

CARLETON COLLEGE

A PROBABILITY PRIMER

SCOTT BIERMAN

(Do not quote without permission)

A Probability Primer

INTRODUCTION

The field of probability and statistics provides an organizing framework for systematically thinking about randomness. Since many decisions regarding how scarce resources are allocated are made when there is uncertainty about the consequences of those decisions, having the help of an organized way of thinking about this uncertainty can be extremely valuable. An introduction to this area is the point of the handout.

FUNDAMENTAL DEFINITIONS

As with all disciplines, probability and statistics has its own language. We begin with a few definitions that are indispensable.

*Definition: **Sample space:** a list of all possible outcomes.*

Example: You buy a stock today. The sample space of stock prices tomorrow consists of all possible stock prices tomorrow (this would be approximated by all non-negative real numbers).

*Definition: **Event:** A set of possible outcomes.*

*Example: You buy a stock today. The **event** that the price of the stock goes up by at least 10% overnight consists of all prices that are 10% higher than the price you paid for the stock.*

Suppose that we let $N(A)$ be the number of times we observe the event A occurring in N trials, while N represents the number of trials. Then the relative frequency of the event A occurring is defined as $\frac{N(A)}{N}$.

*Definition: **Relative Frequency.** The relative frequency of event A is the proportion of times that event A occurs in N trials.*

This immediately brings us to the definition of the probability of an event occurring.

*Definition: **Probability.** The **probability** of event A occurring is the **relative frequency** of event A occurring as the number of trials approaches infinity.*

We will write the probability of event A occurring as: $P(A)$. From this definition it follows that probabilities of events must be between 0 and 1 (inclusive).

SOME USEFUL EVENTS

Some events are **mutually exclusive**. Suppose we are considering two events. These two events are mutually exclusive if it is impossible for the same trial to result in *both* events.

*Definition: **Mutually Exclusive:** Events A and B are mutually exclusive if and only if $P(A \text{ or } B) = P(A) + P(B)$*

This should be read as: The events A and B are mutually exclusive if the probability of event A or event B occurring equals the probability of A occurring plus the probability of B occurring.

Example of two *mutually exclusive events*: One die is rolled. Event A is rolling a 1 or 2. Event B is rolling a 4 or 5. The probability of event A is $1/3$, the probability of event B is $1/3$, but the probability of A or B is $2/3$.

Example of two *non-mutually exclusive events*: One die is rolled. Event A is rolling a 1 or 2. Event B is rolling a 2 or 3. The probability of event A is $1/3$, the probability of event B is $1/3$, but the probability of A or B is not $2/3$. A or B will only occur when a 1, 2, or 3 is rolled. Then means that the probability of event A or B occurring is $1/2$, not $2/3$.

Some events, when taken together, must occur. Collections of events, at least one of which must occur, are called collectively exhaustive.

*Definition: **Collectively Exhaustive:** The events A or B or C are collectively exhaustive if $P(A \text{ or } B \text{ or } C) = 1$*

Example of collectively exhaustive events: One die is rolled. Event A is rolling a number less than 5. Event B is rolling a number greater than 3. Any roll of a die will satisfy either A or B (in fact, a roll of 4 satisfies both).

Some collections of events are mutually exclusive *and* collectively exhaustive.

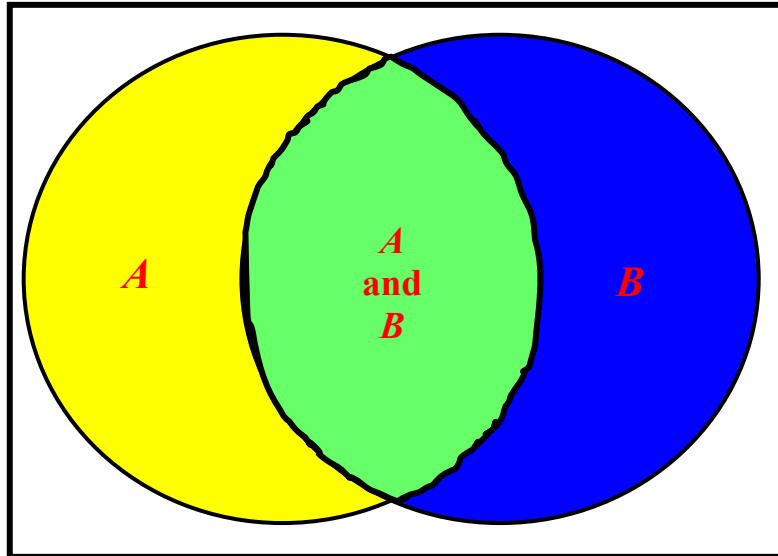
Example of events that are mutually exclusive and collectively exhaustive: You buy a stock today. The event A is that the price of the stock goes up tomorrow. The event B is the price of the stock goes down tomorrow. Event C is the price of the stock does not change.

The terms “mutually exclusive” and “collectively exhaustive” are used so frequently that you can expect to hear them in everyday conversation.

THE BASIC ADDITION PROPERTY OF PROBABILITY

Suppose there are two events; A and B . In many instances we will be interested in calculating what is the probability of event A or event B occurring. We have already seen how to do this for mutually exclusive events, but we have also seen that simply adding the probabilities of the two events does not work for non-mutually exclusive events.

Think of points contained within the rectangle below as the sample space for some random variable.



The points contained within the yellow and green areas we will call event A . The points contained within the blue and green areas we will call event B .

One point is randomly selected from the sample space. If the selection of any point from the sample space is equally likely, then the magnitude of the yellow and green areas relative to the total area as the probability of event A occurring, and the magnitude of the blue and green areas relative to the total area as the probability of event B occurring. The probability of either A or B is the area of all colored regions relative to the total area. But, notice that $P(A \text{ or } B)$ does not equal $P(A)$ plus $P(B)$ because this would count the green area twice. This means,

$$\blacksquare \quad P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

This can also be written

- $P(A \text{ and } B) = P(A) + P(B) - P(A \text{ or } B)$

So, if two events are mutually exclusive, then

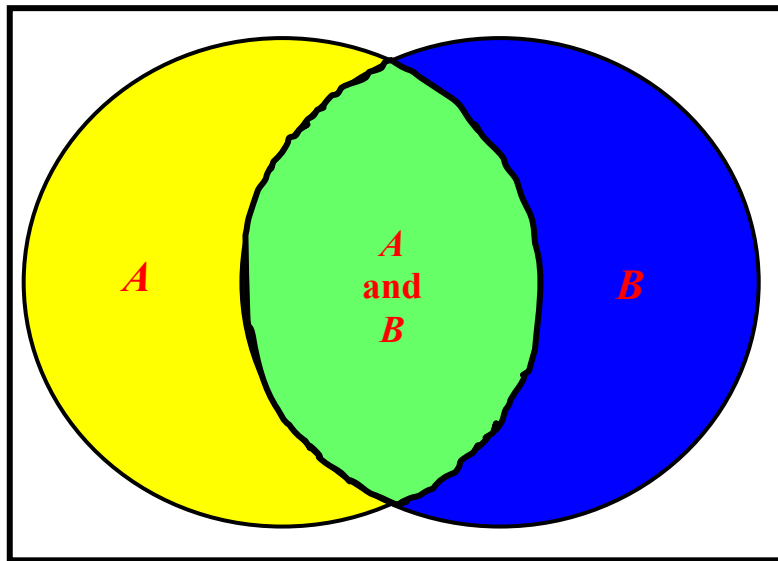
- $P(A \text{ and } B) = 0$

This means there is no intersection between these events.

CONDITIONAL PROBABILITY

In many cases you want to know the probability of some event given the occurrence of some other event. The probability of tomorrow being rainy (without imposing any conditions) is likely different than the probability that tomorrow is rainy given that it is May.

Using the Venn diagram again, consider the probability of B occurring.



Without any conditions, the probability that event B occurs will equal the size of the blue and green areas relative to the area of the entire rectangle. However, if we ask: What is the probability of event B , conditional on event A occurring? We get a different answer. Now, the relevant sample space consists only of points within the yellow and green areas. The

probability of event B occurring given that event A has occurred, equals the green area relative to the yellow area.

This principle can be generalized. The probability of event B occurring conditional on event A equals that probability of events A and B occurring divided by the probability of A occurring.

$$\blacksquare \quad P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$

Example: Suppose a die is rolled. Event A is the roll takes a value less than 4. Event B is that the roll is an odd number. What is the probability of the roll being an odd number given that event A has occurred? We know that Event A will occur when a 1, 2, or 3 is rolled. Event B will occur when a 1, 3, or 5 is rolled. So of the three values that encompass event A , two of them are associated with event B . So the probability of event B occurring given event A is two-thirds. Now use the formula. Events A and B occur when a 1 is thrown or when a 3 is thrown. The probability of one of these happening is $1/3$. Event A occurs when a 1, 2, or 3 is rolled. The probability of event A is $1/2$. So

$$P(A \text{ and } B) = \frac{1}{3}$$

$$P(A) = \frac{1}{2}$$

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{2}{3}$$

You will find that the formula for conditional probability is very useful. It can also be rewritten:

$$\blacksquare \quad P(A \text{ and } B) = P(A) \cdot P(B | A)$$

INDEPENDENT EVENTS

In a casual sense, two events are independent if knowledge that one event has occurred does not cause you to adjust the probability of the other event occurring. More formally,

*Definition: Two events, A and B , are **independent**, if*

$$P(A | B) = P(A)$$

We have seen that by definition

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

This means that if two events are independent:

$$\frac{P(A \text{ and } B)}{P(B)} = P(A)$$

So,

$$\blacksquare \quad P(A \text{ and } B) = P(A) \cdot P(B)$$

Example: Suppose I roll a die and you roll a die. Event A is my roll is a 3. Event B is your roll is a 3. The probability of my roll being a 3 if your roll is a 3 is $1/6$. But this is exactly the same as the probability of my roll being a 3 and having no information about your roll.

This means,

$$P(A | B) = \frac{1}{6}$$

So, the probability of both our rolls being a 3 can be calculated.

$$P(A \text{ and } B) = P(A) \cdot P(B) = \frac{1}{36}$$

BAYES' THEOREM (THE SIMPLE VERSION)

Consider two events A and B . These two events give rise to two other events; *not* A and *not* B . These will be denoted by: \tilde{A} and \tilde{B} .

Notice that the events B and \tilde{B} are mutually exclusive and collectively exhaustive.

So, of course, are A and \tilde{A} .

The question Bayes asked was: How does the probability of B change if we know whether or not A has occurred? Or: How does the observation of whether or not A has occurred cause you to update your probabilistic assessment of the likelihood of B occurring? In short, we want a helpful expression for $P(B | A)$.

We already have the definition of $P(B | A)$,

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$

And, this means,

$$P(A \text{ and } B) = P(A) \cdot P(B | A).$$

But it must also be true that

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

and

$$P(A \text{ and } B) = P(B) \cdot P(A | B).$$

So,

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

Since B and \tilde{B} are mutually exclusive and collectively exhaustive:

$$P(A) = P(A \text{ and } B) + P(A \text{ and } \tilde{B})$$

So,

$$P(A) = P(B) \cdot P(A | B) + P(\tilde{B}) \cdot P(A | \tilde{B})$$

$$\blacksquare \quad P(B | A) = \frac{P(B) \cdot P(A | B)}{P(B) \cdot P(A | B) + P(\tilde{B}) \cdot P(A | \tilde{B})}$$

This is Bayes' theorem. It is nothing more than the definition of conditional probability applied a couple of times and a little bit of algebraic cleverness. The importance of Bayes' theorem is that the informational requirements to calculate $P(B | A)$ from Bayes' theorem are different than those required by the definition of $P(B | A)$.

AN EXAMPLE OF HOW BAYES' THEOREM CAN BE USED

All people at a firm are tested for a medical condition (HIV, for example). Suppose you have the following information about this medical condition and a laboratory test for this condition:

- The chance of a random draw from the population having the medical condition is $\frac{1}{1000}$.
- The chance of a false positive test result from the lab test is $\frac{1}{100}$.
- The chance of a false negative test result for the lab test is $\frac{1}{500}$.

Without the test you rationally believe your chances of having the medical condition are $\frac{1}{1000}$. An important question to someone just tested is; if the test comes back positive, what are the chances that you have the medical condition?

To answer this question, we will formalize the information provided above. Define the following events:

A : You have the medical condition.

\tilde{A} : You do not have the medical condition.

B : You have a positive test result.

\tilde{B} : You have a negative test result.

Since 1 out of every 1000 people have the medical condition: $P(A) = \frac{1}{1000}$. Since there are false positive results in 1 out of every 100 lab tests: $P(B | \tilde{A}) = \frac{1}{100}$. And, since there are false negative results in 1 out of every 500 lab tests: $P(\tilde{B} | A) = \frac{1}{500}$.

But we can go further. The data provided also allow us to calculate. Since 999 out of every 1000 people do not have the medical condition: $P(\tilde{A}) = \frac{999}{1000}$. Since there are 99 correctly positive lab tests out of every 100 positive lab tests: $P(\tilde{B} | \tilde{A}) = \frac{99}{100}$. And, since there are 499 correctly negative lab tests out of every 500 negative lab tests:

$$P(B | A) = \frac{499}{500}.$$

Remember that the person having just received the results of her lab test is interested in calculating the probability of having the medical condition having just heard that she has a positive lab test result. In terms of our notation above, she wants to calculate

$$P(A | B).$$

Bayes' theorem tells us,

$$P(A | B) = \frac{P(A) \cdot P(B | A)}{P(A) \cdot P(B | A) + P(\tilde{A}) \cdot P(B | \tilde{A})}$$

$$\blacksquare \quad P(A | B) = \frac{\frac{1}{1000} \cdot \frac{499}{500}}{\frac{1}{1000} \cdot \frac{499}{500} + \frac{999}{1000} \cdot \frac{1}{100}} = 0.0908 = 9.08\%$$

Having just tested positive for the medical condition there is (only) a 9.08% chance that she actually has the condition.

Is there some intuition behind this result? Suppose there are 1,000,000 people randomly chosen from the population, all of whom are tested for the medical condition. We would expect 1000 of them to have the medical condition. Of the 1000 who have the condition, 2 will not show a positive test result. This is what a false negative lab test means. But, we also know that 998 of the people who have the medical condition will correctly get a positive test result. Of the 999,000 who do not have the condition, 9,990 will receive a positive test result. This is what a false positive means. So, a total of $998+9,980=10,978$ receive a positive test result. But of this group only 998 are actually positive. So, the probability of having the condition if you test positive is $998/10098$, or 0.0988. Pretty close.

PROBABILITY DISTRIBUTIONS

Fundamental to probability distributions are random variables.

*Definition: A **random variable** is a rule that assigns a number to each possible outcome in a chance experiment.*

We will start with discrete random variables (as opposed to continuous random variables).

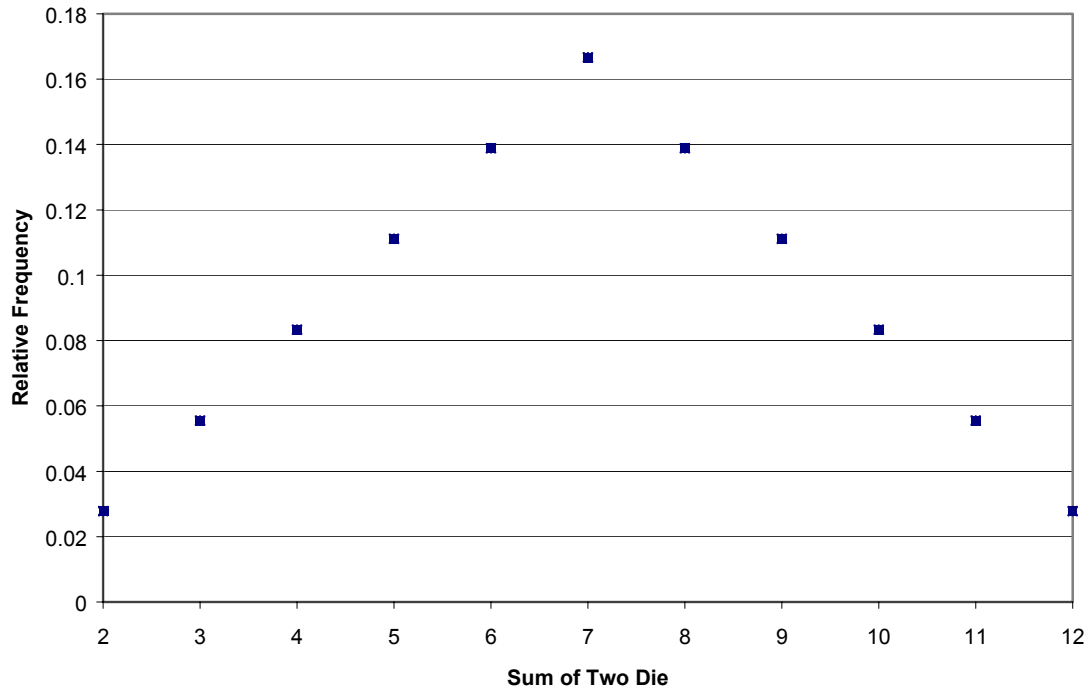
Example: The sum of the two dice.

A probability distribution simply maps the probability of the random variable taking on every possible value to each of those values.

In the example above, the probability distribution is:

RV	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

This can also be plotted. Probability distributions associated with discrete random variables must sum (over all possible values of the random variable) to one.



Now consider a random variable that is continuous. For example, the hourly flow of oil from a well is a random variable that is continuous within some boundaries. Since the probability of any specific number being chosen is infinitely small, it is more useful to think about the probability of a random number falling within a range of values.

For example, it is easy to ask a spreadsheet program such as Excel to randomly select a number between 0 and 50. The computer will then be asked to select any real number between 0 and 50 all of which are equally likely to be drawn. The chances you will get any

particular value exactly is zero, but the chances you will get a value between 10 and 20 is 20% (1 in 5).

As with discrete random variables, it is also possible but more complicated, to put a continuous probability distribution into a diagram. Recall that with a discrete probability distribution:

$$\sum_{i=1}^n P(x_i) = 1$$

where x_i is the i th value that the random variable can take on, $P(\cdot)$ is the probability the random variable takes on a particular value, and n is the number of possible random values. For a continuous probability distribution,

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

where, $\int_a^b f(x)$ is the probability of the random variable, x , falling between the values of a and b , for all a and b .

CUMULATIVE DISTRIBUTION (DENSITY) FUNCTIONS

Related to probability distribution functions are cumulative probability distribution functions. In short they relate the value of a random variable with the probability of the random variable being less than or equal to that value.

For a discrete random variable, the cumulative distribution function is

$$F(x_j) = \sum_{i=1}^j P(x_i) \text{ for all } i \leq j.$$

For a continuous random variable, it is

$$F(a) = \int_{-\infty}^a f(x) dx$$

THE EXPECTED VALUE OF A RANDOM VARIABLE

It is often helpful to have a measure of the central tendency of a random variable. There are several of these that people use regularly; the mean value of a random variable, the median value of a random variable, and the mode of a random variable are the three most important. The one that we will pay the most attention to here is the mean value of a random variable. It is sometimes called the expected value of the random variable.

*Definition: The **Expected Value** of a (discrete) random variable is the sum of all the possible values that random variable can take on each weighted by its probability of occurring.*

$$E(x) = \sum_{i=1}^n x_i P(x_i)$$

*Definition: The **Expected Value** of a (continuous) random variable is*

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

$E(\cdot)$ is often called an “expectations operator” and has a variety of properties that are worth knowing something about. In general, as we functionally transform a random variable, call it $g(x)$, we define,

$$E(g(x)) = \sum_{i=1}^n g(x) \cdot f(x)$$

Therefore,

- $E(ax) = aE(x)$

Proof:

$$g(x) = ax$$

$$E(ax) = \sum_{i=1}^n axP(x)$$

$$E(ax) = a \left(\sum_{i=1}^n xP(x) \right)$$

$$E(ax) = aE(x)$$

- $E(x+a) = E(x) + a$

Proof:

$$g(x) = x + a$$

$$E(x+a) = \sum_{i=1}^n (x+a)P(x)$$

$$E(x+a) = \sum_{i=1}^n xP(x) + aP(x)$$

$$E(x+a) = \sum_{i=1}^n xP(x) + a \sum_{i=1}^n P(x)$$

$$E(x+a) = E(x) + a$$

- $E(ax^2 + bx) = aE(x^2) + bE(x)$

$$E(ax^2 + bx) = \sum_{i=1}^n (ax^2 + bx)P(x)$$

$$E(ax^2 + bx) = \sum_{i=1}^n (ax^2)P(x) + \sum_{i=1}^n (bx)P(x)$$

$$E(ax^2 + bx) = a \sum_{i=1}^n (x^2)P(x) + b \sum_{i=1}^n (x)P(x)$$

$$E(ax^2 + bx) = aE(x^2) + bE(x)$$

- $E((ax+b)^2) = a^2E(x^2) + 2abE(x) + b^2$

Proof:

$$g(x) = (ax+b)^2$$

$$E((ax+b)^2) = E(a^2x^2 + 2abx + b^2)$$

$$E((ax+b)^2) = E(a^2x^2) + E(2abx) + b^2$$

$$E((ax+b)^2) = a^2E(x^2) + 2abE(x) + b^2$$

THE VARIANCE AND STANDARD DEVIATION OF A RANDOM VARIABLE

Another useful way of describing a random variable is to find a measure for its dispersion around the mean value. Commonly the variance or the standard deviation of a random variable is used towards this end.

*Definition: The **variance** of a (discrete) random variable is*

$$E((x - E(x))^2)$$

or,

$$\sum_{i=1}^n (x_i - E(x_i))^2 P(x_i).$$

*Definition: The **standard deviation** of a (discrete) random variable is*

$$\sqrt{E((x - E(x))^2)}$$

or,

$$\sqrt{\sum_{i=1}^n (x_i - E(x_i))^2 P(x_i)}.$$

There is an alternative way of writing variance that is worth deriving and remembering.

$$V(x) = E((x - E(x))^2)$$

$$V(x) = E(x^2 - 2xE(x) + (E(x))^2)$$

$$V(x) = E(x^2) - E(2xE(x)) + E(E(x)^2)$$

$$V(x) = E(x^2) - 2E(xE(x)) + E(E(x)^2)$$

$$V(x) = E(x^2) - 2E(x)E(x) + E(x)^2$$

$$V(x) = E(x^2) - 2E(x)^2 + E(x)^2$$

$$V(x) = E(x^2) - E(x)^2$$

JOINT PROBABILITY DISTRIBUTIONS

There are many instances where we are interested in the relationship between two (or more) random variables. For example, if I own shares of IBM stock and shares of Apple stock, I will certainly be interested in the movement of both stock prices in the future. The degree to which they are likely to go up or down together will be important to me.

A joint probability distribution of two random variables identifies the probability of any pair of outcomes occurring together. The notation: $P(x_1 = 3, x_2 = 3)$ should be read the probability that the random variable x_1 takes on a value of three and that the random variable x_2 takes on a value of three.

Suppose we have a joint probability distribution as shown by the following table. The random variable x_1 takes on values of 1, 2, or 3. The random variable x_2 takes on values of 1, 2, 3, or 4. The number in the cells of the table refers to the joint probability that x_1 equals the value in the associated row and that x_2 equals the value in the associated column.

		x_2			
		1	2	3	4
x_1	1	0.25	0.10	0.05	0.05
	2	0.10	0.05	0.00	0.10
	3	0.20	0.00	0.10	0.00

If this is a well-defined joint probability distribution, the numbers in the cells are required to sum to 1. Or,

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} P(x_i, x_j) = 1$$

MARGINAL DISTRIBUTIONS

Joint probability distributions give rise to a plethora of baby distributions. A class of these is known as marginal distributions. Marginal distributions tell you the probability that one random variable takes on any of its values regardless of the value of the other random variable. For example, the probability that x_2 takes on a value of 1 is equal to the sum of $P(x_1 = 1, x_2 = 1)$, $P(x_1 = 2, x_2 = 1)$, and $P(x_1 = 3, x_2 = 1)$. This equals 0.55. Similarly the probability that x_2 takes on a value of 2 is equal to the sum of $P(x_1 = 1, x_2 = 2)$, $P(x_1 = 2, x_2 = 2)$, and $P(x_1 = 3, x_2 = 2)$. This equals 0.15. Again, the probability that x_2 takes on a value of 3 is equal to the sum of $P(x_1 = 1, x_2 = 3)$, $P(x_1 = 2, x_2 = 3)$, and $P(x_1 = 3, x_2 = 3)$. This equals 0.15. You can confirm the probability that x_2 takes on a value of 4 is also 0.15. In terms of the joint probability table, to find the marginal distribution of the random variable x_2 , we simply sum all the rows for every column. The marginal distribution of x_2 is highlighted below.

		x_2			
		1	2	3	4
x_1	1	0.25	0.10	0.05	0.05
	2	0.10	0.05	0.00	0.10
	3	0.20	0.00	0.10	0.00
Marginal probability distribution of x_2		0.55	0.15	0.15	0.15

The same principle applies to calculate the marginal distribution of x_1 . This is shown below.

		x_2			
		1	2	3	4
x_1	1	0.25	0.10	0.05	0.05
	2	0.10	0.05	0.00	0.10
	3	0.20	0.00	0.10	0.00
		Marginal probability distribution of \mathcal{X}_1			

CONDITIONAL DISTRIBUTIONS

Marginal distributions come in handy in the calculation of “conditional distributions.” Suppose we want to know the probability that x_1 takes on a specific value of 3, conditional on x_2 being equal to 1. We already know from the joint probability distribution function that $P(x_1 = 3, x_2 = 1) = 0.20$, and we know from the marginal distribution function for x_2 that $P(x_2 = 1) = 0.55$. Finally, remember that the definition of a conditional probability is: $P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$. Therefore,

$$P(x_1 = 3 | x_2 = 1) = \frac{P(x_1 = 3, x_2 = 1)}{P(x_2 = 1)} = \frac{0.20}{0.55}.$$

For every value of x_2 , there is a conditional probability distribution function for x_1 . All four of these are shown in the table below.

		x_2			
		This column is the conditional distribution of \mathbf{x}_1 given that \mathbf{x}_2 equals 1	This column is the conditional distribution of \mathbf{x}_1 given that \mathbf{x}_2 equals 2	This column is the conditional distribution of \mathbf{x}_1 given that \mathbf{x}_2 equals 3	This column is the conditional distribution of \mathbf{x}_1 given that \mathbf{x}_2 equals 4
x_1	1	0.25/0.55	0.10/0.15	0.05/0.15	0.05/0.15
	2	0.10/0.55	0.05/0.15	0.00/0.15	0.10/0.15
	3	0.20/0.55	0.00/0.15	0.10/0.15	0.00/0.15

Of course, there are similar conditional distributions for x_2 associated with the three different values for x_1 . These are shown below.

		x_2			
		1	2	3	4
x_1	This row is the conditional distribution of \mathbf{x}_2 given that \mathbf{x}_1 equals 1	0.25/0.45	0.10/0.45	0.05/0.45	0.05/0.45
	This row is the conditional distribution of \mathbf{x}_2 given that \mathbf{x}_1 equals 2	0.10/0.25	0.05/0.25	0.00/0.25	0.10/0.25
	This row is the conditional distribution of \mathbf{x}_2 given that \mathbf{x}_1 equals 3	0.20/0.30	0.00/0.30	0.10/0.30	0.00/0.30

INDEPENDENCE AGAIN

If for any pair of random variables the values in the conditional probability cells do not differ from the unconditional probability, then the random variables are independent.

An example of independence: We assign a value of 1 when a flipped coin comes up heads and a value of 0 when it comes up tails. The random variable x_1 is the sum of the flip of two coins. The random variable x_2 is the sum of the flip of two different coin tosses. The joint probability function for x_1 and x_2 is shown below as are the marginal probability functions. You should be able to verify this.

		x_2			Marginal probability distribution of x_1
		0	1	2	
x_1	0	0.0625	0.125	0.0625	0.25
	1	0.125	0.25	0.125	0.50
	2	0.0625	0.125	0.0625	0.25
Marginal probability distribution of x_2		0.25	0.50	0.25	

From this we can calculate the conditional probabilities for x_1 given x_2 .

		x_2			Marginal probability distribution of x_1
		0	1	2	
x_1	0	0.25	0.25	0.25	0.25
	1	0.5	0.5	0.5	0.50
	2	0.25	0.25	0.25	0.25
Marginal probability distribution of x_2		0.25	0.50	0.25	

Notice that as you read across any row the values are exactly the same (including the value of the marginal probability distribution of x_1). This means the conditional distribution of x_1 given x_2 is equal to the marginal probability of x_1 .

$$P(x_1 | x_2) = P(x_1)$$

This is the definition of independence. Hence, x_1 and x_2 are independent random variables.

We could also check this in the other direction. Find, the conditional probabilities for x_2 given x_1 .

		x_2			Marginal probability distribution of x_1
		0	1	2	
x_1	0	0.25	0.50	0.25	0.25
	1	0.25	0.50	0.25	0.50
	2	0.25	0.50	0.25	0.25
Marginal probability distribution of x_2		0.25	0.50	0.25	

As you read down any column the values are exactly the same (including the value of the marginal probability distribution of x_2). This means the conditional distribution of x_2 given x_1 is equal to the marginal probability of x_2 .

$$P(x_2 | x_1) = P(x_2)$$

CONDITIONAL EXPECTED VALUE

One of the most important tools in empirical economics is a conditional expected value. All regression analysis is based on this concept. Let's return to the joint probability distribution function we saw earlier. This is repeated in the table below.

		x_2			
		1	2	3	4
x_1	1	0.25	0.10	0.05	0.05
	2	0.10	0.05	0.00	0.10
	3	0.20	0.00	0.10	0.00

Remember that the conditional distribution of x_1 given x_2 is the following:

		x_2			
		This column is the conditional distribution of x_1 given that x_2 equals 1	This column is the conditional distribution of x_1 given that x_2 equals 2	This column is the conditional distribution of x_1 given that x_2 equals 3	This column is the conditional distribution of x_1 given that x_2 equals 4
x_1	1	0.25/0.55	0.10/0.15	0.05/0.15	0.05/0.15
	2	0.10/0.55	0.05/0.15	0.00/0.15	0.10/0.15
	3	0.20/0.55	0.00/0.15	0.10/0.15	0.00/0.15

Once you understand how to calculate an expected value and can derive a conditional distribution function, there is no great trick to calculating the conditional expected value of a random variable. What, for example, is the expected value of x_1 conditional on x_2 being equal to 1? Conditional on x_2 being equal to 1 we know that x_1 will equal 1 with a probability of $\frac{0.25}{0.55}$; we know that x_1 will equal 2 with a probability of $\frac{0.10}{0.55}$; and, finally, that x_1 will equal 3 with a probability of $\frac{0.20}{0.55}$. This means

$$E(x_1 | x_2 = 1) = 1 \cdot \frac{0.25}{0.55} + 2 \cdot \frac{0.10}{0.55} + 3 \cdot \frac{0.20}{0.55} = 1.909.$$

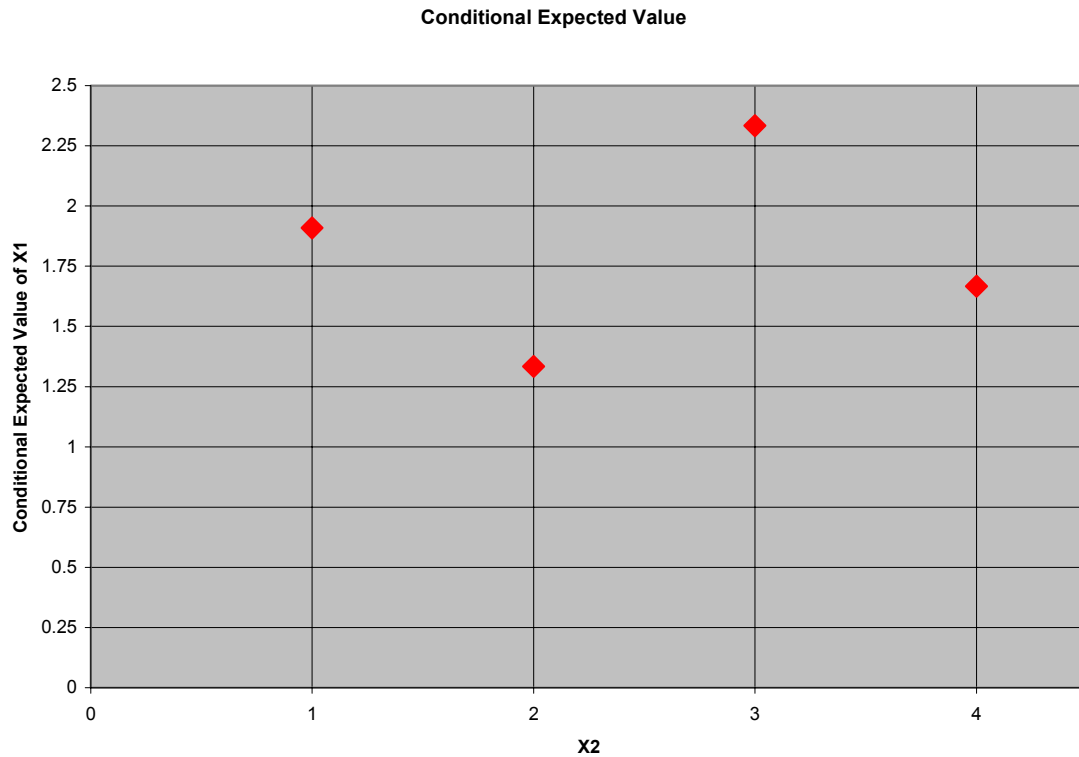
This same type of calculation can be carried out for all other values of x_2 .

$$E(x_1 | x_2 = 2) = 1 \cdot \frac{0.10}{0.15} + 2 \cdot \frac{0.05}{0.15} + 3 \cdot \frac{0.00}{0.15} = 1.333$$

$$E(x_1 | x_2 = 3) = 1 \cdot \frac{0.05}{0.15} + 2 \cdot \frac{0.00}{0.15} + 3 \cdot \frac{0.10}{0.15} = 2.333$$

$$E(x_1 | x_2 = 4) = 1 \cdot \frac{0.05}{0.15} + 2 \cdot \frac{0.10}{0.15} + 3 \cdot \frac{0.00}{0.15} = 1.666$$

The relationship between x_2 and the conditional expected value of x_1 is shown below.



If this conditional expectation was linear we would have a “linear regression function,” something very near and dear to Professor Kanazawa’s heart.

COVARIANCE

The next to last concept we will pay attention to in this handout measures the degree to which two random variables move with each other or against each other. This is often captured by the covariance of the random variables (another important measure that measures the same concept is the correlation coefficient).

*Definition: The **Covariance** of two random variables, x and y , is*

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (x_i - E(x))(y_j - E(y))P(x_i, y_j)$$

or,

$$E((x - E(x))(y - E(y)))$$

With enough patience you ought to be able to show that if the joint probability function for two random variables is given by,

		x_2			
		1	2	3	4
x_1	1	0.25	0.10	0.05	0.05
	2	0.10	0.05	0.00	0.10
	3	0.20	0.00	0.10	0.00

then the covariance between x_1 and x_2 is equal to 1.9. The fact that this number is positive suggests that the two random variables tend to move in the same direction.

You will find that the following manipulation of the definition of covariance is useful.

$$Cov(x, y) = E((x - E(x))(y - E(y)))$$

$$Cov(x, y) = E(xy - xE(y) - yE(x) + E(x)E(y))$$

$$Cov(x, y) = E(xy) - E(x)E(y) - E(x)E(y) + E(x)E(y)$$

$$Cov(x, y) = E(xy) - E(x)E(y)$$

CORRELATION COEFFICIENT

A correlation coefficient between two random variables is linked, as you might imagine, closely to the covariance of those two random variables.

*Definition: The **Correlation Coefficient** between two random variables is*

$$\rho = \frac{\text{Cov}(x, y)}{\sqrt{V(x)}\sqrt{V(y)}}$$

It turns out that ρ must lie between -1 and 1 and is a measure of the degree of linear association between x and y . If for no other reason than it is unit-free it is easier to use than the covariance as a measure of how x and y move together.