm{d}x, \quad X \text{ continuous.} \end{aligned}\right\}$$

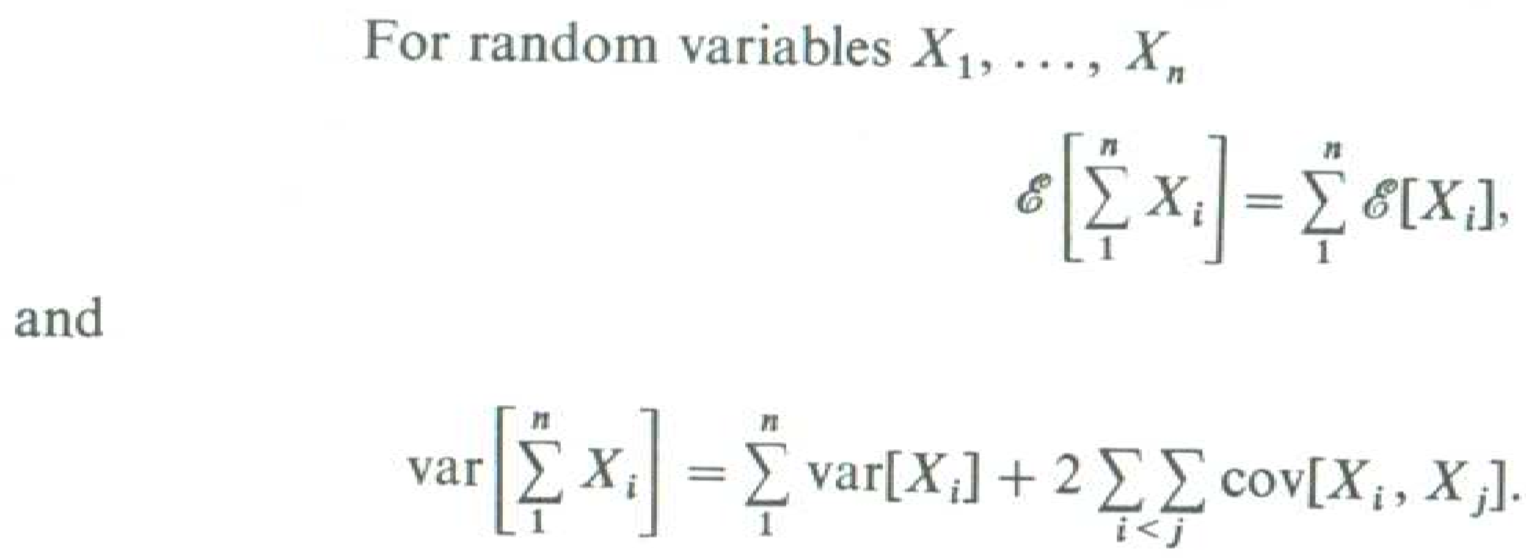
after which the moments of Y are given via taking derivatives:

$$\boxed{E(Y^n) = j^{-n}\, \Phi_Y^{(n)}(0), n = 1, 2, \ldots}$$

- Note that $\boxed{M_Y(t) = \Phi_Y(-jt)}$

6.2 Functions of two or more random variables: sums of random variables

Deriving moments – mean and variances



For random variables $X_1, \ldots, X_n$

$$\mathscr{E}\left[\sum_1^n X_i\right] = \sum_1^n \mathscr{E}[X_i],$$

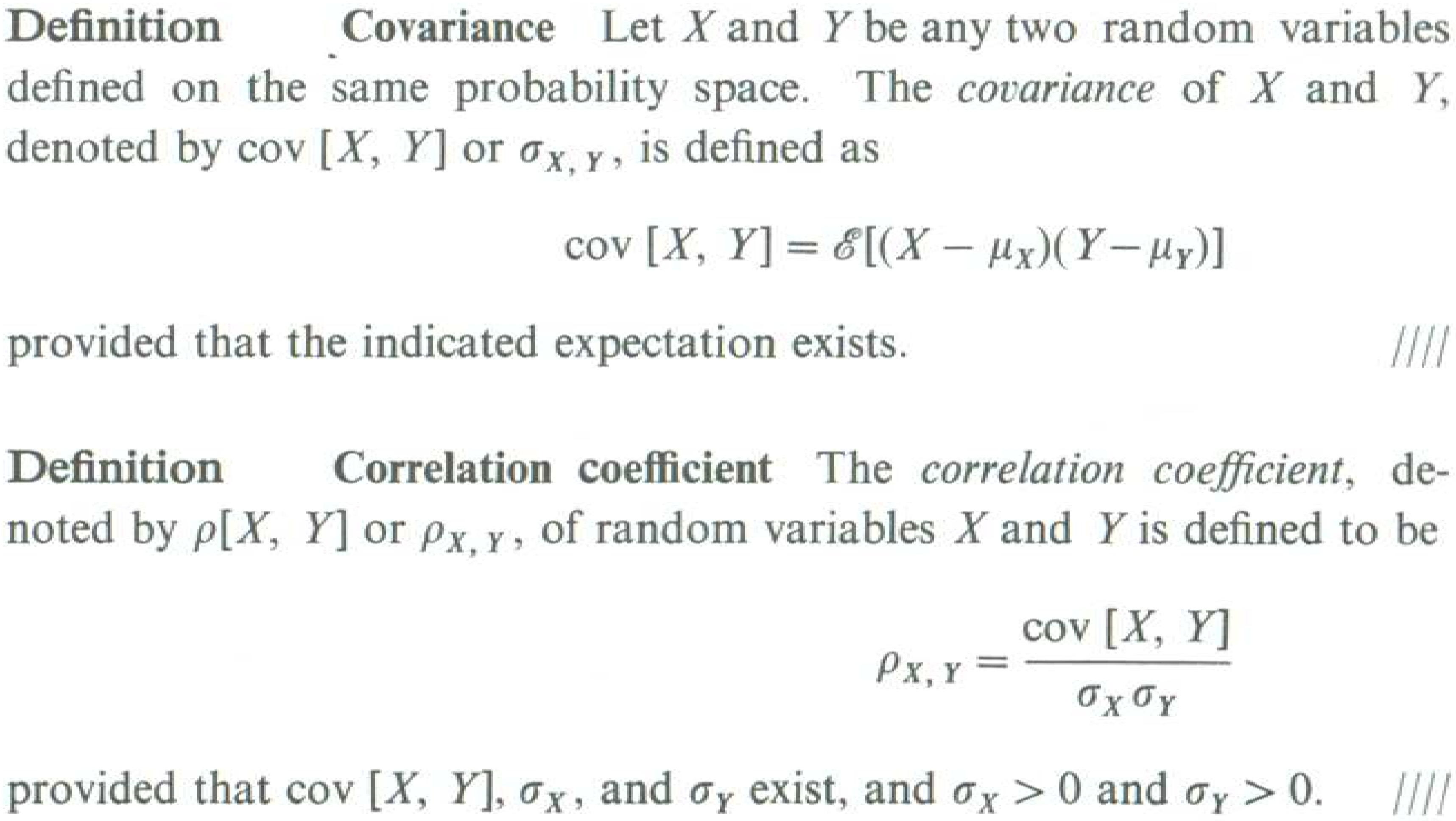
and

$$\operatorname{var}\left[\sum_1^n X_i\right] = \sum_1^n \operatorname{var}[X_i] + 2\sum_{i<j}\sum \operatorname{cov}[X_i, X_j].$$

PROOF That $\mathscr{E}\left[\sum_1^n X_i\right] = \sum_1^n \mathscr{E}[X_i]$ follows from a property of expectation

$$\begin{aligned}
\operatorname{var}\left[\sum_1^n X_i\right] &= \mathscr{E}\left[\left(\sum_1^n X_i - \mathscr{E}\left[\sum_1^n X_i\right]\right)^2\right] = \mathscr{E}\left[\left(\sum_1^n (X_i - \mathscr{E}[X_i])\right)^2\right] \\
&= \mathscr{E}\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - \mathscr{E}[X_i])(X_j - \mathscr{E}[X_j])\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathscr{E}[(X_i - \mathscr{E}[X_i])(X_j - \mathscr{E}[X_j])] \\
&= \sum_{i=1}^n \operatorname{var}[X_i] + 2 \sum_{i<j}\sum \underline{\operatorname{cov}[X_i, X_j].} \qquad ////
\end{aligned}$$

Covariances and correlations



Definition **Covariance** Let X and Y be any two random variables defined on the same probability space. The *covariance* of X and Y, denoted by cov $[X, Y]$ or $\sigma_{X,Y}$, is defined as

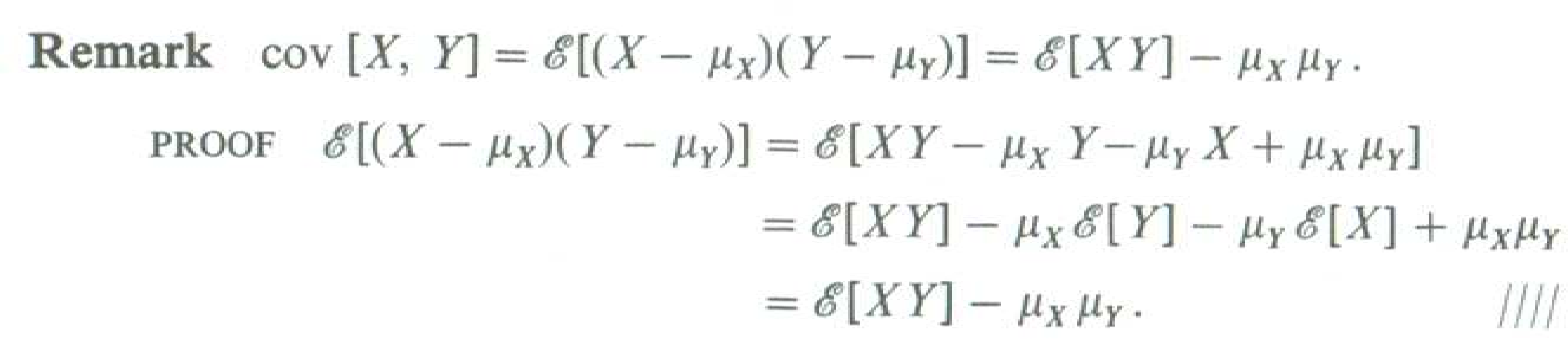
$$\operatorname{cov}[X, Y] = \mathscr{E}[(X - \mu_X)(Y - \mu_Y)]$$

provided that the indicated expectation exists. ////

Definition **Correlation coefficient** The *correlation coefficient*, denoted by $\rho[X, Y]$ or $\rho_{X,Y}$, of random variables X and Y is defined to be

$$\rho_{X,Y} = \frac{\operatorname{cov}[X, Y]}{\sigma_X \sigma_Y}$$

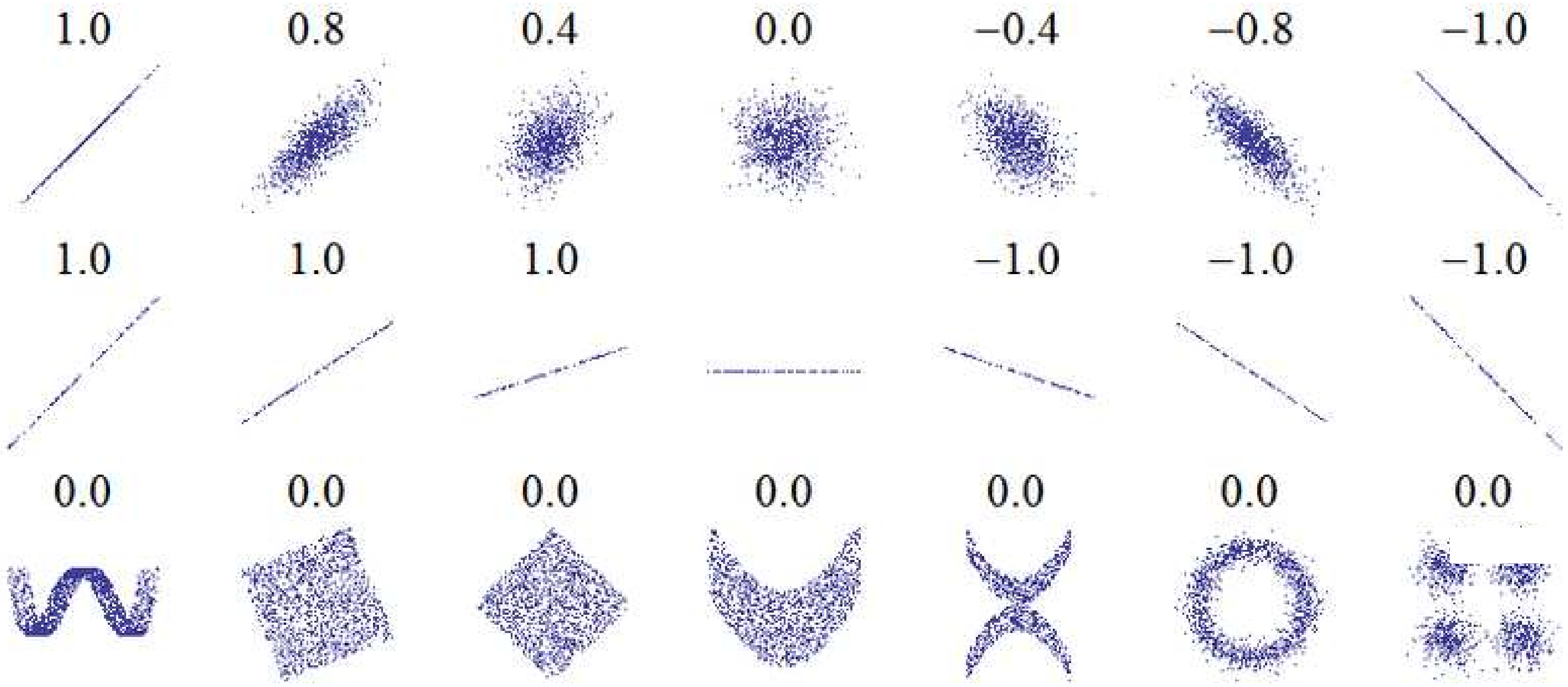
provided that cov $[X, Y]$, σ_X, and σ_Y exist, and $\sigma_X > 0$ and $\sigma_Y > 0$. ////

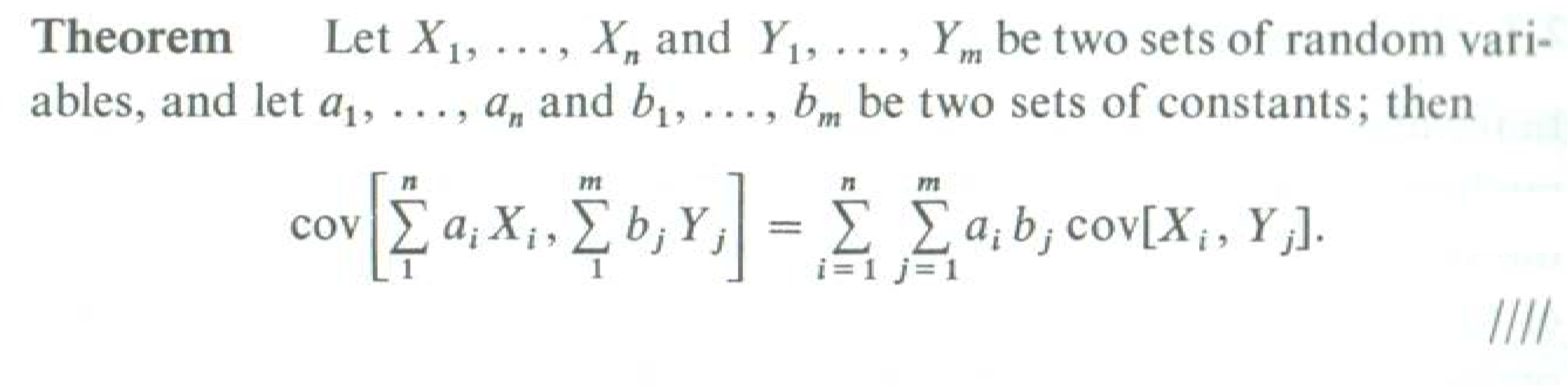


Remark $\operatorname{cov}[X, Y] = \mathscr{E}[(X - \mu_X)(Y - \mu_Y)] = \mathscr{E}[XY] - \mu_X \mu_Y$.

PROOF
$$\begin{aligned} \mathscr{E}[(X - \mu_X)(Y - \mu_Y)] &= \mathscr{E}[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= \mathscr{E}[XY] - \mu_X \mathscr{E}[Y] - \mu_Y \mathscr{E}[X] + \mu_X \mu_Y \\ &= \mathscr{E}[XY] - \mu_X \mu_Y . \qquad //// \end{aligned}$$

- Several sets of (*X*, *Y*) points, with the correlation coefficient of *X* and *Y* for each set, are shown in the following plot. Note that the correlation reflects the noisiness and direction of a *linear relationship* (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom).
- Remark: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero

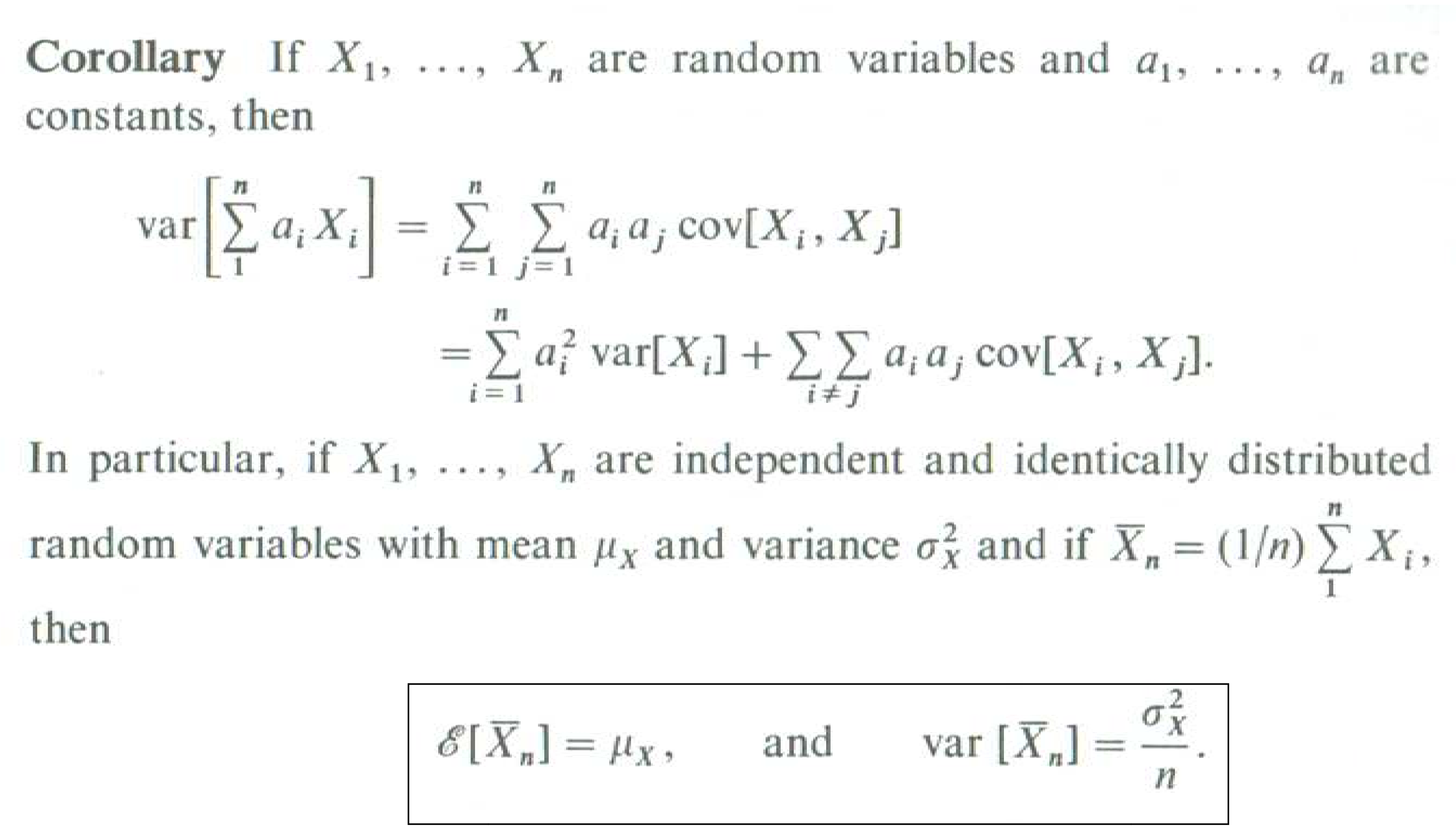

1.0
0.8
0.4
0.0
−0.4
−0.8
−1.0
1.0
1.0
1.0
−1.0
−1.0
−1.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0



Theorem Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two sets of random variables, and let $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ be two sets of constants; then

$$\operatorname{cov}\left[\sum_1^n a_i X_i, \sum_1^m b_j Y_j\right] = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \operatorname{cov}[X_i, Y_j].$$

////



Corollary If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

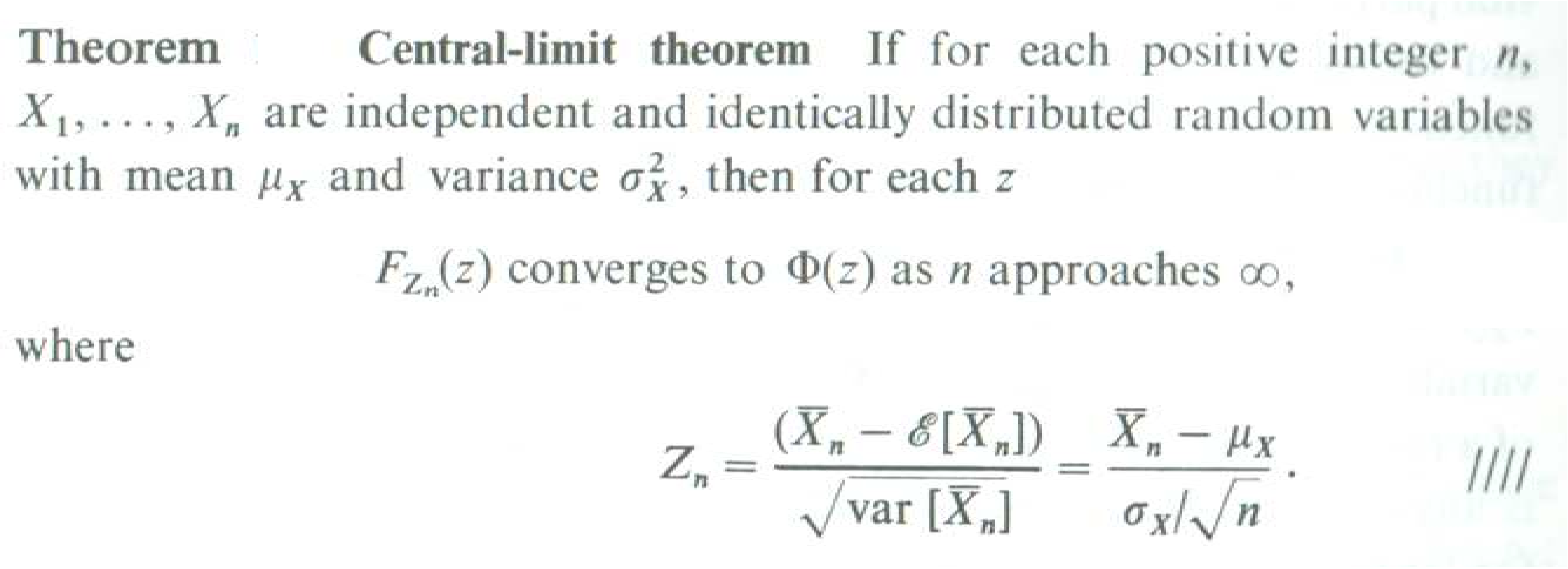
$$\begin{aligned}\operatorname{var}\left[\sum_1^n a_i X_i\right] &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \operatorname{cov}[X_i, X_j] \\ &= \sum_{i=1}^n a_i^2 \operatorname{var}[X_i] + \sum_{i \neq j}\sum a_i a_j \operatorname{cov}[X_i, X_j].\end{aligned}$$

In particular, if $X_1, \ldots, X_n$ are independent and identically distributed random variables with mean μ_X and variance σ_X^2 and if $\bar{X}_n = (1/n)\sum_1^n X_i$, then

$$\boxed{\mathscr{E}[\bar{X}_n] = \mu_X, \qquad \text{and} \qquad \operatorname{var}[\bar{X}_n] = \frac{\sigma_X^2}{n}.}$$

The Central Limit Theorem

- One of the most important theorems of probability theory is the Central Limit Theorem. It gives an approximate distribution of an average.

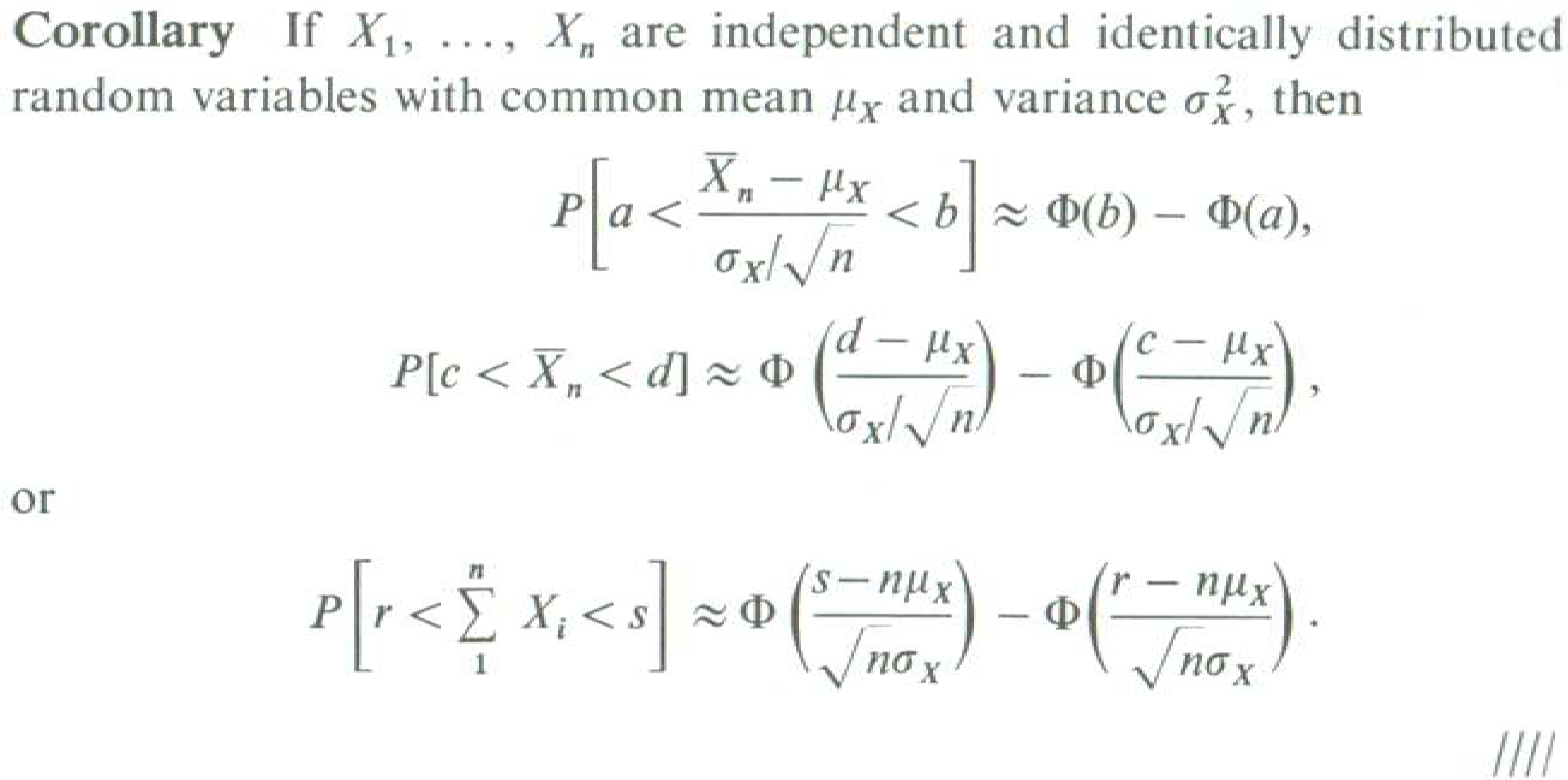


Theorem **Central-limit theorem** If for each positive integer n, $X_1, \ldots, X_n$ are independent and identically distributed random variables with mean μ_X and variance σ_X^2, then for each z

$$F_{Z_n}(z) \text{ converges to } \Phi(z) \text{ as } n \text{ approaches } \infty,$$

where

$$Z_n = \frac{(\overline{X}_n - \mathscr{E}[\overline{X}_n])}{\sqrt{\operatorname{var}[\overline{X}_n]}} = \frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}}. \qquad ////$$



Corollary If $X_1, \ldots, X_n$ are independent and identically distributed random variables with common mean μ_X and variance σ_X^2, then

$$P\left[a < \frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}} < b\right] \approx \Phi(b) - \Phi(a),$$

$$P[c < \overline{X}_n < d] \approx \Phi\left(\frac{d - \mu_X}{\sigma_X/\sqrt{n}}\right) - \Phi\left(\frac{c - \mu_X}{\sigma_X/\sqrt{n}}\right),$$

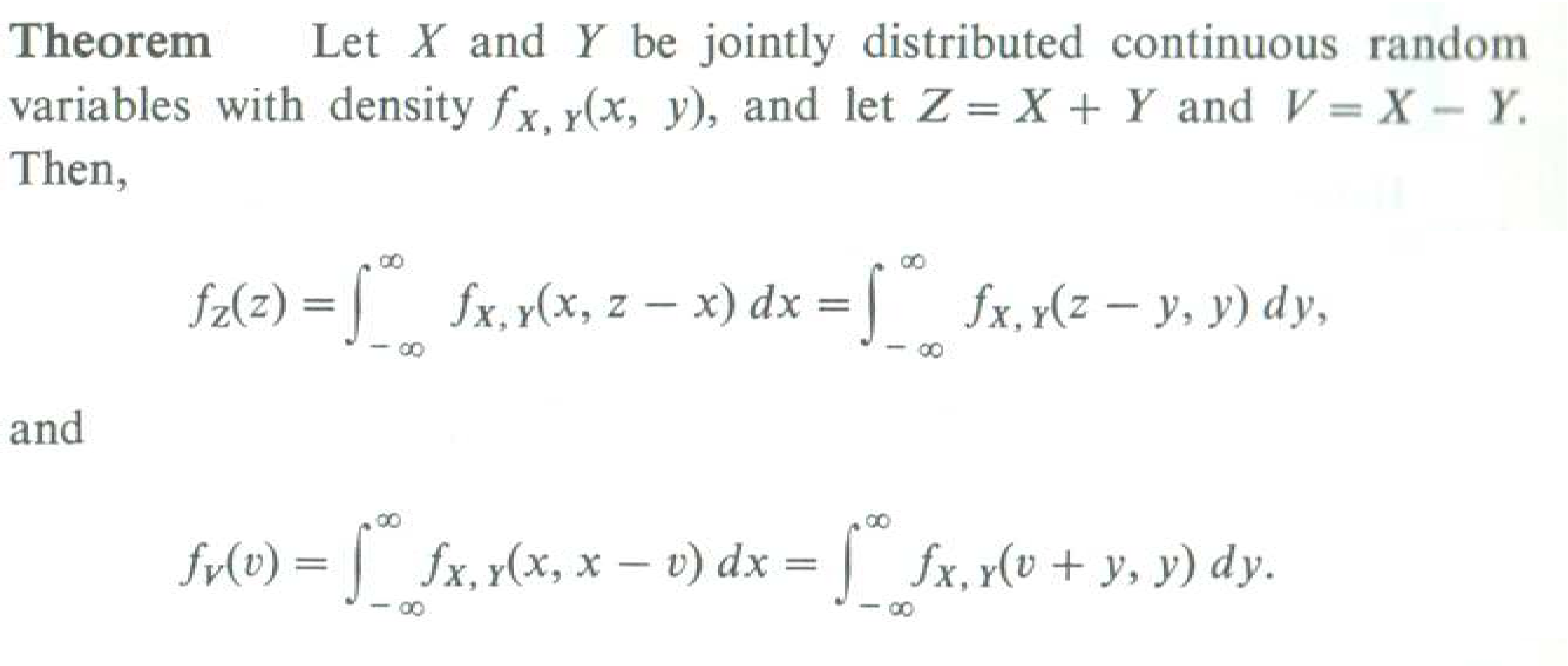
or

$$P\left[r < \sum_1^n X_i < s\right] \approx \Phi\left(\frac{s - n\mu_X}{\sqrt{n}\sigma_X}\right) - \Phi\left(\frac{r - n\mu_X}{\sqrt{n}\sigma_X}\right).$$

////

Remark : In the light of the central limit theorem, our results concerning 1-dimensional random walks is of no surprise: as the number of steps increases, it is expected that position of the particle becomes normally distributed in the limit.

Determine the distribution : (1) Cumulative-distribution function technique



Theorem Let X and Y be jointly distributed continuous random variables with density $f_{X,Y}(x, y)$, and let $Z = X + Y$ and $V = X - Y$. Then,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\, dx = \int_{-\infty}^{\infty} f_{X,Y}(z - y, y)\, dy,$$

and

$$f_V(v) = \int_{-\infty}^{\infty} f_{X,Y}(x, x - v)\, dx = \int_{-\infty}^{\infty} f_{X,Y}(v + y, y)\, dy.$$

PROOF We will prove only the first part ; the others are proved in an analogous manner.

$$F_Z(z) = P[Z \leq z] = P[X + Y \leq z] = \iint_{x+y \leq z} f_{X,Y}(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-x} f_{X,Y}(x, y)\, dy \right] dx$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z} f_{X,Y}(x, u - x)\, du \right] dx$$

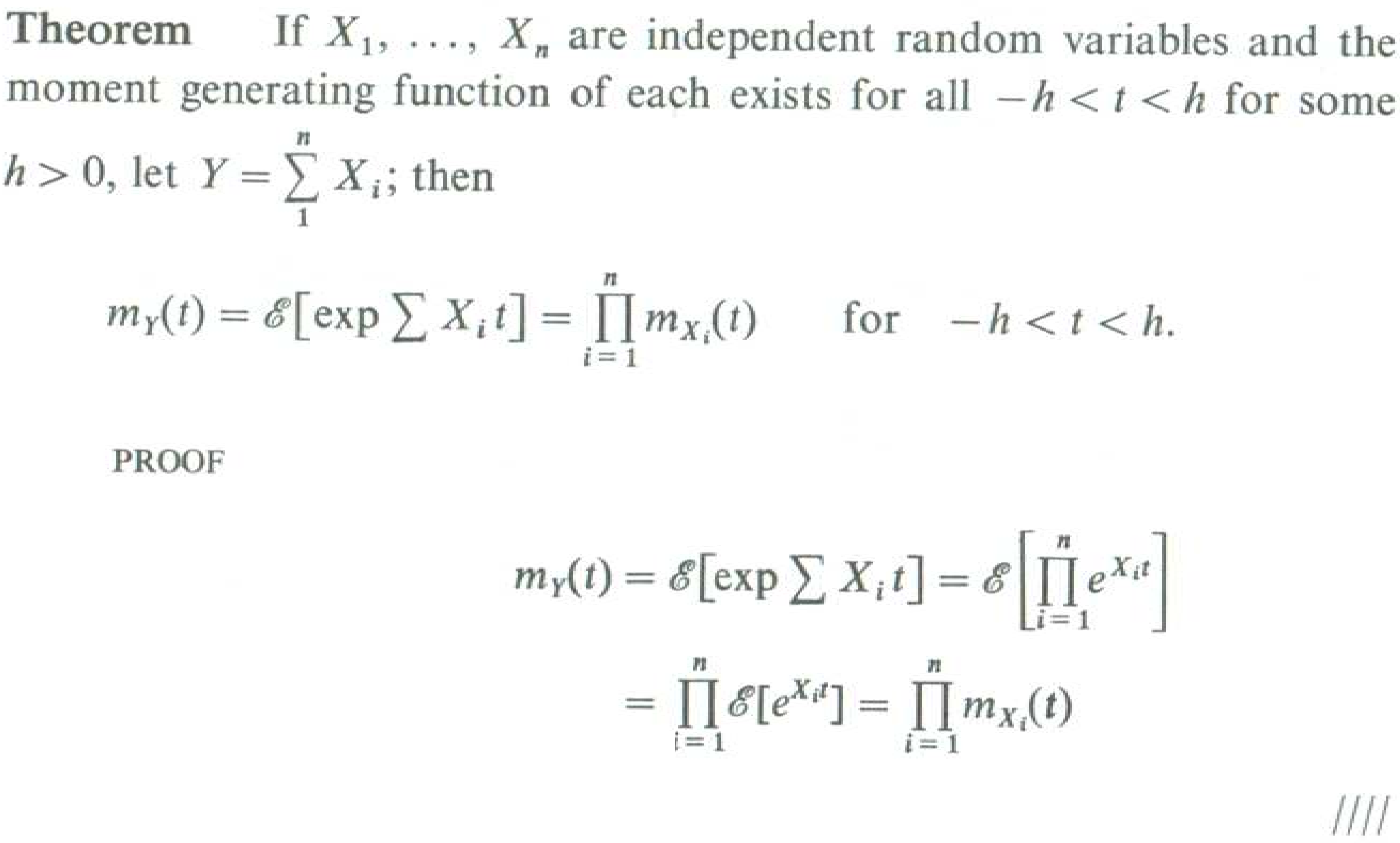
by making the substitution $y = u - x$.

Now

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{d}{dz} \left\{ \int_{-\infty}^{z} \left[\int_{-\infty}^{\infty} f_{X,Y}(x, u - x)\, dx \right] du \right\}$$

$$= \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)\, dx. \qquad ////$$

(2) Moment-generating-function technique



Theorem If $X_1, \ldots, X_n$ are independent random variables and the moment generating function of each exists for all $-h < t < h$ for some $h > 0$, let $Y = \sum_1^n X_i$; then

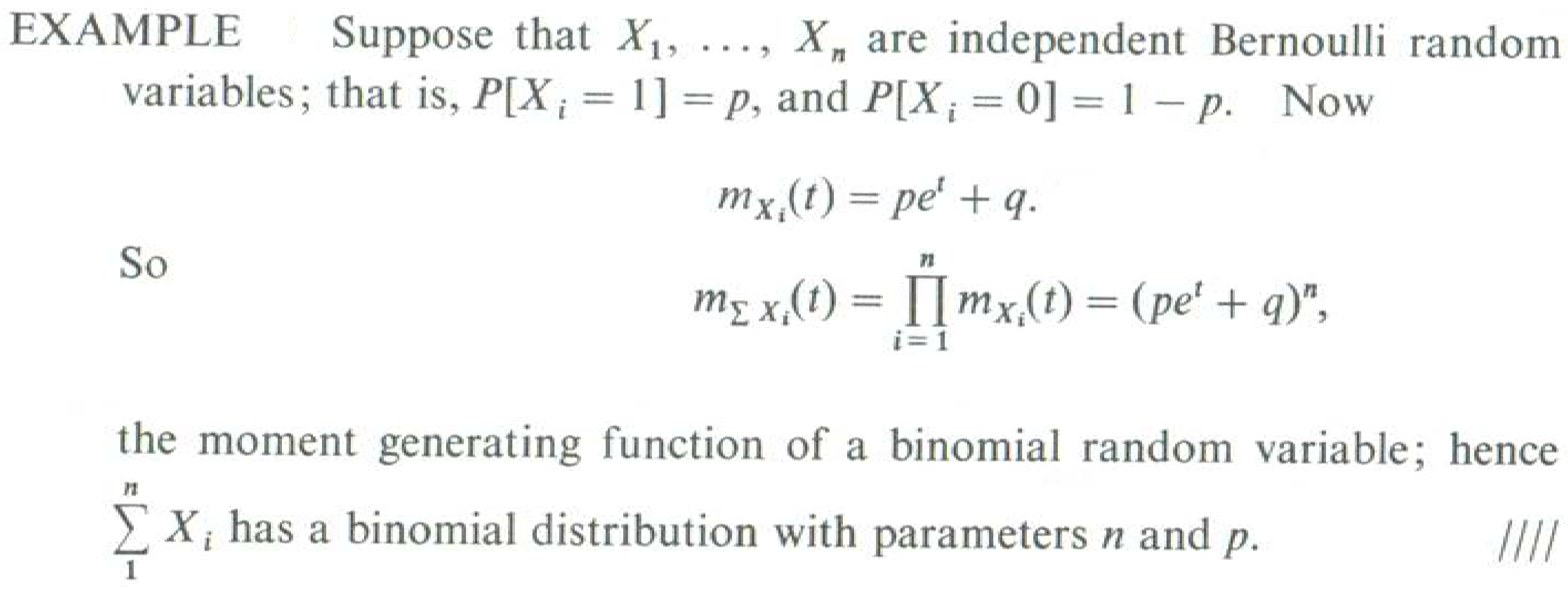
$$m_Y(t) = \mathscr{E}[\exp \sum X_i t] = \prod_{i=1}^{n} m_{X_i}(t) \qquad \text{for} \quad -h < t < h.$$

PROOF

$$m_Y(t) = \mathscr{E}[\exp \sum X_i t] = \mathscr{E}\left[\prod_{i=1}^{n} e^{X_i t}\right]$$

$$= \prod_{i=1}^{n} \mathscr{E}[e^{X_i t}] = \prod_{i=1}^{n} m_{X_i}(t)$$

////



EXAMPLE Suppose that $X_1, \ldots, X_n$ are independent Bernoulli random variables; that is, $P[X_i = 1] = p$, and $P[X_i = 0] = 1 - p$. Now

$$m_{X_i}(t) = pe^t + q.$$

So

$$m_{\sum X_i}(t) = \prod_{i=1}^{n} m_{X_i}(t) = (pe^t + q)^n,$$

the moment generating function of a binomial random variable; hence $\sum_1^n X_i$ has a binomial distribution with parameters n and p. ////

Multivariate transform $M_{X,Y}(s_1, s_2)$ associated with X and Y

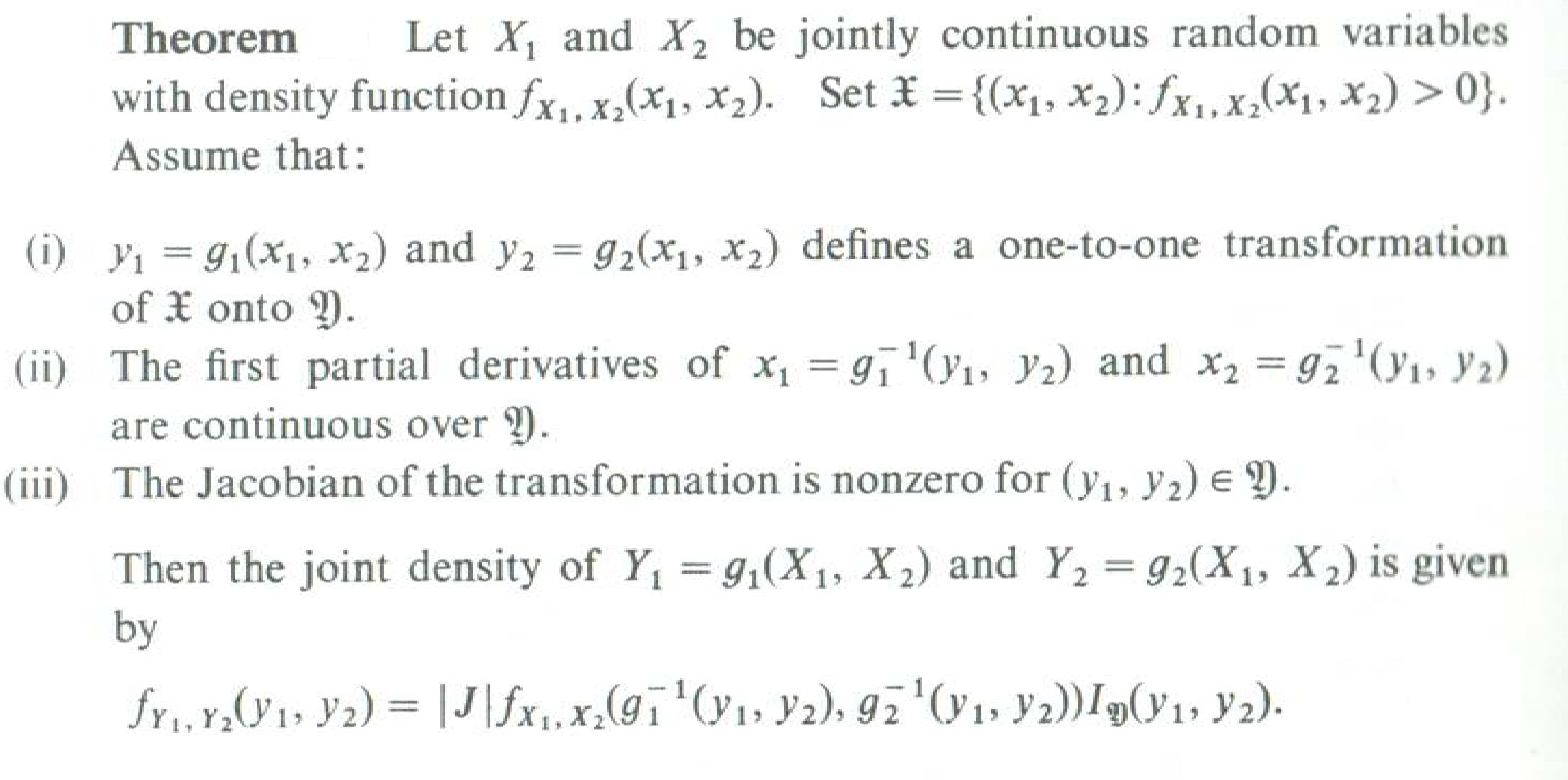
- The multivariate transform $M_{X,Y}(s_1, s_2)$ of X and Y is given by

$$M_{X,Y}(s_1, s_2) = E(e^{s_1 X + s_2 Y})$$

- It is a direct generalization of the moment generating functions we have seen for a single random variable or a sum of independent random variables:
  - If X and Y are independent random variables, and $s_1 = s_2 = t$, then $M_{X+Y} = E(e^{t(X+Y)}) = M_X(t)M_Y(t) = M_{X,Y}(t, t)$
- The function $M_{X,Y}(s_1, s_2)$ is called the **joint moment generating function** of X and Y

(3) The transformation technique– when the number of variables grows

$$Y_1 = g_1(X_1 \qquad X_n), ..., Y_k = g_k(X_1, ..., X_n)$$



Theorem Let X_1 and X_2 be jointly continuous random variables with density function $f_{X_1, X_2}(x_1, x_2)$. Set $\mathfrak{X} = \{(x_1, x_2): f_{X_1, X_2}(x_1, x_2) > 0\}$. Assume that:

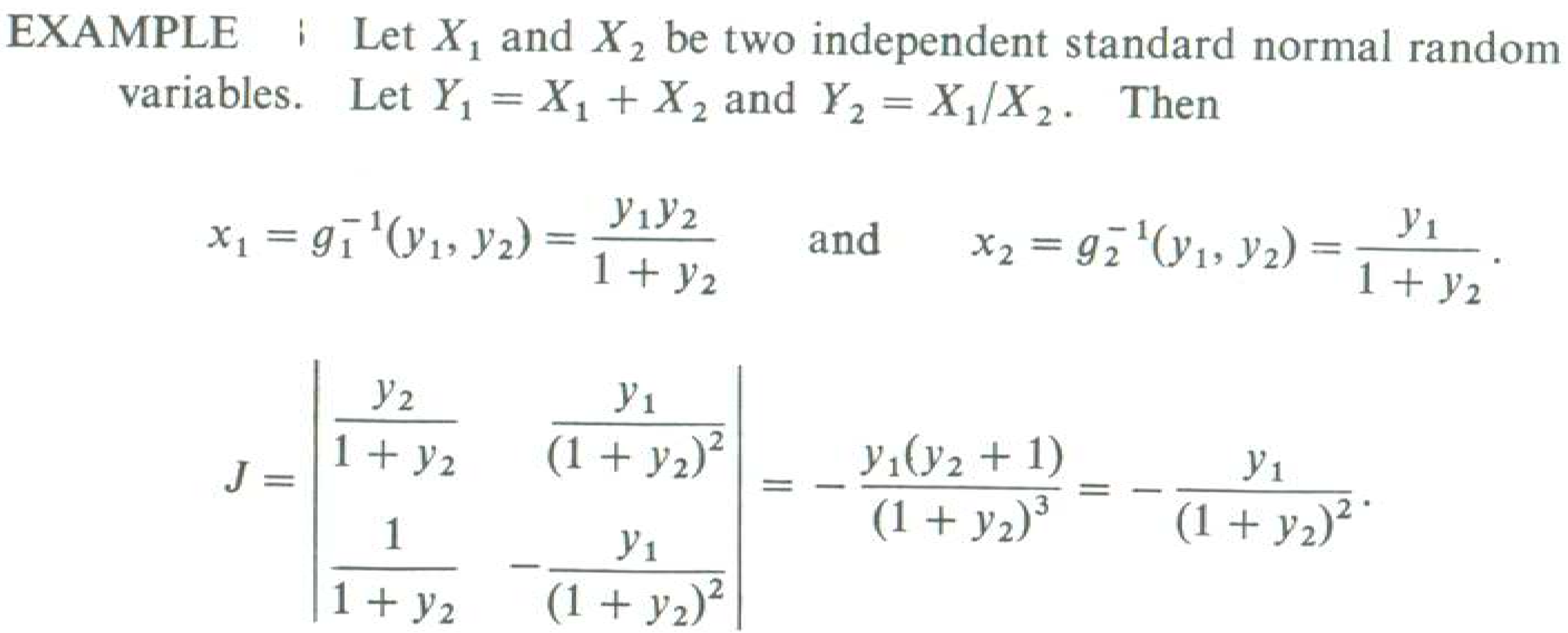
(i) $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ defines a one-to-one transformation of $\mathfrak{X}$ onto $\mathfrak{Y}$.

(ii) The first partial derivatives of $x_1 = g_1^{-1}(y_1, y_2)$ and $x_2 = g_2^{-1}(y_1, y_2)$ are continuous over $\mathfrak{Y}$.

(iii) The Jacobian of the transformation is nonzero for $(y_1, y_2) \in \mathfrak{Y}$.

Then the joint density of $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ is given by

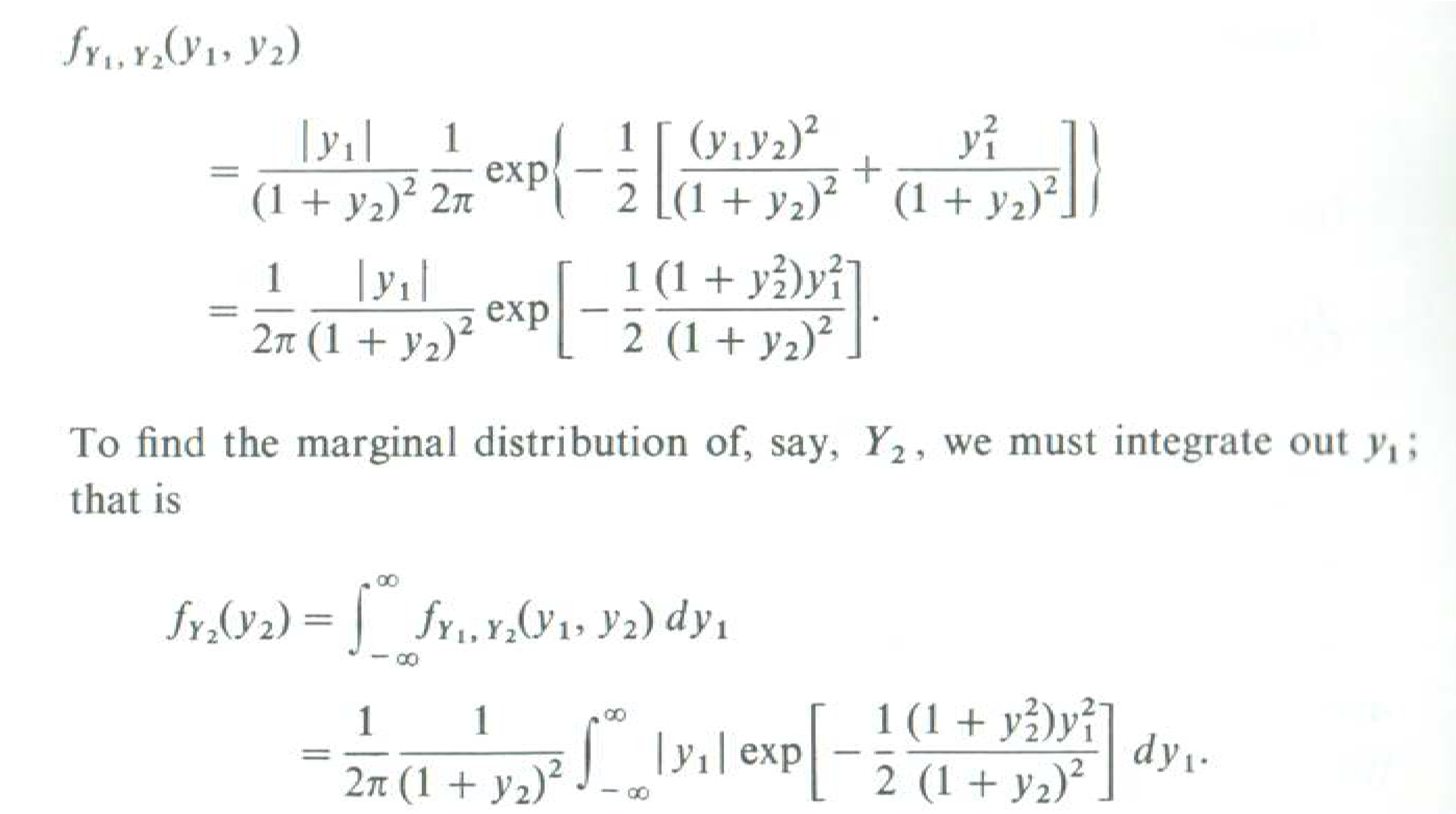
$$f_{Y_1, Y_2}(y_1, y_2) = |J| f_{X_1, X_2}(g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)) I_{\mathfrak{Y}}(y_1, y_2).$$



EXAMPLE Let X_1 and X_2 be two independent standard normal random variables. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1/X_2$. Then

$$x_1 = g_1^{-1}(y_1, y_2) = \frac{y_1 y_2}{1 + y_2} \qquad \text{and} \qquad x_2 = g_2^{-1}(y_1, y_2) = \frac{y_1}{1 + y_2}.$$

$$J = \begin{vmatrix} \dfrac{y_2}{1 + y_2} & \dfrac{y_1}{(1 + y_2)^2} \\ \dfrac{1}{1 + y_2} & -\dfrac{y_1}{(1 + y_2)^2} \end{vmatrix} = -\frac{y_1(y_2 + 1)}{(1 + y_2)^3} = -\frac{y_1}{(1 + y_2)^2}.$$



$$f_{Y_1, Y_2}(y_1, y_2)$$

$$= \frac{|y_1|}{(1+y_2)^2} \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} \left[\frac{(y_1 y_2)^2}{(1+y_2)^2} + \frac{y_1^2}{(1+y_2)^2} \right] \right\}$$

$$= \frac{1}{2\pi} \frac{|y_1|}{(1+y_2)^2} \exp\left[-\frac{1}{2} \frac{(1+y_2^2) y_1^2}{(1+y_2)^2} \right].$$

To find the marginal distribution of, say, Y_2, we must integrate out y_1; that is

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1, Y_2}(y_1, y_2)\, dy_1$$

$$= \frac{1}{2\pi} \frac{1}{(1+y_2)^2} \int_{-\infty}^{\infty} |y_1| \exp\left[-\frac{1}{2} \frac{(1+y_2^2) y_1^2}{(1+y_2)^2} \right] dy_1.$$

Let

$$u = \frac{1}{2}\frac{(1+y_2^2)}{(1+y_2)^2}y_1^2;$$

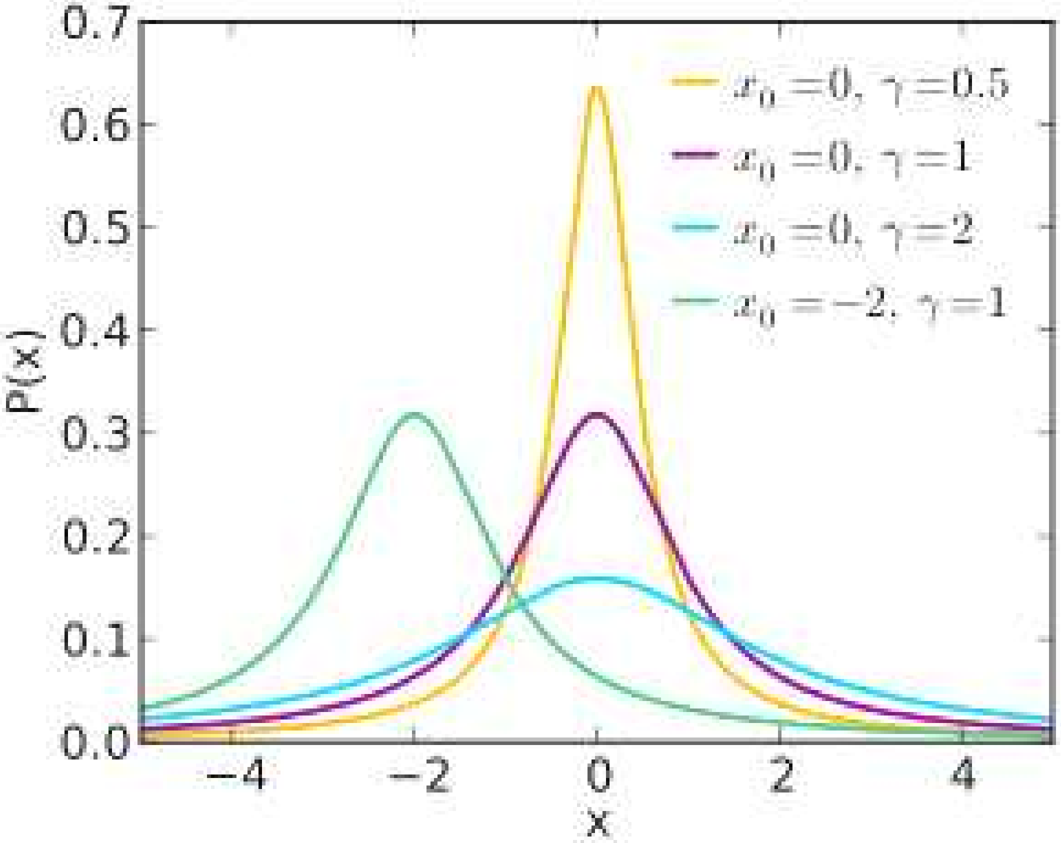
then

$$du = \frac{(1+y_2^2)}{(1+y_2)^2}y_1\,dy_1$$

and so

$$f_{Y_2}(y_2) = \frac{1}{2\pi}\cdot\frac{1}{(1+y_2)^2}\cdot\frac{(1+y_2)^2}{1+y_2^2}(2)\int_0^\infty e^{-u}\,du = \frac{1}{\pi}\cdot\frac{1}{1+y_2^2},$$

a Cauchy density. That is, the *ratio of two independent standard normal random variables has a Cauchy distribution.* ////



where x_0 is the location parameter, specifying the location of the peak of the Cauchy distribution, and γ is the scale parameter (note: mean and standard deviation are undefined!)

6.3 Two or more random variables: multivariate moments

- Let X_1 and X_2 be a jointly distributed random variables (discrete or continuous), then for any pair of positive integers (k_1, k_2) **the joint moment** of (X_1, X_2) of order (k_1, k_2) is defined to be:

$$\mu'_{k_1 k_2} = E(X_1^{k_1} X_2^{k_2})$$

$$= \begin{cases} \sum\limits_{x_1} \sum\limits_{x_2} x_1^{k_1} x_2^{k_2} p\left(x_1, x_2\right) & \text{if } X_1, X_2 \text{ are discrete} \\ \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} x_1^{k_1} x_2^{k_2} f\left(x_1, x_2\right) dx_1 dx_2 & \text{if } X_1, X_2 \text{ are continuous} \end{cases}$$

- Let X_1 and X_2 be a jointly distributed random variables (discrete or continuous), then for any pair of positive integers (k1, k2) the **joint central moment** of (X_1, X_2) of order (k_1, k_2) is defined to be:

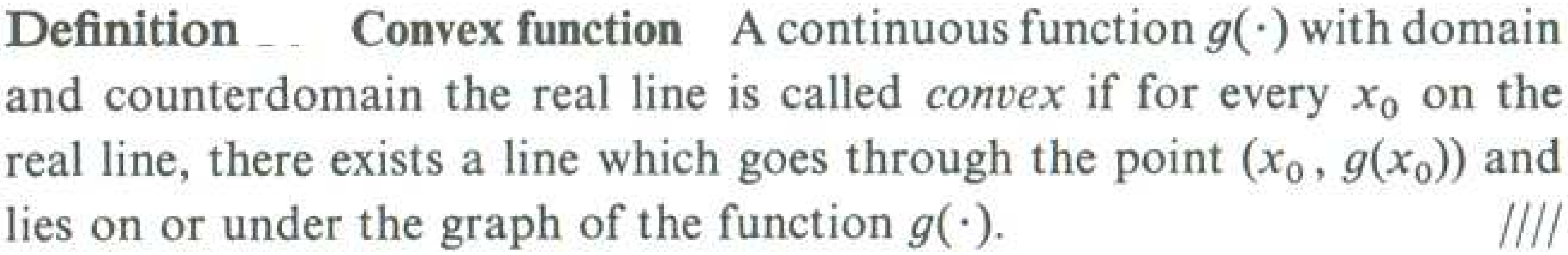
$$\mu_{k_1k_2} = E(\,(X_1 - \mu_1)^{k_1}(X_2 - \mu_2)^{k_2}\,)$$

$$= \begin{cases} \sum_{x_1}\sum_{x_2}(x_1-\mu_1)^{k_1}(x_2-\mu_2)^{k_2}\ p(x_1,x_2) & \text{if } X_1, X_2 \text{ are discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)^{k_1}(x_2-\mu_2)^{k_2}\ f(x_1,x_2)\,dx_1dx_2 & \text{if } X_1, X_2 \text{ are continuous} \end{cases}$$

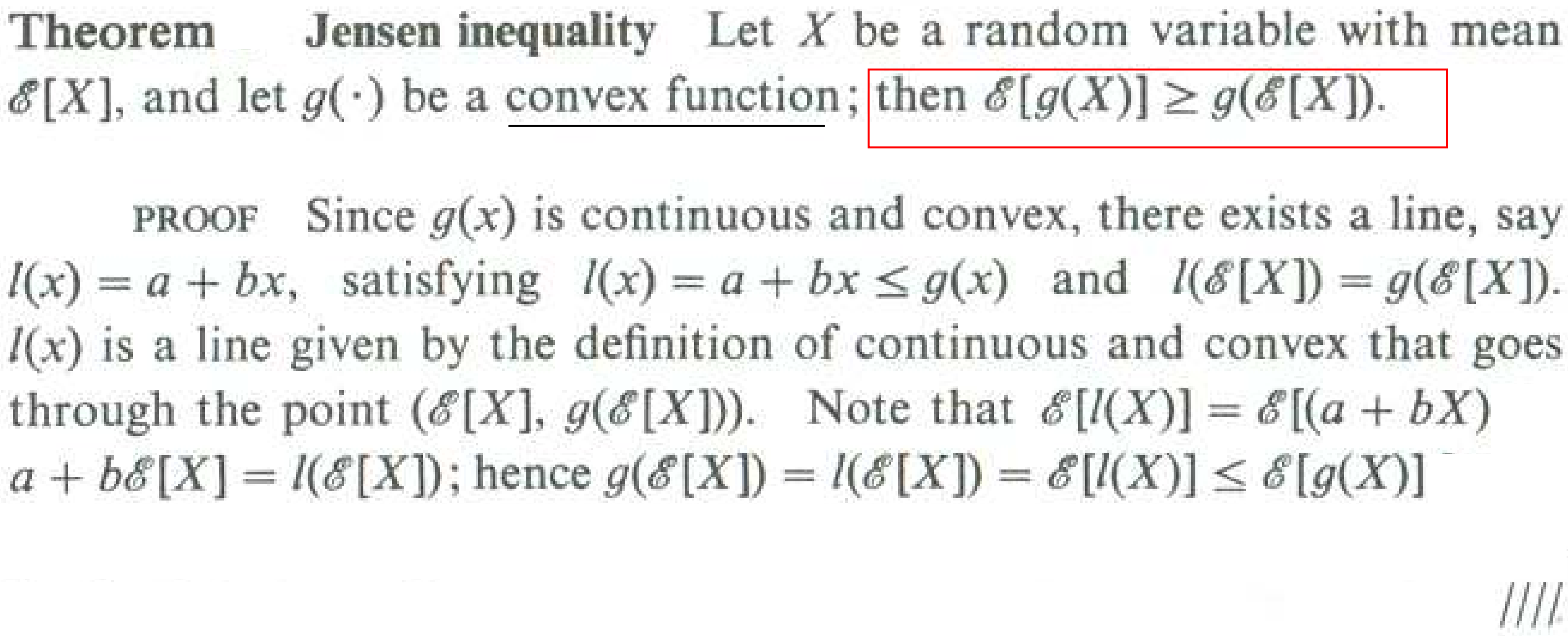
where $\mu_i = E(X_i), i = 1, 2$

7 Inequalities

7.1 Jensen inequality



Definition **Convex function** A continuous function $g(\cdot)$ with domain and counterdomain the real line is called *convex* if for every x_0 on the real line, there exists a line which goes through the point $(x_0, g(x_0))$ and lies on or under the graph of the function $g(\cdot)$. ////



Theorem **Jensen inequality** Let X be a random variable with mean $\mathscr{E}[X]$, and let $g(\cdot)$ be a convex function; then $\mathscr{E}[g(X)] \geq g(\mathscr{E}[X])$.

PROOF Since $g(x)$ is continuous and convex, there exists a line, say $l(x) = a + bx$, satisfying $l(x) = a + bx \leq g(x)$ and $l(\mathscr{E}[X]) = g(\mathscr{E}[X])$. $l(x)$ is a line given by the definition of continuous and convex that goes through the point $(\mathscr{E}[X], g(\mathscr{E}[X]))$. Note that $\mathscr{E}[l(X)] = \mathscr{E}[(a + bX)$
$a + b\mathscr{E}[X] = l(\mathscr{E}[X])$; hence $g(\mathscr{E}[X]) = l(\mathscr{E}[X]) = \mathscr{E}[l(X)] \leq \mathscr{E}[g(X)]$

////

- Note that in general,

$$\mathcal{E}[g(X)] \neq g(\mathcal{E}[X])$$

- **Jensen inequality** can be used to prove the **Rao-Blackwell theorem**
- The latter provides a method for improving the performance of an unbiased estimator of a parameter (i.e. reduce its variance – cfr Chapter 5) provided that a “sufficient” statistic for this estimator is available.
- With g(x)=x^2 (hence g is a convex function), Jensen inequality says

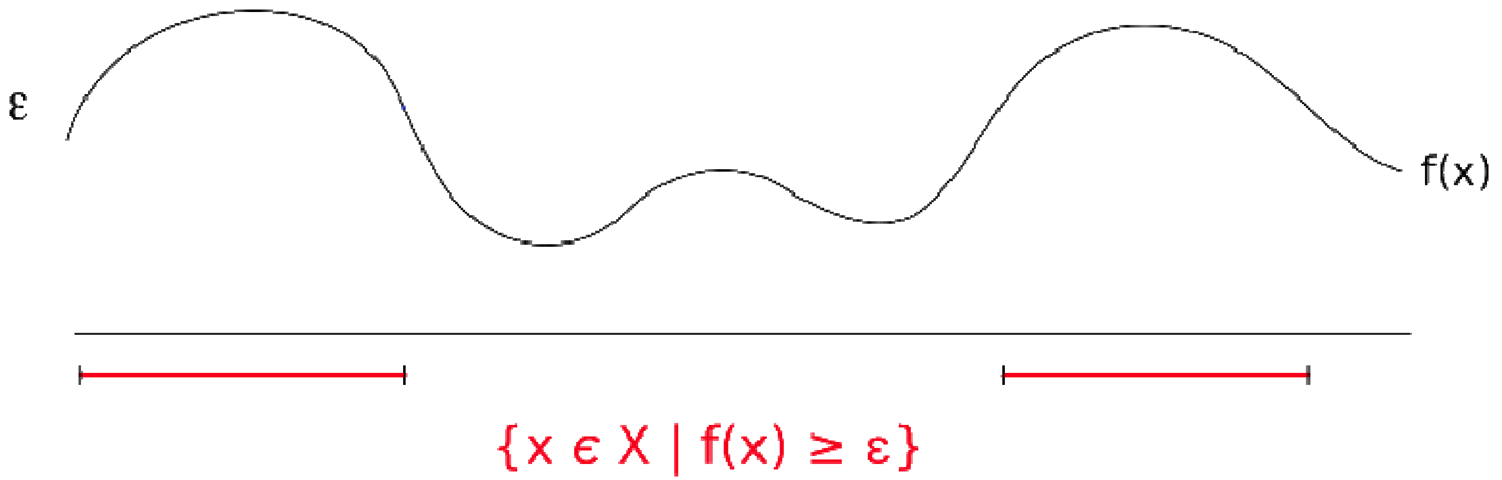
$$\boxed{\mathcal{E}[X^2] \geq (\mathcal{E}[X])^2}$$

and therefore that the variance of X is always non-negative

7.2 Markov’s inequality

- In probability theory, **Markov's inequality** gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant.
- Markov's inequality (and other similar inequalities) relate probabilities to expectations, and provide (frequently) loose but still useful bounds for the cumulative distribution function of a random variable.

- Markov's inequality gives an upper bound for the probability that *X* lies within the set indicated in red.



- Markov's inequality states that for any real-valued random variable *X* and any positive number *a*, we have

$$\boxed{P(X \geq a) \leq E(|X|)/a}$$

Proof:

Clearly, $aI_{(|X|\geq a)} \leq |X|$

Therefore also $E(aI_{(|X|\geq a)}) \leq E(|X|)$

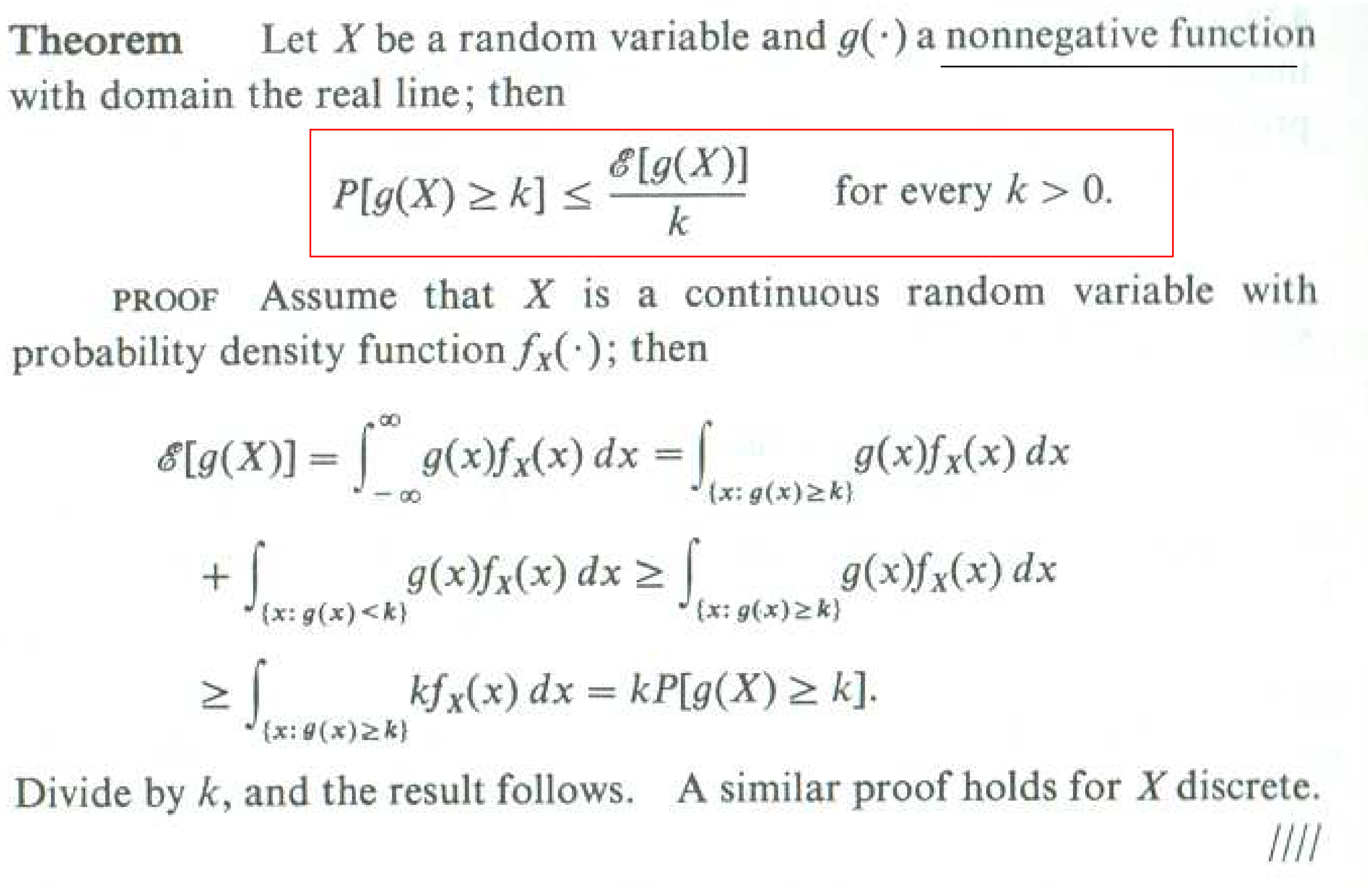
Using linearity of expectations, the left side of this inequality is the same as

$$a\mathrm{E}(I_{(|X|\geq a)}) = a\,\Pr(|X| \geq a).$$

Thus we have $aP(|X| \geq a) \leq E(|X|)$

and since *a* > 0, we can divide both sides by *a*.

7.3 Chebyshev's inequality



Theorem Let X be a random variable and $g(\cdot)$ a nonnegative function with domain the real line; then

$$\boxed{P[g(X) \geq k] \leq \frac{\mathscr{E}[g(X)]}{k} \qquad \text{for every } k > 0.}$$

PROOF Assume that X is a continuous random variable with probability density function $f_X(\cdot)$; then

$$\begin{aligned}\mathscr{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)\,dx = \int_{\{x:\, g(x) \geq k\}} g(x) f_X(x)\,dx \\ &\quad + \int_{\{x:\, g(x) < k\}} g(x) f_X(x)\,dx \geq \int_{\{x:\, g(x) \geq k\}} g(x) f_X(x)\,dx \\ &\geq \int_{\{x:\, g(x) \geq k\}} k f_X(x)\,dx = kP[g(X) \geq k].\end{aligned}$$

Divide by k, and the result follows. A similar proof holds for X discrete.

////

Corollary **Chebyshev inequality** If X is a random variable with finite variance,

$$\boxed{P[|X-\mu_X| \geq r\sigma_X]} = P[(X-\mu_X)^2 \geq r^2\sigma_X^2] \leq \frac{1}{r^2} \qquad \text{for every } r > 0.$$

PROOF Take $g(x) = (x - \mu_X)^2$ and $k = r^2\sigma_X^2$.

////

Remark If X is a random variable with finite variance,

$$\boxed{P[|X-\mu_X| < \mathrm{r}\sigma_X] \geq 1 - \frac{1}{r^2},}$$

which is just a rewriting ////

7.4 Cantelli's inequality – no exam material

- A one-tailed variant of Chebyshev’s inequality with *k* > 0, is

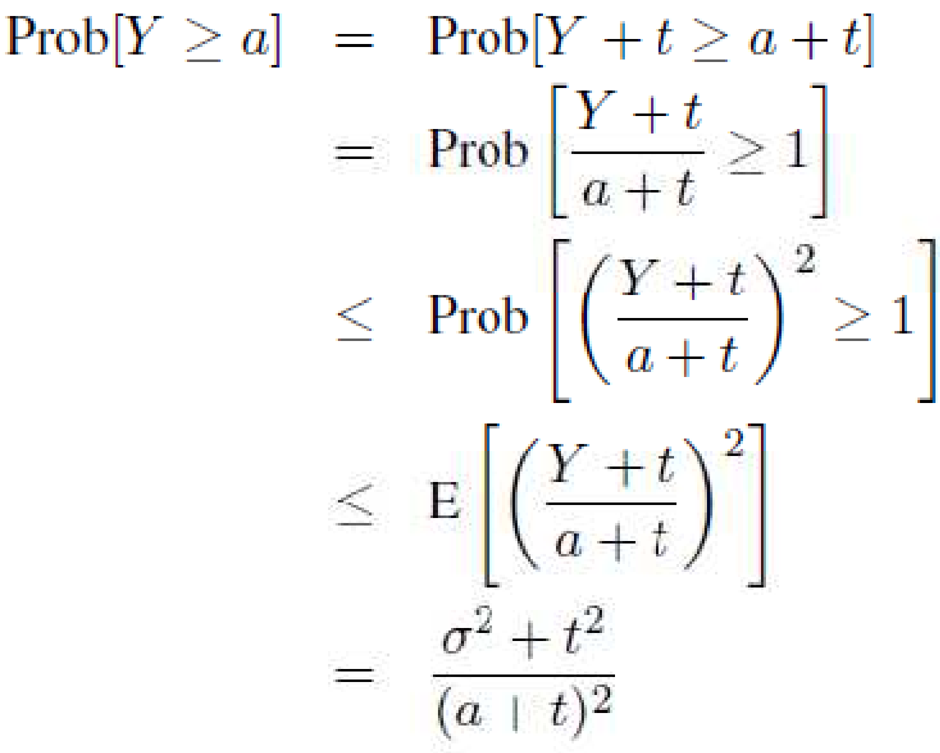
$$P(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Proof:

Without loss of generality, we assume $E(X) = 0$.

Let $Y = X - \mu$ and $Var(Y) = Var(X) = \sigma^2$.

Thus for any t such that $t + a > 0$ we have



$$
\begin{aligned}
\text{Prob}[Y \geq a] &= \text{Prob}[Y + t \geq a + t] \\
&= \text{Prob}\left[\frac{Y+t}{a+t} \geq 1\right] \\
&\leq \text{Prob}\left[\left(\frac{Y+t}{a+t}\right)^2 \geq 1\right] \\
&\leq \text{E}\left[\left(\frac{Y+t}{a+t}\right)^2\right] \\
&= \frac{\sigma^2 + t^2}{(a \mid t)^2}
\end{aligned}
$$

- The second inequality follows from Markov inequality: $P(Z \geq 1) \leq E(|Z|)$
- The above derivation holds for any t such that t+a>0. We can therefore select t to minimize the right-hand side: $t = \sigma^2/a > 0$

- An application: for probability distributions having an expected value and a median, the mean (i.e., the expected value) and the median can never differ from each other by more than one standard deviation.

  To express this in mathematical notation, let μ, m, and σ be respectively the mean, the median, and the standard deviation. Then

$$|\mu - m| \leq \sigma$$

  (There is no need to rely on an assumption that the variance exists, i.e., is finite. This inequality is trivially true if the variance is infinite.)

Proof:

Setting *k* = 1 in the statement for the one-sided inequality gives:

$$\Pr(X - \mu \geq \sigma) \leq \frac{1}{2}.$$

By changing the sign of *X* and so of *μ*, we get

$$\Pr(X \leq \mu - \sigma) \leq \frac{1}{2}.$$

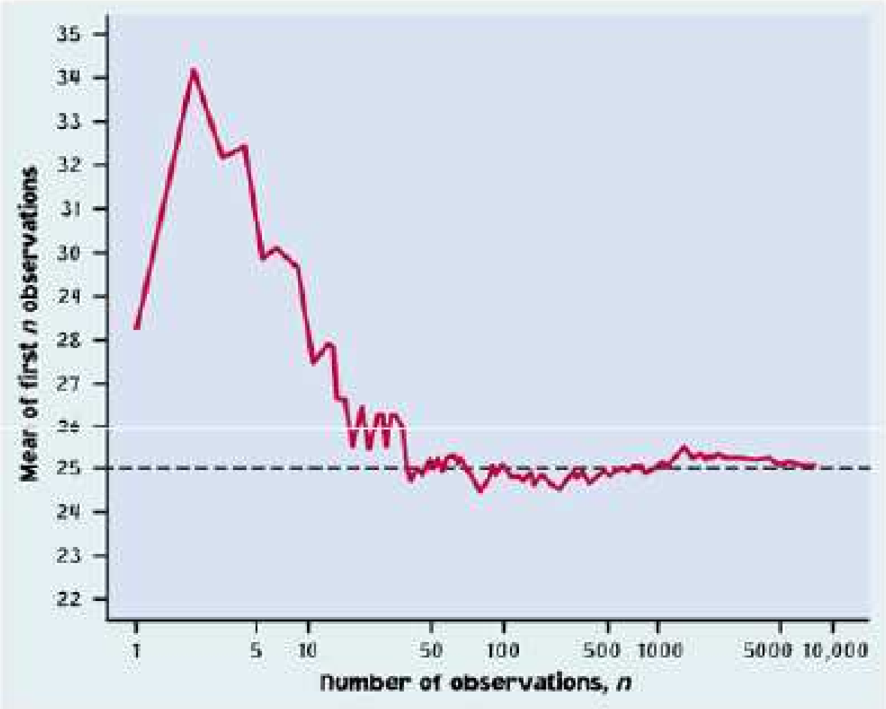
Thus the median is within one standard deviation of the mean.

- Chebyshev inequality can also be used to prove the law of large numbers

7.5 Law of large numbers revisited

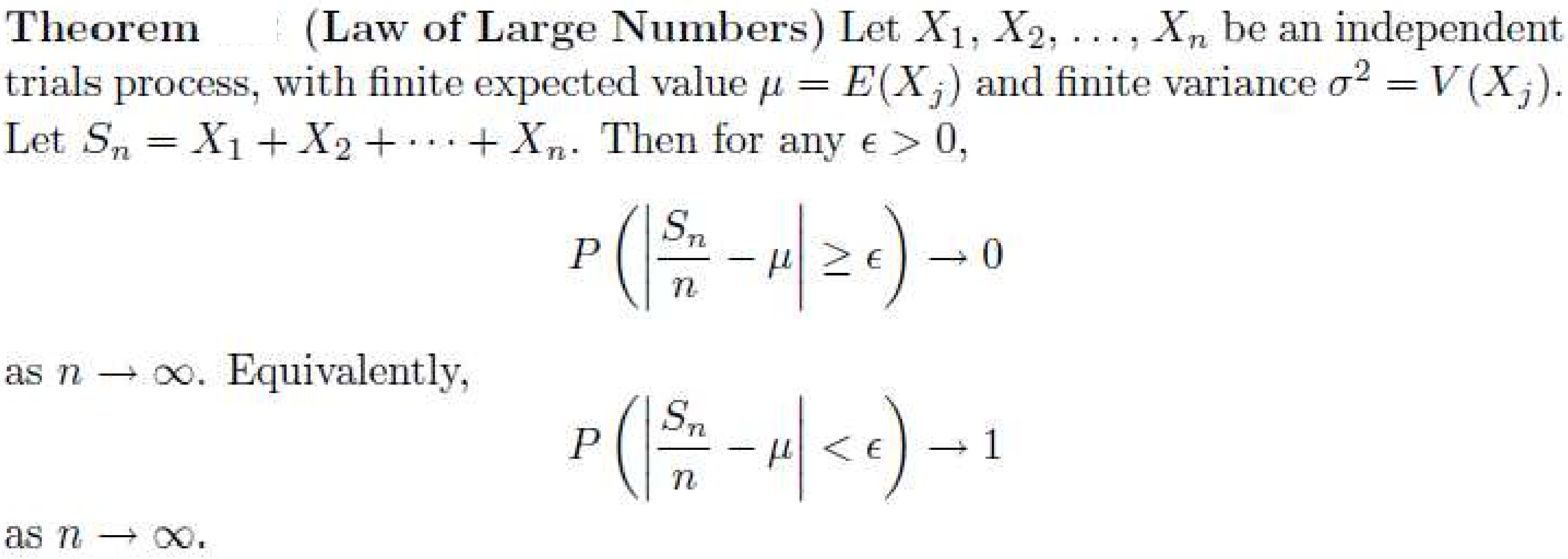
As the number of randomly drawn observations (*n*) in a sample increases, the mean of the sample (*x* bar) gets closer and closer to the population mean μ.

This is the **law of large numbers.** It is valid for any population.



Note: We often intuitively expect predictability over a few random observations, but it is wrong. The law of large numbers only applies to really large numbers.

- Or formulated in the formal way …



Theorem **(Law of Large Numbers)** Let $X_1, X_2, \ldots, X_n$ be an independent trials process, with finite expected value $\mu = E(X_j)$ and finite variance $\sigma^2 = V(X_j)$. Let $S_n = X_1 + X_2 + \cdots + X_n$. Then for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

as $n \to \infty$. Equivalently,

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) \to 1$$

as $n \to \infty$.