Chapter 2

PROBABILITY

2.1 Background

For a discrete random variable, X, it has been emphasized that its mean, μ_X, its variance, σ_X^2, its skewness (β_3) and kurtosis (γ_4) are each constant population parameters which describe respectively a location, spread, lack of symmetry and "peakiness" of the probability distribution of the population of X. In each of the examples in Chapter 1, the population distribution is defined as known. When this is the case, μ_X, σ_X^2, β_3 and γ_4 can be calculated and described, thereby more-or-less completing all that can be said about the location and shape of the population distribution. But, to be sure, it is unusual in practice for the population distribution to be completely known. When a population is not fully known and there is a need to know something additional about it, statistical enquiry proceeds, first, by sampling the population, second, by using information provided by the sample to estimate the population parameters, and third, by the process of statistical inference to expose limits within

which the unknown features of the population most likely lie. To understand how all this may be achieved, it is necessary to understand the theory of probability and its application to sampling, estimation and statistical inference. The ultimate focus of statistical enquiry is on making inferences about a population, based on a subset collected from it, called a sample.

2.2 The Sample Space and Events

The data comprising a sample are to be thought of as generated by an experiment which is repeatable under constant conditions and whose outcomes cannot be predetermined. To fix ideas, consider the simple example of tossing two dice. The possible outcome on any one toss is any one of 36 pairs of numbers $(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,1),(6,2),\ldots,(6,6)$. These points are illustrated in Figure 2.1. Each pair of numbers represents the co-ordinates of a simple event, called a sample point, that is, an event which cannot be decomposed. Thus each simple event corresponds to one, and only one, sample point. The set of all possible sample points forms the sample spa

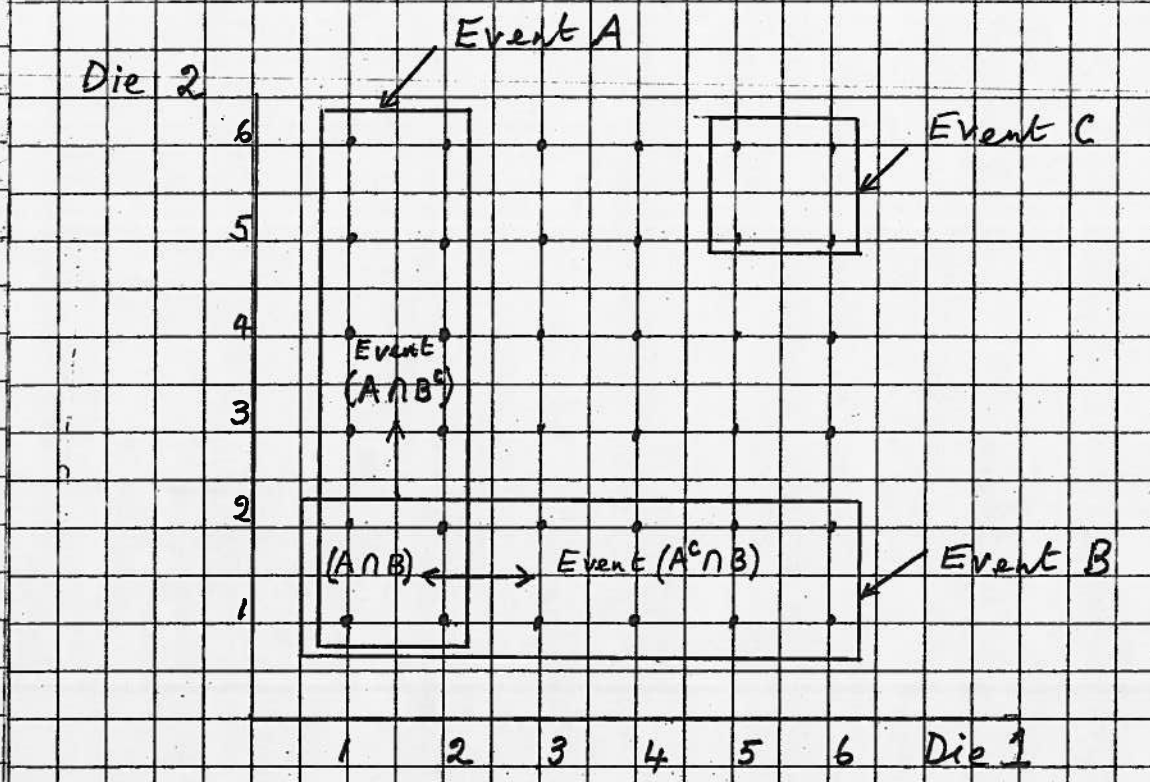


Figure 2.1 : The Sample Space and Events in the Tossing of Two Dice

associated with the Experiment and hence forms the set of all sample points that can arise in the population. The sample space is denoted by S and individual sample points are denoted by s with co-ordinates (s_1, s_2). Thus

$$S = \{s : s = (s_1, s_2);\ s_1 = 1, 2, \ldots, 6;\ s_2 = 1, 2, \ldots, 6\}$$

which reads: S is the set of points s of co-ordinates (s_1, s_2) each of which ranges over the integers $1, 2, \ldots, 6$.

Also illustrated in Figure 2.1 are the compound events A,

B and C. The event A comprises 12 sample points, $a = (a_1, a_2)$, as follows:

$$A = \{a: a = (a_1, a_2),\ a_1 = 1, 2;\ a_2 = 1, 2, \ldots, 6\}.$$

The event A is said to occur if any a in A occurs. Similarly, the event B comprises 12 points b of co-ordinates (b_1, b_2):

$$B = \{b: b = (b_1, b_2);\ b_1 = 1, 2, \ldots, 6;\ b_2 = 1, 2\},$$

and C is defined by

$$C = \{c: c = (c_1, c_2);\ c_1 = 5, 6;\ c_2 = 5, 6\}.$$

It is quite clear from Figure 2.1 that the set of points in Event A intersect with the set of points in Event B, yielding four common sample points. These common sample points define the intersection of events A and B, denoted $A \cap B$:

$$A \cap B = \{i: i = (a_i, b_i);\ a_i = 1, 2;\ b_i = 1, 2\}$$

$A \cap B$ occurs when a sample point belonging to both A and B is observed.

It will be helpful to define an empty set, called the null set, ϕ, that is, a set containing no sample points at all. For example, $A \cap C = \phi$ and $B \cap C = \phi$, because

neither A nor B intersect with . [illegible] events do intersect, they are said to be <u>disjoint or mutually exclusive</u>. It is very important to keep in mind that disjointness or mutual exclusivity is a property of sets.

It is also helpful to define <u>the complement of an event</u>. For example, the complement of A, written A^c or $\bar{A}$, is the event described as: "not A in S". Thus A^c is the event which includes all sample points in S which are not in A:

$$A^c = \{a^c : a^c = (a_1^c, a_2^c);\ a_1^c = 3, 4, 5, 6;\ a_2^c = 1, 2, \ldots, 6\}.$$

Quite clearly,

$$A \cap \bar{A} = \phi = A \cap A^c,$$

$$B \cap \bar{B} = \phi = B \cap B^c,$$

or: an event and its complement are disjoint, because, by definition, they can have no common sample points.

The <u>union of two events</u>, say A and B in Figure 2.1, is denoted by $A \cup B$ and is represented by the 20 individual sample points that fall within the L-shaped perimeter of A and B. Thus $A \cup B$ comprises 8 sample points in $(A \cap B^c)$ plus the 12 sample points

in B, or the 12 points in A plus the 8 in $(A^c \cap B)$:

$$A \cup B = (A \cap B^c) \cup B = A \cup (A^c \cap B).$$

Thus the union of two intersecting events may be written as the union of two disjoint events in two ways. Extending the same idea

$$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B) \qquad (2.1)$$

whereupon $A \cup B$ is written as the union of three disjoint events. This last expression indicates that $A \cup B$ is observed if 'A on its own' $(A \cap B^c$; (8 points) is observed, or 'B on its own' $(A^c \cap B$, (8 points), or 'A and B together' $(A \cap B$; (4 points), making any one of 20 points to be observed.

It should be clear that the union of two disjoint events, say $A \cup C$ or $B \cup C$, will be observed if, respectively, A or C, or B or C is observed. Also $A \cup A^c = S = B \cup B^c = C \cup C^c$.

2.3 The Axioms of Probability

If each of the two dice in Figure 2.1 is perfectly symmetrical, then the same symmetry must apply to the sample space S. Suppose the two dice are tossed 36,000 times. How many times would

each pair of numbers be observed? Given perfect symmetry, it would be expected that each pair of numbers would be observed approximately 1,000 times. Let N represent a large number like 36,000 and let the number of times event A is observed be f_A. Then the relative frequency of observing A in N tosses, given S, is $R(A|S) = (f_A/N)$. Similarly for the event C: $R(C|S) = (f_C/N)$

More generally, what are the properties of relati frequency

Property 1: $R(A|S) = (f_A/N) \geq 0$;

Property 2: $R(S|S) = (N/N) = 1$;

Property 3: If $A \cap C = \phi$,

$$R(A \cup C|S) = \left(\frac{f_A + f_C}{N}\right) = \left(\frac{f_A}{N}\right) + \left(\frac{f_C}{N}\right)$$

$$= R(A|S) + R(C|S).$$

The three axioms of probability are simply idealized statements of relative frequency.

Given a sample space S and any two events defined on it, the probability of any event is $P(\cdot|S)$ which obeys:

Axiom 1: $P(A|S) \geq 0$;

Axiom 2: $P(S|S) = 1$;

Axiom 3: If $A \cap C = \phi$, $P(A \cup C|S) = P(A|S) + P(C|S)$.

2.4 The Law of Addition

From Axioms 1 and 2 in the previous section, it is easy to deduce that, for any arbitrary event A in a sample space S, $0 \leq P(A|S) \leq 1$. When $P(S|S)=1$, S is said to be a certainty. Moreover, from equation (2.1) and Axiom 3, for any two intersecting events A and B in S,

$$P(A \cup B|S) = P(A \cap B^c|S) + P(B|S). \qquad (2.2)$$

But the event A may be written as a disjoint sequence

$$A|S = (A \cap B^c|S) \cup (A \cap B|S)$$

and hence, by Axiom 3,

$$P(A \cap B^c|S) = P(A|S) - P(A \cap B|S). \qquad (2.3)$$

Putting (2.3) into (2.2) there results the Law of Addition:

$$P(A \cup B|S) = P(A|S) + P(B|S) - P(A \cap B|S). \qquad (2.4)$$

Quite clearly, if $A \cap B = \phi$, then (2.4) reduces to Axiom 3 because $P(A \cap B|S) = 0$. In particular, since $A \cap A^c = \phi$ and $A \cup A^c = S$, $P(A^c|S) = 1 - P(A|S)$.

2.5 Conditional Probability and Independence

Returning to the Example of tossing two dice, suppose that the dice have been tossed N times and event A has been observed f_A times. What then is the relative frequency that B has also been observed? The number of times that B has been observed, given that A has

already been observed f_A times, can only be the number of times that $A \cap B$ has been observed, say $f_{A \cap B}$. Thus given A, the relative frequency of observing B, which is denoted $R(B|A)$, is

$$R(B|A) = f_{A \cap B} / f_A = \left(\frac{f_{A \cap B}}{N} \Big/ \frac{f_A}{N} \right)$$

$$= \frac{R(A \cap B|S)}{R(A|S)}.$$

Transforming this last expression into probabilities, there results:

$$P(B|A) = \frac{P(A \cap B|S)}{P(A|S)}$$

which is well defined only if $P(A|S) \neq 0$, whereupon

$$P(A \cap B|S) = P(B|A) P(A|S). \qquad (2.5)$$

Going the other way round, by reversing the roles of B and A,

$$P(A \cap B|S) = P(B|A) P(A|S) = P(A|B) P(B|S), \qquad (2.6)$$

given that $P(B|S) \neq 0$. $P(B|A)$ is the probability that B will be observed, on the condition that A has already occurred, and $P(A|B)$ the probability that A will be observed, on the condition that B has already occurred; $P(A|B)$ and $P(B|A)$ are examples of <u>Conditional probabilities</u>.

Conditional probability leads to a concept which is extremely important in the development of statistical theory, namely, <u>statistical independence</u>. From (2.5) or (2.6), if $P(B|\ ) = P(B|S)$ or $P(A|B)\ P(A|S)$,

then (2.6) reduces to

$$P(A \cap B|S) = P(A|S)\,P(B|S) \tag{2.7}$$

and A and B are said to be <u>independent events</u>. Notice carefully that independence is a property of probabilities which can only be established by satisfying equation (2.7). Students of elementary statistics are often confused between disjointness (or mutual exclusivity) and independence. The former is a property of sets, not probabilities, the latter is a property of probabilities. Events A and B are <u>disjoint if and only if $A \cap B = \phi$</u>, while A and are <u>independent events if and only if equation (2.7) is satisfied.</u>

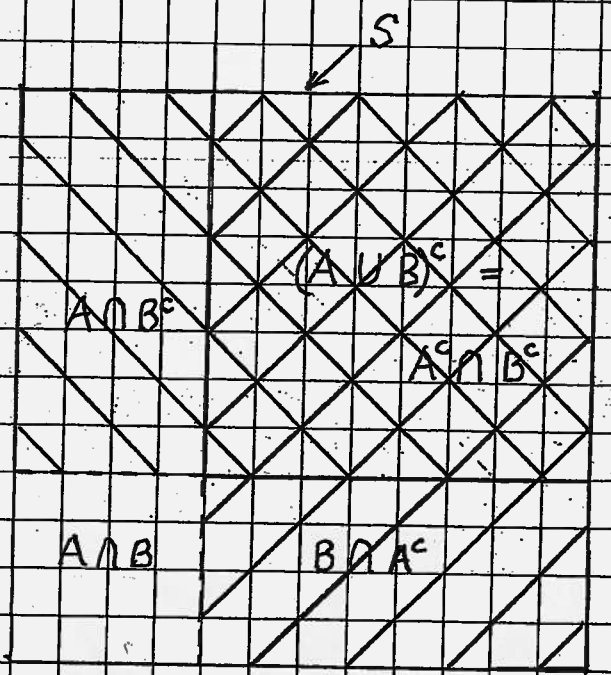
2.6 Complements of Unions and Intersections

It is sometimes necessary to find an expression for the complement of a union or intersection. These will established diagrammatically using a Venn diagram of the two three example. This is displayed in Figure 2.2 which illustrates the theorems that

$$(A \cap B)^c = A^c \cup B^c, \tag{2.8}$$

and

$$(A \cup B)^c = A^c \cap B^c. \tag{2.9}$$



$$(A \cap B)^c = (A \cap B^c) \cup (A^c \cap B^c) \cup (B \cap A^c)$$

$$= A^c \cup B^c$$

$$= \text{▨} + \text{▨}$$

$$(A \cup B)^c = A^c \cap B^c$$

$$= \text{▩}$$

Figure 2.2: $(A \cap B)^c$ and $(A \cup B)^c$ a Venn diagram

Given this pair of results, (2.8) and (9), the following question may be answered: if two events A and B are independent, are their complement, A^c and B^c, independent? By (2.), and dropping conditioning on S, which is henceforth taken as understood,

$$P(A^c \cap B^c) = P[(A \cup B)^c] = 1 - (\cup B)$$

$$= 1 - P(A) - P(B) \quad P(A \cap B), \quad P(A \cap B) \; P(A)P(B),$$

$$= P(A^c) - P(B)(1 \quad P(A))$$

$$= P(A^c)[1 \quad P(B)]$$

$$= P(A^c)P(B) \Rightarrow A \text{ and } B \text{ are independent.}$$

Thus, if A and B are independent so are their complements ^{c} and ^{c}.

If A and B are disjoint, are their complements disjoint? That is if $A \cap B = \phi$, since $A^c \cap B^c = (A \cup B)^c$,

$$P(A \cup B)^c = 1 - P(A \cup B)$$

$$= 1 - P(A) + P(B) + P(A \cap B). \quad (2.10)$$

But $P(A \cap B) = 0$. Therefore, from (2.10),

$$P(A^c \cap B^c) = P(A^c) - P(B)$$

$$= P(A^c) + (1 - P(B^c))$$

$$= P(A^c) + P(B^c) - 1.$$

Thus, if A and B are disjoint, their complements, A^c and B^c, are NOT disjoint, unless $B = A^c$ (whose complement is A), whereupon

$$P(A^c \cap B^c) = P(A^c \cap A) = \phi.$$

2.7 Bayes' Theorem

Let $H_1, H_2, \ldots, H_k$ be a series of disjoint events which cover the entire sample space, that is, $H_i \cap H_j = \phi \ \forall i \neq j$ and $\bigcup_{i=1}^{k} H_i = S$. When these two

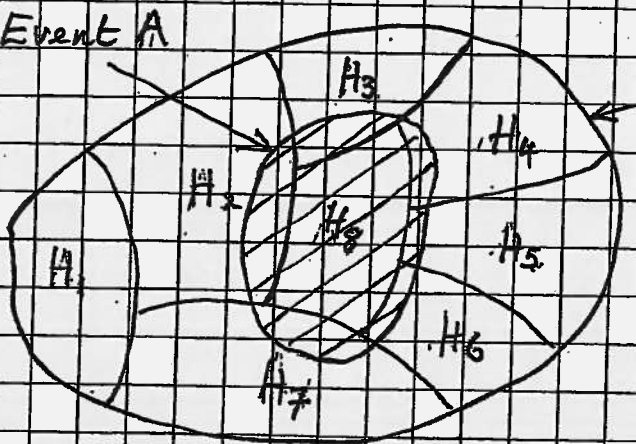


Figure 2.3. A Sample Space for Bayes Theorem.

conditions are met, there is an exhaustive set of disjoint events, and $\sum_{i=1}^{k} P(H_i) = P(S) = 1$. In addition to the exhaustive set of disjoint events, there is another, separate, event A. Then

$$P(A) = \sum_{i=1}^{k} P(A \cap H_i) = \sum_{i=1}^{k} P(A|H_i) P(H_i).$$

Now consider the various H_i as causes leading to the occurrence of the event A, then interest centres on which H_i ($i = 1, 2, \ldots, k$) is the most likely cause of observing A. Consider a particular H_q. Given that A has occurred, the largest $P(H_q|A)$ is required, in order to find the most likely cause of A from among the various $H_1, H_2, \ldots, H_k$. Now,

$$P(H_q|A) = \frac{P(H_q \cap A)}{P(A)} = \frac{P(A|H_q) P(H_q)}{\sum_i P(A|H_i) P(H_i)}, \qquad (2.10)$$

and, since $P(A)$ is fixed by virtue of it being observed,

$$P(H_q|A) \propto P(A|H_q) P(H_q). \qquad (2.11)$$

$P(H_q|A)$ will vary as H_q varies over $H_1, H_2, \ldots, H_k$ with A remaining fixed because it has been observed. This is entirely conventional. On the right-hand side of (2.11), $P(A|H_q)$ also varies with H_q, but no H_q is the conditioning event, while A is the subject whose probability is to be found, notwithstanding it has already been observed. This is clearly unconventional and hence is given the name likelihood, to distinguish it from a conventional conditional probability. The final

element, $P(H_q)$, is called a prior probability, being the probability that H_q is the cause of observing A, before A has in fact been observed.

Both equations (2.10) and (2.11) are statements of Bayes' Theorem, which in (2.11) may be stated as: the posterior probability after A has occurred $[P(H_q|A)]$ is proportional to the likelihood $[P(A|H_q)]$ times the prior probability $[P(H_q)]$. To illustrate an application of Bayes' Theorem, a simple example will now be considered.

Example 2.1

There are three companies which deliver radios to a warehouse: Smith Ltd., Jones Inc., and Snodgrass PLC. Smith delivers 30% of all deliveries, Jones 45% and Snodgrass 25%. If Smith delivers, 2% of his radios are damaged, if Jones delivers, 1.5% are damaged and if Snodgrass delivers, 2.5% of his deliveries are damaged. An inspector goes to the warehouse and randomly selects a radio which turns out to be damaged. Which company delivered the damaged radio and what is the associated probability.

To answer the question let H_1 = Smith, H_2 = Jone and H_3 = Snodgrass. Event A = damaged radios. It follows that

$P(A|H_1)P(H_1) = 0.02 \times 0.30 = 0.006$

$P(A|H_2)P(H_2) = 0.015 \times 0.45 = 0.00675$

$P(A|H_3)P(H_3) = 0.025 \times 0.25 = 0.00625$

$\sum_{i=1}^{3} P(A|H_i)P(H_i) = \text{Sum} = 0.019$

The largest of $P(A|H_i)P(H_i)$ is $P(A|H_2)P(H_2) = 0.00675$, and from these calculations:

$P(H_1|A) = P(A|H_1)P(H_1) / \sum_i P(A|H_i)P(H_i) = (0.006/0.019) = 0.3158$

$P(H_2|A) = P(A|H_2)P(H_2) / \sum_i P(A|H_i)P(H_i) = (0.00675/0.019) = 0.3553$

$P(H_3|A) = P(A|H_3)P(H_3) / \sum_i P(A|H_i)P(H_i) = (0.00625/0.019) = 0.3289$

Thus the most likely company to have delivered the damaged nails is Jones Inc. with $P(H_2|A) = 0.3553$ or roughly 36%. However for Smith Ltd., $P(H_1|A) = 32\%$ and for Snodgrass P C, $P(H_3|A)$ 3%

Thus there is not a great deal of difference between the largest at 36% and the other two at 3% and 3%.

2.8 Bivariate Tables

Many of the problems that arise in elementary statistics are in the form of bivariate table. For example, consider the effect of New Methods of Teaching on Students. efore new methods, students

were assigned as top marks, medium and low marks and, after a period of new Teaching methods, the same students were better off, they did not change or they were worse off. The figures are arranged in Figure 2.1. Each three rows are added together and each

Table 2.1: The Effect of New Teaching Methods on Performance

	BEFORE			
AFTER	Top T	Medium M	Low L	Marginal Totals
Better B	200	80	70	350
No change NC	110	150	80	340
Worse W	80	80	150	310
Marginal Totals	390	310	300	1,000

three columns; and each of the three column totals sum to the grand total of 1,000, as is the sum of the three row totals. The same Table is now converted into relative frequencies or probabilities by dividing each of the cells and each of the marginal totals by the grand total of 1,000. This lead to Table 2.2 from which may be calculated various Compound probabilities. For example:

a) $P(M \cup NC) = P(M) + P(NC) - P(M \cap NC)$

$= 0.31 + 0.34 - 0.15$

$= 0.50$

Table 2.2: The Effect of New Teaching Methods of Performance: Individual & Marginal Probabilities

	BEFORE			
AFTER	Top (T)	Medium (M)	Low (L)	Marginal
Better (B)	0.20 $= P(B \cap T)$	0.08 $= P(B \cap M)$	0.07 $P(B \cap L)$	0.35 $= P(B)$
No change (NC)	0.11 $= P(NC \cap T)$	0.15 $= P(NC \cap M)$	0.08 $= P(NC \cap L)$	0.34 $= P(NC)$
Worse (W)	0.08 $= P(W \cap T)$	0.08 $= P(W \cap M)$	0.15 $= P(W \cap L)$	0.31 $= P(W)$
Marginal	0.39 $= P(T)$	0.31 $= P(M)$	0.30 $= P(L)$	1.00

b) $P(B|L) = \frac{P(B \cap L)}{P(L)} = \frac{0.07}{0.30} = 0.2\dot{3}$;

c) $P(T|B) = \frac{P(B \cap T)}{P(B)} = \frac{0.20}{0.35} = 0.5714$.

d) Are W and L independent events? $P(W \cap L) = 0.15$ and $P(W)P(L) = 0.31 \times 0.30 = 0.093$. Since $P(W \cap L) \neq P(W)P(L)$ W and L are not independent.

If now instead of classifications, like Top, Medium, Low and Better, No change and Worse, actual numbers are assigned, then the bivariate table permits bivariate expectations to be calculated... For example, consider the number of credit cards a person has and the number of credit purchases the same person has made in a week. These are displayed in Figure 2.3, with X

representating the number of credit cards and Y representing the number of purchases in a week.

Fig 2.3: The Number of Credit Cards and the Purchases in a Week

Number of credit cards	Number of Purchases in a Week					
	Y					
X	0	1	2	3	4	$p(x_i)$
1	0.08 $P(x=1 \cap y=0)$	0.13 $P(x=1 \cap y=1)$	0.09 $P(x=1 \cap y=2)$	0.06 $P(x=1 \cap y=3)$	0.03 $P(x=1 \cap y=4)$	0.39 $P(x=1)$
2	0.03 $P(x=2 \cap y=0)$	0.08 $P(x=2 \cap y=1)$	0.08 $P(x=2 \cap y=2)$	0.09 $P(x=2 \cap y=3)$	0.07 $P(x=2 \cap y=4)$	0.35 $P(x=2)$
3	0.01 $P(x=3 \cap y=0)$	0.03 $P(x=3 \cap y=1)$	0.06 $P(x=3 \cap y=2)$	0.08 $P(x=3 \cap y=3)$	0.08 $P(x=3 \cap y=4)$	0.26 $P(x=3)$
$p(y_j)$	0.12 $P(y=0)$	0.24 $P(y=1)$	0.23 $P(y=2)$	0.23 $P(y=3)$	0.18 $P(y=4)$	1.00

In general, X takes on values x_i, $i = 1, 2, \ldots n_x$ and Y takes on values y_j, $j = 1, 2, \ldots, n_y$; in Fig 2.3, $n_x = 3$ and $n_y = 5$. $P(X=x_i \cap Y=y_j) = P(x_i, y_j)$; $0 \leq P(x_i, y_j) \leq 1$ and $\sum_i \sum_j P(x_i, y_j) = 1$. The marginal distribution of $Y = \sum_i P(x_i, y_j) = P(y_j)$ and the marginal distribution of $X = \sum_j P(x_i, y_j) = P(x_i)$. The conditional probability of x_i given y_j is $P(x_i \mid y_j) = P(x_i \cap y_j) / P(y_j)$; similarly, $P(y_j \mid x_i) = P(x_i, y_j) / P(x_i)$. For example, the conditional distribution of X, given $y = 2$, $= P(x_i \cap y=2) / P(y=2)$

$= \frac{0.09}{0.23}, \frac{0.08}{0.23}, \frac{0.06}{0.23} = 0.3913, 0.3478, 0.2609$, the sum of which is one.

+44-

Are $x=2$, $y=2$ independent? Well, $p(x=2 \cap y=2) = 0.08$ and $p(x=2)p(y=2) = 0.35 \times 0.23 = 0.0805$ which is not (quite) independent.

The population means and variances will be calculated from Fig. 2.3:

$E(X) = 1\times0.39 + 2\times0.35 + 3\times0.26 = 0.39 + 0.70 + 0.78$

$= 1.87$ cards $= \mu_x =$ mean of X. $\mu_x^2 = 3.4969$

$E(X^2) = 1\times0.39 + 4\times0.35 + 9\times0.26 = 0.39 + 1.40 + 2.34$

$= 4.13$

$\sigma_x^2 = E(X^2) - \mu_x^2 = 4.13 - 3.4969 = 0.6331 =$ variance of X

$\sigma_x \doteq 0.7967 =$ standard deviation of X

$E(Y) = 0\times0.12 + 1\times0.24 + 2\times0.23 + 3\times0.23 + 4\times0.18$

$= 0.24 + 0.46 + 0.69 + 0.72$

$= 2.11$ purchas per week $= \mu_Y =$ mean of Y. $\mu_Y^2 = 4.4521$

$E(Y^2) = 1\times0.24 + 4\times0.23 + 9\times0.23 + 16\times0.18$

$= 0.24 + 0.92 + 2.07 + 2.88$

$= 6.11$

$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = 6.11 - 4.4521 = 1.6579 =$ variance of Y

$\sigma_Y \doteq 1.2876$

If there is a bivariate distribution in X and Y, there is a mean and a variance of each of X and Y, and in addition there is a

<u>covariance</u> of X and $Y = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_X)] = E(XY) - \mu_X\mu_Y$

Since μ_X and μ_Y are known, $\mu_X\mu_Y = 1.87 \times 2.11 = 3.9457$.

$E(XY) = \sum_i \sum_j x_i y_j p(x_i, y_j)$ which, leaving out the element $= 0$,

$$\begin{aligned}
&= 1\times1\times0.13 + 1\times2\times0.09 + 1\times3\times0.06 + 1\times4\times0.03 \\
&\quad + 2\times1\times0.08 + 2\times2\times0.08 + 2\times3\times0.09 + 2\times4\times0.07 \\
&\quad + 3\times1\times0.03 + 3\times2\times0.06 + 3\times3\times0.08 + 3\times4\times0.08, \\
&= 0.13 + 0.18 + 0.18 + 0.12 \\
&\quad + 0.16 + 0.32 + 0.54 + 0.56 \\
&\quad + 0.09 + 0.36 + 0.72 + 0.96, \\
&= 4.32.
\end{aligned}$$

$\sigma_{XY} = E(XY) - \mu_X\mu_Y = 4.32 - 3.9457 = 0.3743.$

In summary, X and Y are arranged with means, variances and covariances as follows:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix}\right] = \left[\begin{pmatrix} 1.87 \\ 2.11 \end{pmatrix}, \begin{pmatrix} 0.6331 & 0.3743 \\ 0.3743 & 1.6579 \end{pmatrix}\right].$$

Finally, the square of the correlation $\rho_{XY}^2 \equiv \left(\frac{\sigma_{XY}^2}{\sigma_X^2\,\sigma_Y^2}\right)$ and $0 \leq \rho_{XY}^2 \leq 1$.

Thus, $\rho_{XY}^2 = \frac{0.140100049}{1.049561649} \doteq 0.1335$; $\rho = 0.3653$ where $-1 \leq \rho \leq +1$.