STA250 Probability and Statistics

Chapter 4 Notes

Discrete and Continuous Random Variables and Their Probability Functions

Asst. Prof. Abdullah YALÇINKAYA

Ankara University, Faculty of Science, Department of Statistics

ayalcinkaya@ankara.edu.tr

2023

STA250 Probability and Statistics

Reference Book

This lecture notes are prepared according to the contents of

"**PROBABILITY & STATISTICS FOR ENGINEERS & SCIENTISTS** by Walpole, Myers, Myers and Ye"

A Random Variable

- **A random variable is a function that associates a real number with each element in the sample space. Consequently, a random variable can be used to identify numerical events that are of interest in an experiment.**
- **If X denote a random variable, we shall use a capital letter. Its corresponding small letter , x in this case, for one of its values.**
- ***X* random variable is shown as follows,**

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

$$w \ \rightarrow X(w)$$

where:

D_X: The set of X values.

There are two types of random variables:

Discrete and Continuous Random Variables.

A Random Variable

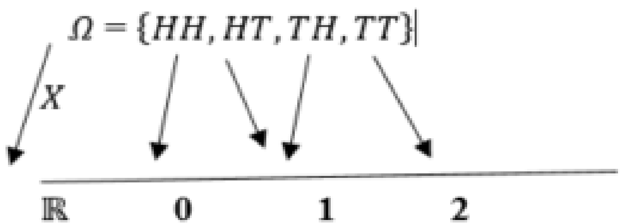
- **Two balls** are drawn in succession without replacement from an urn containing 4 red (R) balls and 3 black (B) balls.
- The possible outcomes and the values y of the random variable Y , where Y is the number of red balls, are

Sample Space	y
RR	2
RB	1
BR	1
BB	0

A Random Variable

Example 1: Flip a coin two times.
Let random variable X denote the number of tails in each sample point.
A = even of at least 1 Tail (T) occurring.

Sample Space:
$S = \{TT, TH, HT, HH\}$
n(S) = 4 (Number of Elements)



$$D_X = \{0, 1, 2\}$$

$$P(X > 2) = 0$$

$$P(X \geq 1) = P(X = 1) + P(X = 2)$$
$$= P(\{TH, HT\}) + P(\{TT\})$$
$$= 2/4 + 1/4 = 3/4$$

A Random Variable

- **Example 2: Components can either be defective (D) or not (N).**
  - What is the sample space for this situation?
    $S = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$
  - Let random variable X denote the number of **defective components** in each sample point.

  - $P(X \leq 2) = ?$ $P(X = 0) + P(X = 1) + P(X = 2)$

What are the values of $P(X = x), for\ x = 0, 1, 2, 3$?

- $D_X = \{0,1,2,3\}$ its values of X random variable.
- *NNN (1 of 8) → X=0;*
- *NND, NDN, DNN (3 of 8) → X=1*
- *NDD, DND, DDN (3 of 8)→ X=2;*
- *DDD (1 of 8) → X=3*
- $P(X = 0) = \frac{1}{8} = .125$
- $P(X = 1) = \frac{3}{8} = .375$
- $P(X = 2) = \frac{3}{8} = .375$
- $P(X = 3) = \frac{1}{8} = .125.$

A Random Variable

- **A discrete sample space has a finite or countably infinite number of points (outcomes).**
- **Example of countably infinite: experiment consists of flipping a coin until a heads occurs.**
  - S = ?
  - S = {H, TH, TTH, TTTH, TTTTH, TTTTTH, …}
  - S has a countably infinite number of sample points.

Discrete Probability Distribution

- **The probability that X takes on the value x, $P(X = x)$, is defined as the sum of the probabilities of all sample points in S that are assigned the value x. It is denoted $f(x) = P(X = x)$.**

- **The set of ordered pairs $(x, f(x))$ is called probability function or probability function of X.**

- **For any discrete probability distribution, the following must be true**
  - $f(x) \geq 0 \qquad x \in D_x$
  - $\sum_{x \in D_x} f(x) = 1$
  - $P(X = x) = f(x)$

Cumulative Distribution & Plotting

- **The cumulative distribution function, denoted *F(x)*, of a discrete random variable *X* with probability distribution *f(x)* is**
  - $F(x) = P(X \leq x)$

  ***F(x)* is calculated as follows,**
  - $F(x) = \sum_{t \leq x} f(t)$
- **It is useful to plot both a probability distribution and the corresponding cumulative distribution.**
  - Typically, the values of f(x) versus x are plotted using a <u>probability histogram</u>.
  - Cumulative distributions are also plotted using a similar type of histogram/step function.

Example 1

- **The probability function of X random variable is given as;**
  - $f(x) = cx^2 \quad D_X = \{-2, -1, 1, 2\}$

a) $\boldsymbol{c} = ?$

b) **Obtain the probability function, table.**

c) **Find the probabilities.**

$P(X > 2) = ?$ $\quad P(X \geq 1) = ?$ $\quad P(0 < X \leq 2) = ?$

Solution 1

a) $\sum_{-2}^{2} cx^2 = 1 \rightarrow c = \frac{1}{10}$

- **The probability function of X random variable is given as;**
  - $f(x) = \dfrac{x^2}{10}$ $\quad D_X = \{-2, -1, 1, 2\}$

b) The probability Function and Table

$$f(x) = \begin{cases} \frac{4}{10}, & x = -2 \\ \frac{1}{10}, & x = -1 \\ \frac{1}{10}, & x = 1 \\ \frac{4}{10}, & x = 2 \end{cases}$$

$X = x$	-2	-1	1	2
$P(X = x)$	$\frac{4}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{4}{10}$

c) $P(X > 2) = 0$

$$P(X \geq 1) = P(X = 1) + P(X = 2) = f(1) + f(2) = \frac{1}{10} + \frac{4}{10} = \frac{5}{10}$$

$$P(0 < X \leq 2) = P(X = 1) + P(X = 2) = f(1) + f(2) = 5/10$$

Discrete Probability Distribution

- If a car agency sells 50% of its inventory of a certain foreign car equipped with side airbags, find a formula for the probability distribution of the number of cars with side airbags among the next 4 cars sold by the agency.
- ***Solution* :** Since the probability of selling an automobile with side airbags is 0.5, the $2^4 = 16$ points in the sample space are equally likely to occur. Therefore, the denominator for all probabilities, and also for our function, is 16.
- To obtain the number of ways of selling 3 cars with side airbags, we need to consider the number of ways of partitioning 4 outcomes into two cells, with 3 cars with side airbags assigned to one cell and the model without side airbags assigned to the other. This can be done in $\binom{4}{3} = 4$ ways.
- In general, the event of selling x models with side airbags and $4-x$ models without side airbags can occur in $\binom{4}{x}$ ways, where x can be 0, 1, 2, 3, or 4.
- Thus, the probability distribution $f(x) = P(X = x)$ is

$f(x) = \frac{1}{16}\binom{4}{x}$, for $x = 0, 1, 2, 3, 4$.

Example 2

- The probability function
  - $f(x) = P(X = x) = \frac{1}{16}\binom{4}{x}$ $\qquad x = 0,1,2,3,4$
  - Using $F(x)$, verify that $f(2) = 3/8$.

Find the cumulative distribution function of the random variable X.

$$f(0) = \frac{1}{16} \quad f(1) = \frac{1}{4} \quad f(2) = \frac{3}{8} \quad f(3) = \frac{1}{4} \quad f(4) = \frac{1}{16}$$

Solution 2

$$F(0) = f(0) = \frac{1}{16},$$

$$F(1) = f(0) + f(1) = \frac{5}{16},$$

$$F(2) = f(0) + f(1) + f(2) = \frac{11}{16},$$

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

$$F(4) = f(0) + f(1) + f(2) + f(3) + f(4) = 1.$$

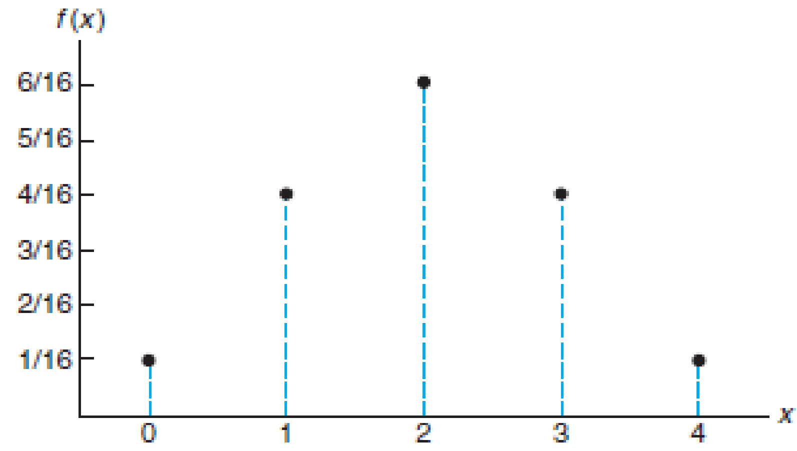


Figure 3.1: Probability mass function plot.

$$F(x) = \begin{cases} 0, & \text{for } x < 0, \\ \frac{1}{16}, & \text{for } 0 \le x < 1, \\ \frac{5}{16}, & \text{for } 1 \le x < 2, \\ \frac{11}{16}, & \text{for } 2 \le x < 3, \\ \frac{15}{16}, & \text{for } 3 \le x < 4, \\ 1 & \text{for } x \ge 4. \end{cases}$$

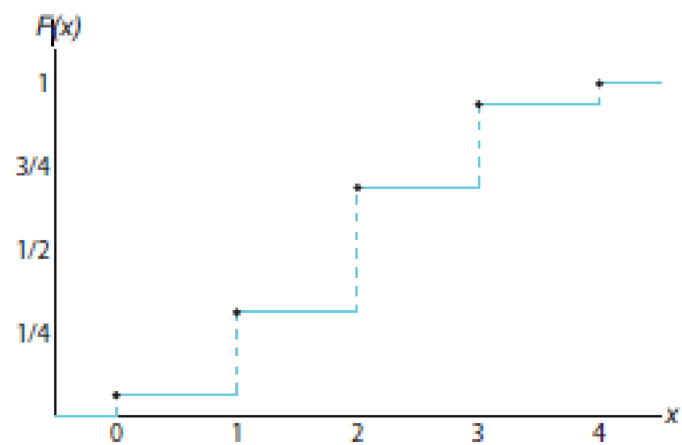


Figure 3.3: Discrete cumulative distribution function.

$$f(2) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{3}{8}.$$

Continuous Probability Distributions

- **A continuous random variable has a probability of 0 of assuming *exactly* any of its values.**
  - $P(X = x) = 0.$
  - Otherwise the probabilities couldn't sum to 1.

- **That is, it does not matter whether we include an endpoint of the interval or not. This is not true, though, when X is discrete.**
  - Since the probability of any individual point is 0,
    $P(a < X < b) = P(a \le X \le b)$
    on, the endpoints can be included or not.

- **In dealing with continuous variables, $f(x)$ is usually called <u>the probability density function</u>, or simply the density function, of X.**

Continuous Probability Distributions

- The probability that X assumes a value between a and b is equal to the shaded area under the density function between the ordinates at $x = a$ and $x = b$, and from integral calculus is given by

$$P(a < X < b) = \int_a^b f(x)\,dx.$$

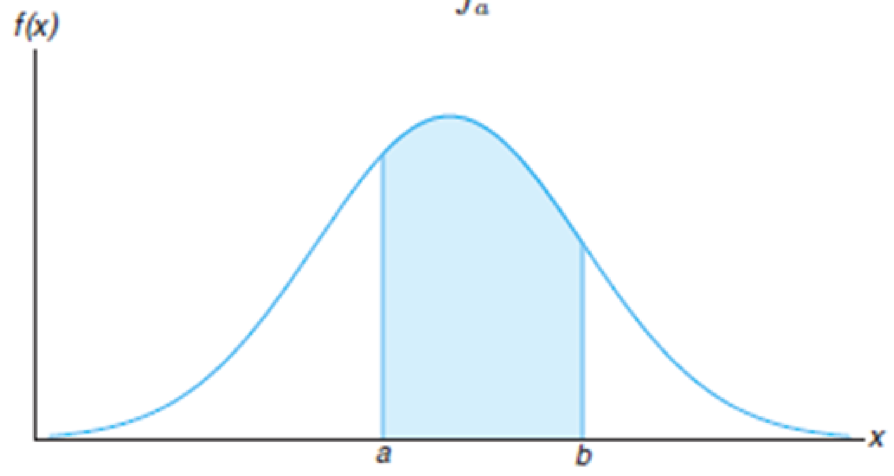


Figure 3.5: $P(a < X < b)$.

Continuous Distributions

- The function *f(x)* is a probability density function (pdf) for the continuous random variable *X*, defined over the set of real numbers, if
  1. $f(x) \geq 0, for\ all\ x \in R$
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$
  3. $P(a < X < b) = \int_{a}^{b} f(x)dx.$

- The cumulative distribution function *F(x)* of a continuous random variable *X* with density function *f(x)* is
  - $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \qquad -\infty < x < \infty$
  - $P(a < X \leq b) = F(b) - F(a)$
    - If discrete, must use "$a < X$", and not "$a \leq X$", above.
  - $f(x) = \frac{dF(x)}{dx}$

Example 1

- **Suppose that the error in the reaction temperature, in C, for a controlled laboratory experiment is a continuous random variable X having the probability density function**

$$f(x) = \begin{cases} \dfrac{x^2}{3}, & -1 < x < 2, \\ 0\,, & elsewhere \end{cases}$$

  - Verify that $f(x)$ is a density function.
  - Find $P(0 < X \leq 1)$.

- **Solution:**
  - Obviously, $f(x) \geq 0$. To verify condition $\int_{-\infty}^{\infty} f(x)dx = 1$.
  - $\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{2} \frac{x^2}{3} dx = \frac{x^3}{9}(2|-1) = \frac{8}{9} + \frac{1}{9} = 1.$

Example 1

$$f(x) = \begin{cases} \dfrac{x^2}{3}, & -1 < x < 2, \\ 0, & elsewhere \end{cases}$$

- Find $P(0 < X \leq 1)$.

Solution:

$$P(0 < X \leq 1) = \int_0^1 \frac{x^2}{3} dx = \frac{x^3}{9} | = \frac{1}{9}$$

Example 2

- **For the density function of Example 1. Find $F(x)$, and use it to evaluate $P(0 < X \leq 1)$.**

Solution: For $-1 < x < 2$,

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-1}^{x} \frac{t^2}{3} dt = \frac{t^3}{9} \Big| = \frac{x^3 + 1}{9}.$$

Therefore,

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{x^3+1}{9}, & -1 < x \leq 2, \\ 1, & x \geq 2. \end{cases}$$

The cumulative distribution function *F(x)* is expressed in Figure 3.6.

Now,

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

Which agrees with the result obtained by using the density function in Example 1.

Example 2

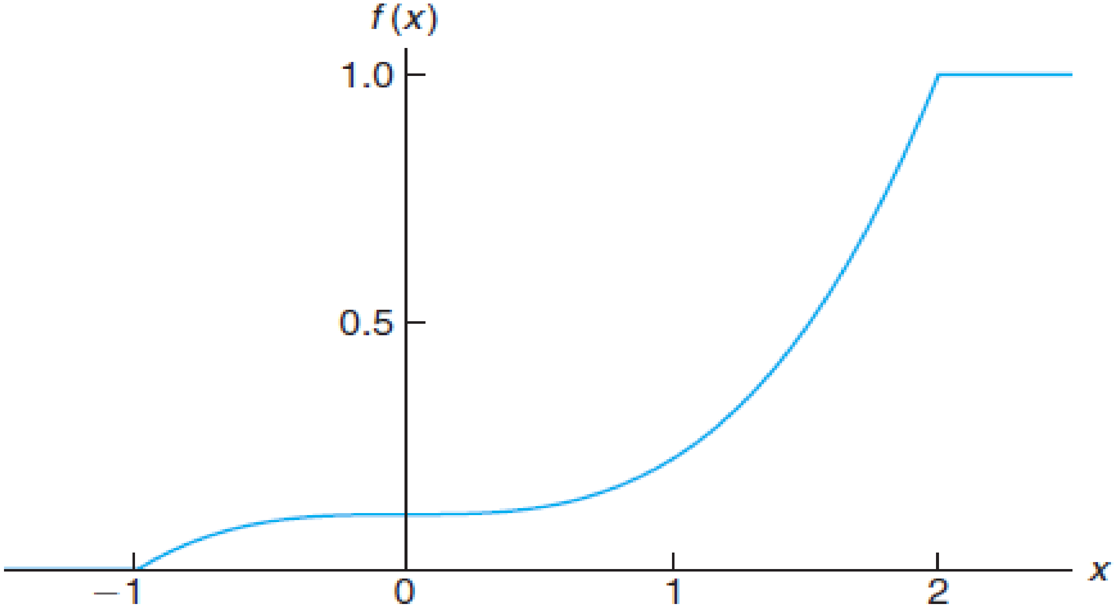


Figure 3.6: Continuous cumulative distribution function.

Example 3

The Department of Energy (DOE) puts projects out on bid and generally estimates what a reasonable bid should be. Call the estimate *b*. The DOE has determined that the density function of the winning (low) bid is

$$f(y) = \begin{cases} \dfrac{5}{8b}, & \dfrac{2}{5}b \le y \le 2b \\ 0, & elsewhere \end{cases}$$

- Find $F(y)$ and use it to determine the probability that the winning bid is less than the DOE's preliminary estimate b.
- **Solution:** For $\frac{2b}{5} \le y \le 2b$;

$$F(y) = \int_{2b/5}^{y} \frac{5}{8b} dy = \frac{5t}{8b}\bigg| = \frac{5y}{8b} - \frac{1}{4}$$

Thus;

$$F(y) = \begin{cases} 0, & y < \dfrac{2}{5}b \\ \dfrac{5y}{8b} - \dfrac{1}{4}, & \dfrac{2}{5}b \le y < 2b \\ 1, & y \ge 2b \end{cases}$$

To determine the probability that the winning bid is less than the preliminary bid estimate *b*, we have

$$P(Y \le b) = F(b) = \frac{5}{8} - \frac{1}{4} = \frac{3}{8}$$

Joint Probability Distributions

- **Given a pair of discrete random variables on the same sample space, X and Y, the joint probability distribution of X and Y is**

$$f(x, y) = P(X = x, Y = y)$$

$f(x, y)$ equals the probability that both x and y occur.

- **The usual rules hold for joint probability distributions:**
  - $f(x, y) \geq 0$ for all (x, y)
  - $\sum_x \sum_y f(x, y) = 1$
  - For any region A in the xy plane,
    $P[(X, Y) \in A] = \sum \sum_A f(x, y)$
- **For continuous joint probability distributions, the sums above are replaced with integrals.**

Joint Probability Distributions

- **Two ballpoint pens** are selected at random from a box that contains **3 blue pens**, **2 red pens**, and **3 green pens**.
- If ***X* is the number of blue pens** selected and ***Y* is the number of red pens** selected, find
- (a) the joint probability function $f(x, y)$,
- (b) $P[(X, Y) \in A]$, where A is the region $\{(x, y)|x + y \leq 1\}$.
- ***Solution* :** The possible pairs of values (x, y) are (0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0).
  - Because of selecting Two pens, $x + y$ must be equal 2.
- (a) Now, $f(0, 1)$, for example, represents the probability that a red and a green pens are selected. The total number of equally likely ways of selecting any 2 pens from the 8 is $\binom{8}{2}= 28$.
- The number of ways of selecting 1 red from 2 red pens and 1 green from 3 green pens is $\binom{2}{1}\binom{3}{1} = 6$.
- Hence, $f(0, 1) = \frac{6}{28} = \frac{3}{14}$.

Joint Probability Distributions

- Similar calculations yield the probabilities for the other cases, which are presented in Table. Note that the probabilities sum to 1.

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

it will become clear that the joint probability distribution of Table can be represented by the formula

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

for $x = 0,1,2;\ y = 0,1,2;\ 0 \le x + y \le 2.$

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

The probability that (X,Y) fall in the region A is

$$P[(X,Y) \in A] = P(X+Y \le 1) = f(0,0) + f(0,1) + f(1,0)$$
$$= 3/28 + 3/14 + 9/28 = 9/14$$

Marginal Distributions

- **The <u>marginal distribution</u> of X alone or Y alone can be calculated from the joint distribution function as follows:**
  - $g(x) = \sum_y f(x,y)$ and $h(y) = \sum_x f(x,y)$ if discrete
  - $g(x) = \int_y f(x,y)dy$ and $h(y) = \int_x f(x,y)dx$ if continuous
- **In other words, for example, $g(x) = P(X = x)$ is the sum (or integral) of $f(x,y)$ over all values of y.**

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

Show that the column and row totals of Table give the marginal distribution of *X* alone and of *Y* alone.

Marginal Distributions

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

Show that the column and row totals of Table give the marginal distribution of *X* alone and of *Y* alone.

Solution **:** For the random variable *X*, we see that

$$g(0) = f(0,0) + f(0,1) + f(0,2) = 3/28 + 3/14 + 1/28 = 5/14,$$
$$g(1) = f(1,0) + f(1,1) + f(1,2) = 9/28 + 3/14 + \ 0 \ = 15/28,$$
$$g(2) = f(2,0) + f(2,1) + f(2,2) = 3/28 + \ 0 \ + \ 0 \ = 3/28,$$

which are just the column totals of Table. In a similar manner we could Show that the values of *h*(*y*) are given by the row totals. In tabular form, these marginal distributions may be written as follows:

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

y	0	1	2
$h(y)$	$\frac{15}{28}$	$\frac{3}{7}$	$\frac{1}{28}$

Conditional Distributions

- **For either discrete or continuous random variables, X and Y,** the conditional distribution of Y given that X = x is

$$f(y \mid x) = f(x, y)/g(x) \quad \text{if} \quad g(x) > 0$$

- and the conditional distribution of X given that Y=y is

$$f(x \mid y) = f(x, y)/h(y) \quad \text{if} \quad h(y) > 0$$

- **X and Y are statistically independent if**

$$f(x, y) = g(x)\, h(y)$$

for all x and y within their range.

- **A similar equation holds for n mutually statistically independent jointly distributed random variables.**

Statistical Independence

- **The definition of <u>independence</u> is as before:**
  - Previously, P (A | B) = P (A) and P (B | A) = P (B).
  - How about terms of the conditional distribution?
  - $f(x \mid y) = g(x)$ and $f(y \mid x) = h(y)$.
  - The other way to demonstrate independence?
  - $f(x, y) = g(x)\, h(y) \qquad \forall\, x, y$ in range.
- **Similar formulas also apply to more than two mutually independent random variables.**

Conditional Distribution

- If we wish to find the probability that the discrete random variable X falls between a and b when it is known that the discrete variable $Y = y$, we evaluate

$$P(a < X < b \mid Y = y) = \sum_{a<x<b} f(x|y)$$

- where the summation extends over all values of X between a and b. When X and Y are continuous, we evaluate

$$P(a < X < b \mid Y = y) = \int_a^b f(x|y)dx$$

Conditional Distribution

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

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

y	0	1	2
$h(y)$	$\frac{15}{28}$	$\frac{3}{7}$	$\frac{1}{28}$

Find the conditional distribution of X given that $Y = 1$, and use it to determine $P(X = 0 \mid Y = 1)$.

Solution : We need to find $f(x|y)$, where $y = 1$. First, we find that

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

$$\boxed{f(0|1) = \frac{7}{3} f(0,1) = \frac{7}{3}\frac{3}{14} = \frac{1}{2}} \quad \boxed{f(1|1) = \frac{7}{3} f(1,1) = \frac{7}{3}\frac{3}{14} = \frac{1}{2}} \quad \boxed{f(2|1) = \frac{7}{3} f(2,1) = \frac{7}{3} 0 = 0}$$

The conditional distribution of X, given that $Y = 1$,

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

Therefore, if it is known that 1 of the 2 pen refills selected is red, we have a probability equal to 1/2 that the other refill is not blue.

$$P(X = 0 \mid Y = 1) = 1/2$$

Next Lesson

- **Discrete Probability Distributions**

See you☺