1 Random Variables and Probability Distributions

1.1 Concept of a Random Variable

- It is often important to allocate a numerical description to the outcome of a statistical experiment.
- These values are random quantities determined by the outcome of the experiment.
- **Definition 3.1**:
  A **random variable** is a function that associates a real number with each element in the sample space.
- We use a capital letter, say X, to denote a random variable and its corresponding small letter , x in this case, for one of its value.
- One and only one numerical value is assigned to each sample point X.
- **Example 3.1**: Two balls are drawn in succession without replacement from an box containing 4 red balls and 3 black balls.
- The possible outcomes and the values y of the random variable Y, where Y is the number of red balls, are

Sample Space	y
RR	2
RB	1
BR	1
BB	0

Example: Number of defective (D) products when 3 products are tested.

Outcomes in Sample Space	x: value of X
DDD	3
DDN	2
DND	2
DNN	1
NDD	2
NDN	1
NND	1
NNN	0

- **Example 3.3**: Components from the production line are defective or not defective.
- Define the random variable X by

$$X = \left\{ \begin{array}{ll} 1, & if\ the\ component\ is\ defective \\ 0, & if\ the\ component\ is\ not\ defective \end{array} \right\}$$

- This random variable is categorical in nature.
- **Example 3.5**: A process will be evaluated by sampling items until a defective item is observed.
- Define X by the number of consecutive items observed

Sample Space	x
D	1
ND	2
NND	3
⋮	⋮

- According to the countability of the sample space which is measurable, it can be either discrete or continuous.
- **Discrete random variable:** If a random variable take on only a countable number of distinct values.
  - If the set of possible outcomes is countable
  - Often represent count data, such as the number of defectives, highway fatalities
- **Continuous random variable**: If a random variable can take on values on a continuous scale.
  - often represent measured data, such as heights, weights, temperatures, distance or life periods
- **Definition 3.2**:

  Discrete sample space: If a sample space contains a finite number of possibilities or an unending sequence with as many elements as there are whole numbers.
- **Definition 3.3**:

  Continuous sample space: If a sample space contains an infinite number of possibilities equal to the number of points on a line segment.

1.2 Discrete Probability Distributions

- A discrete random variable assumes each of its values with a certain probability.
- Frequently, it is convenient to represent all the probabilities of a random variable X by a formula;

$$f(x) = P(X = x), \quad f(3) = P(X = 3)$$

- **Definition 3.4**:

  The set of ordered pairs $(x, f(x))$ is a **probability function** (**probability mass function**, or **probability distribution)** of the discrete random variable X if for each possible outcome x,

  1. $f(x) \geq 0$,
  2. $\sum f(x) = 1$,
  3. $P(X = x) = f(x)$.

- The probability distribution of a discrete random variable can be presented in the form of a mathematical formula, a table, or a graph-probability histogram or barchart.

Example: Let X be the random variable: number of heads in 3 tosses of a fair coin.

Sample Space	x
TTT	0
TTH	1
THT	1
THH	2
HTT	1
HTH	2
HHT	2
HHH	3

$P(X = x)$: Probability that outcome is a specific x value.

x	0	1	2	3
P(X=x)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

- **Example 3.8**: A shipment of 8 similar microcomputers to a retail outlet contains 3 that are defective.

- If a school make a random purchase of 2 of these computers.
- Find the probability distribution for the number of defectives.

x	0	1	2
f(x)	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$

$$f(0) = P(X = 0) = \frac{\binom{3}{0}\binom{5}{2}}{\binom{8}{2}} = \frac{10}{28}$$

$$f(1) = P(X = 1) = \frac{\binom{3}{1}\binom{5}{1}}{\binom{8}{2}} = \frac{15}{28}$$

$$f(2) = P(X = 2) = \frac{\binom{3}{2}\binom{5}{0}}{\binom{8}{2}} = \frac{3}{28}$$

- **Definition 3.5**:

  The **Cumulative distribution function** $F(x)$ of a discrete random variable X with probability distribution $f(x)$ is

$$F(x) = P(X \le x) = \sum_{t \le x} f(t), \ \ for \ \ -\infty < x < \infty$$

- **Example 3.10**: Find the cumulative distribution of the random variable X in Example 3.9.

$f(0) = \frac{1}{16}, f(1) = \frac{4}{16}, f(2) = \frac{6}{16}, f(3) = \frac{4}{16}, f(4) = \frac{1}{16},$

$F(0) = f(0) = \frac{1}{16}$
$F(1) = f(0) + f(1) = \frac{5}{16}$
$F(2) = f(0) + f(1) + f(2) = \frac{11}{16}$
$F(3) = f(0) + f(1) + f(2) + f(3) = \frac{15}{16}$
$F(4) = f(0) + f(1) + f(2) + f(3) + f(4) = 1$

$$F(x) = \left\{ \begin{matrix} 0 \; for \; x < 0 \\ \frac{1}{16} \; for \; 0 \leq x < 1 \\ \frac{5}{16} \; for \; 1 \leq x < 2 \\ \frac{11}{16} \; for \; 2 \leq x < 3 \\ \frac{15}{16} \; for \; 3 \leq x < 4 \\ 1 \; for \; x \geq 4 \end{matrix} \right\}$$

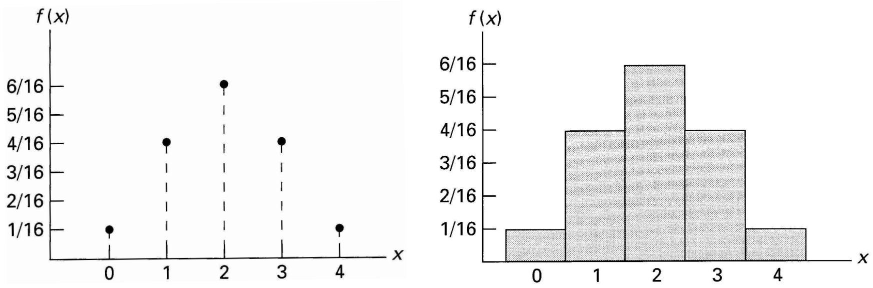

Figure 1: Bar chart and probability histogram

[

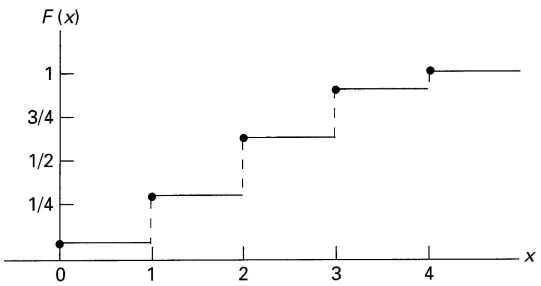

Figure 2: Discrete cumulative distribution.

1.3 Continuous Probability Distributions

- A continuous random variable has a probability of zero of assuming exactly any of its values.

$$P(a < X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a \leq X \leq b)$$

- **Example**: Height of a random person. $P(X = 178\ cm) = 0$. No assuming exactly.

- With continuous random variables we talk about the probability of x being in some interval, like $P(a < X < b)$, rather than x assuming a precise value like $P(X = a)$.

- Its probability distribution cannot be given in tabular form, but can be stated as a formula a function of the numerical values of the continuous random variables.

- Some of these functions are shown below:

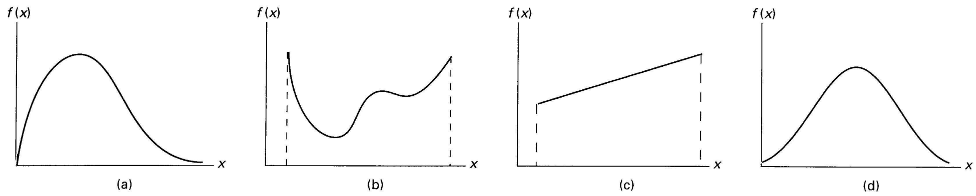

Figure 3: Typical density functions.

- **Definition 3.6**:

  The function $f(x)$ is a **probability density function** (or **density function**, **p.d.f**) for the continuous random variable X, defined over the set of real numbers R, if

  1. $f(X) \geq 0,\ for\ all\ x \in R$
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$
  3. $P(a < X < b) = \int_a^b f(x)dx$

A probability density function is constructed so that the area under its curve bounded by the x axis is equal to 1.

- **Example 3.11**: Suppose that the error in reaction temperature in °C is a continuous random variable X having the probability density function

$$f(x) = \begin{Bmatrix} \frac{x^2}{3}\ for\ -1 < x < 2 \\ 0,\ elsewhere \end{Bmatrix}$$

- Verify $\int_{-\infty}^{\infty} f(x)dx = 1$

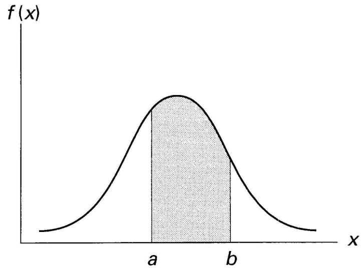

Figure 4: $P(a < X < b)$

- $\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{2} \frac{x^2}{3}dx = \frac{x^3}{9}j_{-1}^{2} = \frac{8}{9} + \frac{1}{9} = 1$

- Find $P(0 < X < 1)$
  - $P(0 < X < 1) = \int_{0}^{1} \frac{x^2}{3}dx = \frac{x^3}{9}j_{0}^{1} = \frac{1}{9}$

- **Definition 3.7**:

> The **cumulative function** $F(x)$ of a continuous random variable X with density function $f(x)$ is
>
> $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \ for \ -\infty < x < \infty$$

- An immediate consequence:
  - $P(a < X < b) = F(b) - F(a)$
  - $f(x) = \frac{dF(x}{dx})$, if the derivative exists

Example 3.12: For the density function of Example 3.6 find $F(x)$, and use it to evaluate $P(0 < X \leq 1)$.

For $-1 < x < 2$

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{t^2}{3}dt$$

$$= \frac{t^3}{9}j_{-1}^{x} = \frac{x^3+1}{9}$$

$$F(x) = \left\{ \begin{array}{l} 0, \ x \leq -1 \\ \frac{x^3+1}{9}, -1 \leq x < 2 \\ 1, x \geq 2 \end{array} \right\}$$

$$P(0 < X \leq 1) = F(1) - F(0)$$
$$= \frac{2}{9} - \frac{1}{9} = \frac{1}{9}$$

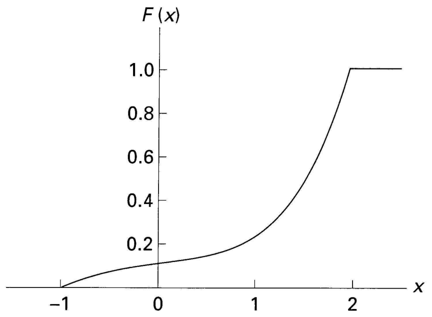

Figure 5: Continuous cumulative distribution function.

1.4 Joint Probability Distribution

- In some experiment, we might want to study simultaneous outcomes of several random variables.
- If X and Y are two discrete random variables, the probability distribution for their simultaneous occurrence can be represented by a function with values $f(x, y)$
- **Definition 3.8**:

The function $f(x, y)$ is a **joint probability distribution** (or **probability mass function**) of the discrete random variables X and Y if

1. $f(x, y) \geq 0$, for all (x, y)
2. $\sum_x \sum_y f(x, y) = 1$
3. $P(X = x, Y = y) = f(x, y)$

For any region A in the xy-plane,

$$P[(X, Y) \in A] = \sum \sum_{A} f(x, y)$$

- **Example 3.14**: Two refills for a ballpoint pen are selected at random from a box that contains 3 blue refills, 2 red refills, and 3 green refills. If X is the number of blue refills and Y is the number of red refills selected, find

- the joint probability function $f(x, y)$

$$f(x,y) = \frac{\binom{3}{x}\binom{2}{y}\binom{3}{2-x-y}}{\binom{8}{2}}$$

- $P[(X, Y) \in A]$, where A is the region $\{(x, y)|x + y \leq 1\}$.

$$P[(X,Y) \in A] = P(X+Y \leq 1)$$
$$= f(0,0) + f(0,1) + f(1,0)$$
$$= \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}$$

$f(x,y)$		x 0	1	2	Row Totals
	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
y	1	$\frac{3}{14}$	$\frac{3}{14}$		$\frac{3}{7}$
	2	$\frac{1}{28}$			$\frac{1}{28}$
Column Totals		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

- **Definition 3.9**:

The function $f(x, y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x, y) \geq 0$, for all (x, y)
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$
3. $P[(X, Y) \in A] = \int\int_A f(x, y)dxdy$

For any region A in the xy-plane,

- **Example 3.15**: A candy company distributes boxes of chocolates with a mixture of creams, toffees, and nuts coated in both light and dark chocolate.

- For randomly selected box, let X and Y, respectively, be the proportions of the light and dark chocolates that are creams.
- The joint density function is as follows:

$$f(x,y) = \left\{ \begin{array}{l} \frac{2}{5}(2x+3y),\ 0 \le x \le 1,\ 0 \le y \le 1 \\ 0,\ elsewhere \end{array} \right\}$$

- Verify $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_0^1\int_0^1 \frac{2}{5}(2x+3y)dxdy$$
$$= \int_0^1 (\frac{2x^2}{5} + \frac{6xy}{5})|_{x=0}^{x=1} dy = \int_0^1 (\frac{2}{5} + \frac{6y}{5})dy = (\frac{2y}{5} + \frac{3y^2}{5})|_0^1$$
$$= \frac{2}{5} + \frac{3}{5} = 1$$

- $P[(X,Y) \in A]$, where A is the region $(x,y)|0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}$,

$$P[(X,Y) \in A] = P(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{1}{2})$$
$$= \int_{\frac{1}{4}}^{\frac{1}{2}}\int_0^{\frac{1}{2}} \frac{2}{5}(2x+3y)dxdy = \int_{\frac{1}{4}}^{\frac{1}{2}} (\frac{2x^2}{5} + \frac{6xy}{5})|_{x=0}^{x=\frac{1}{2}} dy$$
$$= \int_{\frac{1}{4}}^{\frac{1}{2}} (\frac{1}{10} + \frac{3y}{5})dy = (\frac{y}{10} + \frac{3y^2}{10})|_{\frac{1}{4}}^{\frac{1}{2}} = \frac{13}{160}$$

- **Definition 3.10**:

The **marginal distributions** of X alone and of Y alone are

$$g(x) = \sum_y f(x,y)\ and\ h(y) = \sum_x f(x,y)$$

for the discrete case

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy\ and\ h(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

for the continuous case

Example 3.16: Show that the column and row totals of the following table give the marginal distribution of X alone and of Y alone.

$f(x,y)$		x 0	1	2	Row Totals
	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
y	1	$\frac{3}{14}$	$\frac{3}{14}$		$\frac{3}{7}$
	2	$\frac{1}{28}$			$\frac{1}{28}$
Column Totals		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Solution:

$$P(X=0)=g(0)=\sum_{y=0}^{2} f(0,y)=f(0,0)+f(0,1)+f(0,2)$$
$$=\frac{3}{28}+\frac{3}{14}+\frac{1}{28}=\frac{5}{14}$$
$$P(X=1)=g(1)=\sum_{y=0}^{2} f(1,y)=f(1,0)+f(1,1)+f(1,2)$$
$$=\frac{9}{28}+\frac{3}{14}+0=\frac{15}{28}$$
$$P(X=2)=g(2)=\sum_{y=0}^{2} f(2,y)=f(2,0)+f(2,1)+f(2,2)$$
$$=\frac{3}{28}+0+0=\frac{3}{28}$$

x	0	1	2
g(x)	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$

- **Example 3.17**: Find $g(x)$ and $h(y)$ for the following joint density function.

$$f(x,y)=\left\{ \begin{array}{l} \frac{2}{5}(2x+3y),\ 0\le x\le 1,\ 0\le y\le 1 \\ 0,\ elsewhere \end{array} \right\}$$

- $g(x)$

$$=\int_{-\infty}^{\infty} f(x,y)dy=\int_{0}^{1}\frac{2}{5}(2x+3y)dy$$
$$=(\frac{4xy}{5}+\frac{6y^2}{10})j_{y=0}^{y=1}=\frac{4x+3}{5}$$

for $0\le x\le 1,\ 0\le y\le 1$ and $g(x)=0$, elsewhere

- $h(y)$

$$= \int_{-\infty}^{\infty} f(x,y)dx = \int_0^1 \frac{2}{5}(2x+3y)dx$$
$$= (\frac{2x^2}{5} + \frac{6yx}{5})j_{x=0}^{x=1} = \frac{2+6y}{5}$$
for $0 \le x \le 1,\ 0 \le y \le 1$ and $h(y) = 0$, elsewhere

- **Definition 3.11**:

Let X and Y be two random variables, discrete or continuous. The **conditional distribution** of the random variable Y, given that $X = x$, is
$$f(y|x) = \frac{f(x,y)}{g(x)}, g(x) > 0$$
Similarly, the conditional distribution of the random variable X, given that $Y = y$, is
$$f(x|y) = \frac{f(x,y)}{h(y)}, h(y) > 0$$

- Evaluate the probability that X falls between a and b given that Y is known.

$$P(a < X < b|Y = y) = \sum_x f(x|y),\ for\ the\ discrete\ case$$

$$P(a < X < b|Y = y) = \int_a^b f(x|y),\ for\ the\ continuous\ case$$

- **Example 3.18**: Referring to Example 3.14, find the conditional distribution of X, given that $Y = 1$, and use it to determine $P(X = 0|Y = 1)$.

- Solution:

$$h(y=1) = \sum_{x=0}^{2} f(x,1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}$$
$$f(x|1) = \frac{f(x,1)}{h(1)} = \frac{7}{3}f(x,1), x = 0,1,2$$
$$f(0|1) = \frac{7}{3}f(0,1) = \frac{7}{3} * \frac{3}{14} = \frac{1}{2}$$
$$f(1|1) = \frac{7}{3}f(1,1) = \frac{7}{3} * \frac{3}{14} = \frac{1}{2}$$

$$f(2|1) = \frac{7}{3} f(2,1) = \frac{7}{3} * 0 = 0$$

$$\Longrightarrow P(X = 0|Y = 1) = f(0|1) = \frac{1}{2}$$

x	0	1	2
f(x—1)	$\frac{1}{2}$	$\frac{1}{2}$	0

- **Example 3.19**: The joint density for the random variables (X, Y), where X is the unit temperature change and Y is the proportion of spectrum shift that a certain atomic particle produces is

$$f(x, y) = \left\{ \begin{array}{l} 10xy^2, \ 0 < x < y < 1 \\ 0, \ elsewhere \end{array} \right\}$$

- Find the marginal densities $g(x), h(y)$, and the conditional density $f(y|x)$.

$$g(x) = \int_{-\infty}^{\infty} f(x, y)dy = \int_{x}^{1} 10xy^2 dy = \frac{10x(1 - x^3)}{3}$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y)dx = \int_{0}^{y} 10xy^2 dx$$

$$f(y|x) = \frac{f(x, y)}{g(x)} = \frac{10xy^2}{\frac{10x(1-x^3)}{3}} = \frac{3y^2}{(1 - x^3)}$$

- Find the probability that the spectrum shifts more than half of the total observations, given the temperature is increased to 0.25 unit.

$$P(Y > \frac{1}{2}|X = 0.25) = \int_{1/2}^{1} f(y|0.25)dy = \int_{1/2}^{1} \frac{3y^2}{(1 - 0.25^3)} dy = \frac{8}{9}$$

- **Definition 3.12**:

Let X and Y be two random variables, discrete or continuous, with joint probability distribution $f(x, y)$ and marginal distributions $g(x)$ and $h(y)$, respectively. The random variables X and Y are said to be **statistically independent** if and only if

$$f(x, y) = g(x)h(y), \ for \ all \ (x, y) \ within \ their \ range$$

- **Example 3.21**: Show that the random variables of Example 3.14 are not statistically independent.

$$f(0, 1) = \frac{3}{14}, g(0) = \sum_{y=0}^{2} f(0, y) = \frac{5}{14}, \ h(1) = \sum_{x=0}^{2} f(x, 1) = \frac{3}{7}$$

$$\Longrightarrow f(0,1) \neq g(0) * h(1)$$

therefore X and Y are not statistically independent.

- **Example**: In a binary communications channel, let X denote the bit sent by the transmitter and let Y denote the bit received at the other end of the channel. Due to noise in the channel we do not always have $Y = X$. A joint probability distribution is given as

		x 0	1	$h(y)$
y	0	0.45	0.03	0.48
	1	0.05	0.47	0.52
	$g(x)$	0.5	0.5	

		x 0	1	$h(y)$
y	0	f(0,0)	f(1,0)	h(0)
	1	f(0,1)	f(1,1)	h(1)
	$g(x)$	g(0)	g(1)	

- X and Y are not independent because

$$f(0,0) \neq g(0)h(0) \Longrightarrow 0.45 \neq 0.5 * 0.48$$

- $P(X = x, Y = y) = P[(X = x) \cap (Y = y)]$: it is the probability that $X = x$ and $Y = y$ simultaneously.
- $f(0,0) = P(X = 0, Y = 0) = P[(X = 0) \cap (Y = 0)]$
- So $g(0) = P[X = 0]$

$$= P[(X = 0) \cap (Y = 0)] + P[(X = 0) \cap (Y = 1)] = f(0,0) + f(0,1)$$

- $\Longrightarrow P[Y = 0|X = 0] = \frac{P[(X=0)\cap(Y=0)]}{P[X=0]} = \frac{f(0,0)}{g(0)}$
- Sent 0 & Received 0: NO error.

$$P[Y = 0|X = 0] = \frac{f(0,0)}{g(0)} = \frac{0.45}{0.9} = 0.9$$

- Sent 1 & Received 0: ERROR

$$P[Y = 0|X = 1] = \frac{f(1,0)}{g(1)} = \frac{0.03}{0.5} = 0.06$$

- Sent 0 & Received 1: ERROR

$$P[Y=1|X=0] = \frac{f(0,1)}{g(0)} = \frac{0.05}{0.5} = 0.1$$

- Sent 1 & Received 1: NO error.

$$P[Y=1|X=1] = \frac{f(1,1)}{g(1)} = \frac{0.47}{0.5} = 0.94$$

- Notice that

$$P[Y=0|X=0] + P[Y=1|X=0] = 1$$

$$P[Y=0|X=1] + P[Y=1|X=1] = 1$$

Definition 3.13:

Let $X_1, X_2, \ldots, X_n$ be n random variables, discrete or continuous, with joint probability distribution $f(x_1, x_2, \ldots, x_n)$ and marginal distributions $f(x_1), f(x_2), \ldots, f(x_n)$, respectively. The random variables $X_1, X_2, \ldots, X_n$ are said to be **mutually statistically independent** if and only if

$$f(x_1, x_2, \ldots) = f_1(x_1) f_2(x_2) \ldots f_n(x_n)$$

for all $(x_1, x_2, \ldots, x_n)$ within their range.

- **Example 3.22**: Suppose that the shelf life, in years, of a certain perishable food product packaged in cardboard containers is a random variable whose probability density function is given by

$$f(x,y) = \left\{ \begin{array}{l} e^{-x},\ x > 0 \\ 0,\ elsewhere \end{array} \right\}$$

- Let $X_1, X_2, \ldots, X_n$ represent the shelf lives for three of these containers selected independently and find $P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$

- Solution:

$$f(x_1, x_2, x_3) = f(x_1) f(x_2) f(x_3) = e^{-x_1 - x_2 - x_3}$$

  for $x_1, x_2, x_3 > 0$ and $f(x_1, x_2, x_3) = 0$ elsewhere

$$P(X_1 < 2, 1 < X_2 < 3, X_3 > 2) = \int_2^{\infty} \int_1^3 \int_0^2 e^{-x_1 - x_2 - x_3} dx_1 dx_2 dx_3$$

$$= (1 - e^{-2})(e^{-1} - e^{-3}) e^{-2} = 0.0372$$