UNIVERSITY OF JOHANNESBURG

Statistics 2B (Practical)

Practical compiled by Mr. V van Appel, Department of Statistics - University of Johannesburg
Practical video done by Mr. T Hansragh, Department of Statistics - University of Johannesburg

Practical 1: Bivariate Data

https://youtu.be/xJy5t59-RAo

The relationship between 2 variables is often of interest. For example, are height and weight related? Are age and heart rate related? Are income and taxes paid related? Is a new drug better than an old drug? Does a batter hit better as a switch hitter or not? Does the weather depend on the previous days weather? Exploring and summarizing such relationships is the current goal.

1.1 Handling bivariate categorical data

The table command will summarize bivariate data in a similar manner as it summarized univariate data. Suppose a student survey is done to evaluate if students who smoke study less. The data recorded is:

Person	Smokes	Amount of Studying
1	Y	less than 5 hours
2	N	5 - 10 hours
3	N	5 - 10 hours
4	Y	more than 5 hours
5	N	more than 5 hours
6	Y	less than 5 hours
7	Y	5 - 10 hours
8	Y	less than 5 hours
9	N	more than 5 hours
10	Y	5 - 10 hours

We can handle this in R by creating two vectors to hold our data, and then using the table() command.

```
> smokes = c("Y","N","N","Y","N","Y","Y","Y","N","Y")
> amount = c(1,2,2,3,3,1,2,1,3,2)
> table(smokes,amount)
```

```
      amount
smokes 1 2 3
     N 0 2 2
     Y 3 2 1
```

We see that there may be some relationship.

What would be nice to have are the marginal totals and the proportions. For example, what proportion of smokers study 5 hours or less. We know that this is $3/(3+2+1) = 1/2$, but how can we do this in R?

The command prop.table() will compute this for us. It needs to be told the table to work on, and a number to indicate if you want the row proportions (a 1) or the column proportions (a 2) the default is to just find proportions.

```
> tmp=table(smokes,amount) # store the table
> old.digits = options("digits") # store the number of digits
> options(digits=3) # only print 3 decimal places
> prop.table(tmp,1) # the rows sum to 1 now
```

```
      amount
smokes     1     2     3
     N 0.000 0.500 0.500
     Y 0.500 0.333 0.167
```

```
> prop.table(tmp,2) # the columns sum to 1 now
```

```
      amount
smokes     1     2     3
     N 0.000 0.500 0.667
     Y 1.000 0.500 0.333
```

```
> prop.table(tmp)   # all the numbers sum to 1
```

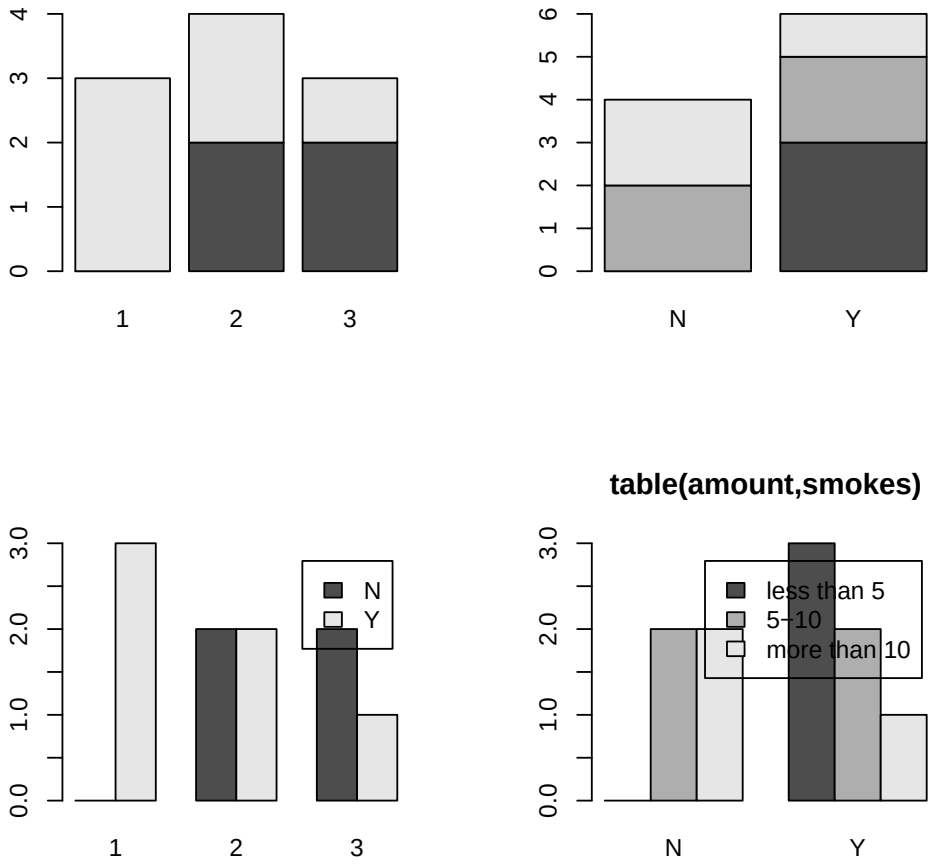
```
      amount
smokes   1   2   3
     N 0.0 0.2 0.2
     Y 0.3 0.2 0.1
```

```
> #options(digits=old.digits) # restore the number of digits
```

1.1.1 Plotting tabular data

You might wish to graphically represent the data summarized in a table. For the smoking example, you could plot the amount variable for each of No or Yes, or the No and Yes variable for each level of smoking. In either case, you can use a barplot. We simply call it in the appropriate manner.

```
> par(mfrow=c(2,2))
> barplot(table(smokes,amount))
> barplot(table(amount,smokes))
> smokes=factor(smokes) # for names
> barplot(table(smokes,amount),
+ beside=TRUE, # put beside not stacked
+ legend.text=T) # add legend
> barplot(table(amount,smokes),main="table(amount,smokes)",beside=TRUE,
+ legend.text=c("less than 5","5-10","more than 10"))
```



Notice in figure 10 the importance of order when making the table. Essentially, barplot plots each row of data. It can do it in a stacked manner (the default), or besides (by setting beside=TRUE). The attribute legend.text adds the legend to the graph. You can change the names, but the default of legend.text=T is easiest if you have a factor labelling the rows of the table command.

1.2 Multivariate Distributions

We have built up quite a catalogue of distributions, discrete and continuous. They were all univariate, however, meaning that we only considered one random variable at a time. We can

imagine nevertheless many random variables associated with a single person: their height, their weight, their wrist circumference (all continuous), or their eye/hair color, shoe size, whether they are right handed, left handed, or ambidextrous (all categorical), and we can even surmise reasonable probability distributions to associate with each of these variables. But there is a difference: for a single person, these variables are related. For instance, a person's height betrays a lot of information about that person's weight. The concept we are hinting at is the notion of dependence between random variables. It is the focus of this chapter to study this concept in some detail. Along the way, we will pick up additional models to add to our catalogue. Moreover, we will study certain classes of dependence, and clarify the special case when there is no dependence, namely, independence.

What do you need you to know?

- the basic notion of dependence and how it is manifested with multiple variables (two, in particular)
- joint versus marginal distributions/expectation (discrete and continuous)
- some numeric measures of dependence
- conditional distributions, in the context of independence and exchangeability
- some details of at least one multivariate model (discrete and continuous)
- what it looks like when there are more than two random variables present

Example 1.1 *Roll a fair die twice. Let X be the face shown on the first roll, and let Y be the face shown on the second roll. We have already seen this example many times. For this example, it suffices to define*

$$f_{XY}(x, y) = \frac{1}{36}, \qquad x = 1, \ldots, 6; y = 1, \ldots, 6 \tag{1.1}$$

The marginal pmf's are given by $f_X(x) = \frac{1}{6}$, $x = 1, 2, \ldots, 6$, and $f_Y(y) = \frac{1}{6}$, $y = 1, 2, \ldots, 6$, since

$$f_X(x) = \sum_{y=1}^{6} \frac{1}{36}, \quad x = 1, \ldots, 6, \tag{1.2}$$

and the same computation with the letters switched works for Y. In the previous example, and in many other ones, the joint support can be written as a product set of the support of X "times" the support of Y, that is, it may be represented as a cartesian product set, $f_{XY} = f_X \times f_Y$. As we shall see presently in Section 3.4 (Rice), this form is a necessary condition for X and Y to be independent. We next investigate just such an example.

Example 1.2 *Let the random experiment again be to roll a fair die twice, except now let us define the random variables U and V by U = the maximum of the two rolls, and V = the sum of the two rolls. We see that the sample space of U is $S_U = \{1, 2, \ldots, 6\}$ and the sample space of V is $S_V = \{2, 3, \ldots, 12\}$. We may represent the sample space with a matrix, and for each entry in the matrix we may calculate the value that U assumes. The result is in the left half of Table 1.1. We can use the table to calculate the marginal pmf of U, because from Example 1.1 we know that each entry in the matrix has probability 1/36 associated with it. For instance, there is only one outcome in the matrix with $U = 1$, namely, the top left corner. This single entry has probability*

Table 1.1: Maximum U and sum V of a pair of dice rolls (X, Y)

U	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

(a) $U = max(X, Y)$

V	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

(b) $V = X + Y$

1/36, therefore, it must be that $f_U(1) = P(U = 1) = 1/36$. *Similarly we see that there are three entries in the matrix with* $U = 2$, *thus* $f_U(2) = 3/36$. *Continuing in this fashion we will find the marginal distribution of* U *may be written:*

$$f_U(u) = \frac{2u - 1}{36}, \qquad u = 1, 2, \ldots, 6. \tag{1.3}$$

We may do a similar thing for V*; see the right half of Table 1.1. Collecting all of the probability we will find that the marginal pmf of* V *is*

$$f_V(v) = \frac{6 - |v - 7|}{36}, \qquad v = 2, 3, \ldots, 12. \tag{1.4}$$

We may collapse the two matrices from Table 1.1 into one, big matrix of pairs of values (u, v). *The result is shown in Table 1.2.*

Table 1.2: Joint values of $U = max(X, Y)$ and $V = X + Y$

(U, V)	1	2	3	4	5	6
1	(1,2)	(2,3)	(3,4)	(4,5)	(5,6)	(6,7)
2	(2,3)	(2,4)	(3,5)	(4,6)	(5,7)	(6,8)
3	(3,4)	(3,5)	(3,6)	(4,7)	(5,8)	(6,9)
4	(4,5)	(4,6)	(4,7)	(4,8)	(5,9)	(6,10)
5	(5,6)	(5,7)	(5,8)	(5,9)	(5,10)	(6,11)
6	(6,7)	(6,8)	(6,9)	(6,10)	(6,11)	(6,12)

$(1, 2)$ *appears twice, but* $(2, 3)$ *appears only once. We can make more sense out of this by writing a new table with* U *on one side and* V *along the top. We will accumulate the probability just like we did in Example 1.1. See Table 1.3.*

The outcomes of U *are along the left and the outcomes of* V *are along the top. Empty entries in the table have zero probability. The row totals (on the right) and column totals (on the bottom) correspond to the marginal distribution of* U *and* V*, respectively.*

The joint support of (U, V) *is concentrated along the main diagonal; note that the nonzero entries do not form a rectangle. Also notice that if we form row and column totals we are doing exactly the same thing as the marginal distribution, so that the marginal distribution of* U *is the list of totals in the right "margin" of the Table 1.3, and the marginal distribution of* V *is the list of totals in the bottom "margin".*

Table 1.3: The joint pmf of (U, V). The outcomes of U are along the left and the outcomes of V are along the top. Empty entries in the table have zero probability. The row totals (on the right) and column totals (on the bottom) correspond to the marginal distribution of U and V, respectively.

(U,V)	2	3	4	5	6	7	8	9	10	11	12	Total
1	1/36											1/36
2		2/36	1/36									3/36
3			2/36	2/36	1/36							5/36
4				2/36	2/36	2/36	1/36					7/36
5					2/36	2/36	2/36	2/36	1/36			9/36
6						2/36	2/36	2/36	2/36	2/36	1/36	11/36
Total	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

1.2.1 How to do it with R

We will show how to do Example 1.2 using R; it is much simpler to do it with R than without. First we set up the sample space with the expand.grid() function (under the library(PASWR)). Next, we add random variables U and V with the apply() function. We take a look at the data frame (probability space) to make sure that everything is operating according to plan.

Try this for Homework. The Solution will be discussed in the You-tube video!

You should check that the answers that we have obtained exactly match the same (somewhat laborious) calculations that we completed in Example 1.2.

1.3 Covariance and Correlation

There are two very special cases of joint expectation: the covariance and the correlation. These are measures which help us quantify the dependence between X and Y.

Definition 1.1 *The covariance of X and Y is*

$$Cov(X, Y) = E[(x - E(x))(Y - E(Y))] \tag{1.5}$$

By the way, there is a shortcut formula for covariance which is almost as handy as the shortcut for the variance:

$$Cov(X, Y) = E(XY) - E(X)E(Y) \tag{1.6}$$

The proof is left as an Exercise.

The Pearson product moment correlation between X and Y is the covariance between X and Y rescaled to fall in the interval $[-1, 1]$. It is formally defined by

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{1.7}$$

The correlation is usually denoted by $\rho_{X,Y}$ or simply ρ if the random variables are clear from context. There are some important facts about the correlation coefficient:

- The range of correlation is $-1 \leq \rho_{X,Y} \leq 1$
- Equality holds above ($\rho_{X,Y} = \pm 1$) if and only if Y is a linear function of X with probability one.

Example 1.3 *We will compute the covariance for the discrete distribution in Example 1.2. The expected value of U is*

$$E(U) = \sum_{u=1}^{6} u f_U(u) = \sum_{u=1}^{6} u \frac{2u-1}{36} = \frac{161}{36} \tag{1.8}$$

and the expected value of V is

$$E(V) = \sum_{v=2}^{12} v f_V(v) = \sum_{v=2}^{12} v \frac{6 - |7-v|}{36} = 7 \tag{1.9}$$

and the expected value of UV is

$$E(UV) = \sum_{u=1}^{6} \sum_{v=2}^{12} uv f_{U,V}(u,v) = \frac{308}{9} \tag{1.10}$$

Therefore the covariance of (U, V) is

$$Cov(U,V) = E(XY) - E(X)E(Y) = \frac{308}{9} - \frac{161}{36} \times 7 = \frac{35}{12}. \tag{1.11}$$

All we need now are the standard deviations of U and V to calculate the correlation coefficient (omitted).

1.3.1 How to do it with R

There are not any specific functions designed for multivariate expectation. This is not a problem, though, because it is easy enough to do expectation the long way - with column operations. We just need to keep the definition in mind. For instance, we may compute the covariance of (U, V) from Example 1.3.

Try this for Homework. The Solution will be discussed in the You-tube video!

Compare this answer to what we got in Example 1.3.

All of this can be done in a few lines as follows: (see You-tube video)

1.4 Independent Random Variables

We recall from Chapter 3 (Rice) that the events A and B are said to be independent when

$$P(A \cap B) = P(A) \times P(B) \tag{1.12}$$

If it happens that

$$P(X = x; Y = y) = P(X = x) \times P(Y = y), \quad \text{for every} \quad x \in S_X,\, y \in S_Y \tag{1.13}$$

then we say that X and Y are independent random variables. Otherwise, we say that X and Y are dependent. Using the pmf notation from above, we see that independent discrete random variables satisfy

$$f_{X,Y}(x,y) = f_X(x) \times f_Y(y) \quad \text{for every} \quad x \in S_X,\, y \in S_Y \tag{1.14}$$

then we say that X and Y are independent.

Example 1.4
In Example 1.1 we considered the random experiment of rolling a fair die twice. There we found the joint pmf to be

$$f_{X,Y}(x,y) = \frac{1}{36} \quad \text{for every} \quad x = 1,\ldots,6,\, y = 1,\ldots,6 \tag{1.15}$$

and we found the marginal pmf's $f_X(x) = 1/6,\ x = 1,2,\ldots,6$, and $f_Y(y) = 1/6,\ y = 1,\ldots,6$. Therefore in this experiment X and Y are independent since for every x and y in the joint support the joint pmf satisfies

$$f_{X,Y}(x,y) = \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = f_X(x) \times f_Y(y) \tag{1.16}$$

Example 1.5
In Example 1.2 we considered the same experiment but different random variables U and V. We can prove that U and V are not independent if we can find a single pair (u,v) where the independence equality does not hold. There are many such pairs. One of them is (6, 12):

$$f_{U,V}(6,12) = \frac{1}{36} \neq \frac{11}{36} \times \frac{1}{36} = f_U(6) \times f_V(12) \tag{1.17}$$

Independent random variables are very useful to the mathematician. They have many, many, tractable properties. We mention some of the more important ones.

Proposition 1.1 *If X and Y are independent, then for any functions u and v,*

$$E[u(X)v(Y)] = E[u(X)]E[v(Y)]\,: \tag{1.18}$$