

Statistics and Data Analysis

B01.1305

Professor William Greene

Phone: 212.998.0876

Office: KMC 7-78

Home page: www.stern.nyu.edu/~wgreene

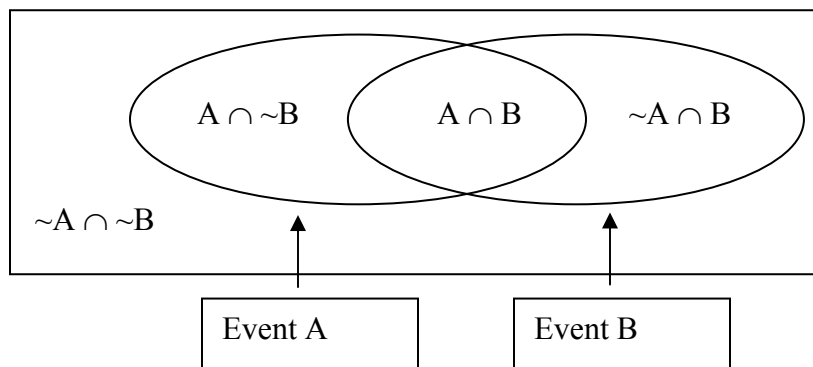
Email: wgreene@stern.nyu.edu

Course web page: www.stern.nyu.edu/~wgreene/Statistics/Outline.htm

Sample Probability Problems

Example 1. Probability Theory

For two events A and B, $P(A) = 0.30$, $P(B) = 0.42$, $P(A \cup B) = 0.61$. What is $P(\sim A \cup \sim B)$? A Venn diagram helps. For two events, there are four regions, $A \cap B$, $A \cap \sim B$, $\sim A \cap B$, $\sim A \cap \sim B$.



A general result is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can deduce from the numbers given that $.61 = .30 + .42 - P(A \cap B)$ so $P(A \cap B) = .11$. The .11 is the area of the region where A and B overlap. The event $\sim A \cup \sim B$ is everything that is outside A or outside B, which is everything that is not in this overlap region. Therefore, $P(\sim A \cup \sim B) = 1 - P(A \cap B) = .89$.

Example 2. Basic Probabilities

Consider a case in which there are 100 employees of a modest-sized firm. Let's suppose the cross-classified information is this:

	Drives to work	Uses mass transit	TOTAL
Male	45	15	60
Female	23	17	40
TOTAL	68	32	100

The category “Drives to work” refers to driving alone. People in car pools are in the “Uses mass transit” category. Suppose that one employee is selected at random.

A. Find

$$\begin{aligned} P(\text{female}) &= 40/100 \\ P(\text{drives to work}), &= 68/100 \\ P(\text{female or drives to work}) &= 40/100 + 68/100 - 23/100 = 85/100. \end{aligned}$$

B. Now, find $P(\text{Drives to work} \mid \text{Male})$.

If A and B are two events with $P(B) > 0$, then

$$P(A \mid B) = \text{“probability of } A, \text{ given that } B \text{ has already occurred,“} = P(A \cap B)/P(B)$$

Certainly $P(\text{male}) = 60/100 = 0.6$.

Then

$$\begin{aligned} P(\text{Drives to work} \mid \text{Male}) &= P(\text{Drives to work} \cap \text{Male})/P(\text{Male}) \\ &= (45/100) / (60/100) = 45/60 = 3/4 = 75\% \end{aligned}$$

This has applied the definition literally. One could also use the “obviousness” idea; there are 60 men and exactly 45 of them drive to work. The probability must be $45/60 = 3/4 = 75\%$.

Note that $P(\text{Male} \mid \text{Drives to work}) = 45 / 68 \approx 0.6618 \approx 66\%$ and this is certainly *not* equal to $P(\text{Drives to work} \mid \text{Male}) = 45/60 = 0.75 = 75\%$.

Example 3. Random Draws and Counting to Obtain Probabilities

An urn (jar) contains 10 balls:

6 green balls 5 balls striped 1 ball solid
4 red balls 3 balls striped 1 ball solid

Some people find it helpful to have this information in a table.

	Solid	Striped	TOTAL
Green	1	5	6
Red	1	3	4
TOTAL	2	8	10

If you select a ball at random, find

$$\begin{aligned} P(\text{green}) &= 6/10 = 3/5 \\ P(\text{red}) &= 4/10 = 2/5 \\ P(\text{solid} \mid \text{red}) &= 1/4 \text{ [also note, } (1/10)/(4/10)\text{]} \\ P(\text{red} \mid \text{solid}) &= 1/2 \text{ [(1/10)/(1/20)]} \\ P(\text{striped} \mid \text{green}) &= 5/6 \text{ [(5/10)/(6/10)]} \\ P(\text{red} \mid \text{green}) &= 0 (?) \\ P(\text{solid} \mid \text{green}) &= 1/6 \text{ [(1/10)/(6/10)]} \end{aligned}$$

Example 4: Independent Events. For cars at an inspection, the probability that a car will fail emissions test is 0.12. The probability that it will fail both emissions test and brake test is 0.02. Given that it fails the emissions test, what is the probability that it will fail the brake test?

$$P(\text{fail brake} \mid \text{fail emission}) = P(\text{fail brake and emission})/P(\text{fail emission}) = .02/.12 = 1/6.$$

Example 5. Product Rule. A firm produces chips in four plants, creatively named P1, P2, P3 and P4. Production rates and defect rates at the four plants are as follows:

Plant:	P1	P2	P3	P4
Production	35%	16%	21%	28%
Defect Rate	2%,	3%	7%	1%

(Why would the firm continue to produce at Plant 3 with its atypically high defect rate? Perhaps the marginal cost of production at Plant 3 is significantly lower than the others, so the expected cost of the defects could be lower.) Let D be the event that a chip is produced with a defect. Compute P(D). If a defective chip is produced, what is the probability that it is produced in P3?

The problem asks for P(D). The data above give $P(D|P1) = .02$ and so on. We also have $P(P1) = .35$, $P(P2) = .16$, $P(P3) = .21$ and $P(P4) = .28$. We require

$$\begin{aligned}
 P(D) &= P(D \cap P1) + P(D \cap P2) + P(D \cap P3) + P(D \cap P4) \\
 &= P(D|P1)P(P1) + P(D|P2)P(P2) + P(D|P3)P(P3) + P(D|P4)P(P4) \\
 &= .02(.35) + .03(.16) + .07(.21) + .01(.28) \\
 &= .0070 + .0048 + .0147 + .0028 \\
 &= .0293 = 2.93\%.
 \end{aligned}$$

The second question asks for $P(P3|D)$. Use the definition,

$$P(P3|D) = P(D \cap P3)/P(D) = P(D|P3)P(P3)/P(D) = .07(.21)/.0293 = .5015.$$

Example 6. Conditional Probabilities. In a certain city, taxis are easily identified by rooftop lights. There are two taxi companies in this city, Green and Blue, with cabs colored accordingly. One night, there is a hit-and-run accident involving a taxi. The vehicle was clearly a taxi, as multiple witnesses saw the rooftop light. However, only one witness (Ralph) claimed to be able to identify the cab by color. Ralph claimed that this was a Green cab. We would like to know the probability that the accident was really committed by a Green cab. Here are some facts:

- (1) Ralph was given an identification test under lighting conditions similar to those the night of the accident. When the test cab was blue, he identified it correctly 80% of the time.
- (2) When the test cab was green, he identified it correctly 80% of the time. A number of other people were given the same identification test, and they also produced results close to the 80% that Ralph got.
- (3) In this town, 85% of the taxis are Blue.

(1) means $P(\text{Ralph says Blue} | \text{Cab is Blue}) = .80$

(2) means $P(\text{Ralph says Green} | \text{Cab is Green}) = .80$

(3) states $P(\text{Cab is Blue}) = .85$.

(4) this means that $P(\text{Cab is Green}) = .15$ since all cabs are either Blue or Green

From the facts, we can deduce immediately:

(1) and (3) imply $P(\text{Ralph says Blue and Cab is Blue}) = .80 \times .85 = .68$ (Product rule)

(2) and (4) imply $P(\text{Ralph says Green and is Cab is Green}) = .80 \times .15 = .12$

This gives us, with the immediate facts bold and underlined.

		Ralph says	
Cab really is	Blue	Green	Total
Blue	<u>.68</u>	.17	<u>.85</u>
Green	.03	<u>.12</u>	.15
Total	.71	.29	1.00

In the first row, we see that $P(\text{Ralph says Green and Cab is Blue}) = .85 - .68 = .17$

In the second row, we see that $P(\text{Ralph says Blue and Cab is Green}) = .15 - .12 = .03$.

Adding down the columns, we can see that $P(\text{Ralph says Blue}) = .71$ and $P(\text{Ralph says Green}) = .29$.

This is all we need. We now want $P(\text{Cab is Green} \mid \text{Ralph says Green})$

$$= P(\text{Ralph says Green and Cab is Green}) / P(\text{Ralph says Green}) = .12 / .29 = .4138.$$

Example 7. Conditional Probability. At Amherst College (not the one in Massachusetts), 72% of business students read the Wall Street Journal (28% read the National Enquirer) while only 20% of nonbusiness students read the Wall Street Journal. (With the change of ownership, the two newspapers will probably resemble each other.) A total of 14% of the students are business majors. Find the probability that a student who reads the Wall Street Journal is a business major. Find also the probability that a student who does not read the Wall Street Journal is a nonbusiness student.

Denote by B the “event” business student and $\sim B$ by nonbusiness student. Denote by W the “event” reads the Wall Street Journal. The facts give us

$$P(W|B) = .72,$$

$$P(\sim W|B) = .28$$

$$P(W|\sim B) = .20$$

$$P(B) = .14$$

We can see immediately that $P(\sim B) = 1 - .14 = .86$. The question asks for $P(B|W)$ and $P(\sim B|\sim W)$. This can be an application of Bayes Theorem, but it is just as simple to deduce the necessary information first. We will need the joint and marginal probabilities,

	B	$\sim B$	Total
W	.1008	.1720	.2728.
$\sim W$.0392	.6880	.7272
Total	<u>.1400</u>	.8600	1.0000

Using the product rule,

$$P(W \text{ and } B) = P(W|B)P(B) = .72(.14) = .1008$$

$$P(\sim W \text{ and } B) = P(\sim W | B)P(B) = .28(.14) = .0392$$

This fills the first column.

$$P(W \text{ and } \sim B) = P(W|\sim B)P(\sim B) = .20(.86) = .1720$$

Adding across, $P(W) = .1008 + .1720 = .2728$. This fills the first row. We can find the values in the second row just by subtraction. $P(\sim W) = 1 - P(W) = 1 - .2728 = .7272$.

Finally, $P(\sim W \text{ and } \sim B)$ is found by making the second column sum to .8600; it is .6880.

Then, $P(B|W) = P(B \text{ and } W) / P(W) = .1008 / .2728 = .3695$ and

$P(\sim B|\sim W) = P(\sim B \text{ and } \sim W) / P(\sim W) = .6880 / .7272 = .9461$.

Example 8. Using Complementary Probabilities

Sometimes in order to find the probability that an event occurs, it is much simpler to find the probability that it does not occur. Here are a couple examples:

Example 8.1. You are going to play roulette ten times. Each time you will make the same bet, all red numbers. What is the probability that you win at least once? Directly, this is

$$P(\text{at least 1 win}) = P(1 \text{ win}) + P(2 \text{ wins}) + P(3 \text{ wins}) + \dots + P(10 \text{ wins}).$$

This promises to be formidable, but there is a much easier way. Winning at least once means NOT losing every time. So, to find the desired probability, we can just find

$$P(\text{at least 1 win}) = 1 - P(0 \text{ wins})$$

and $P(0 \text{ wins})$ is easy. The spins are independent and the probability of losing on any given spin is $P(\text{lose}) = P(\text{not red}) = P(\text{black or green}) = 18/38 + 2/38 = 20/38$ or $10/19$. Therefore,

$$P(\text{lose every time}) = (10/19)^{10} = .001631.$$

$$\text{Therefore, } P(\text{at least 1 win}) = 1 - .001631 = .998369.$$

Example 8.2. (A classic – the birthday problem). There are (assume) 50 people in this class. What is the probability that at least two of them have the same birthday? This one promises to be extremely messy. $P(\text{at least two same}) = P(\text{exactly two have the same}) + P(\text{three have the same}) + \dots$ a lot of combinations of twos and threes and so on. But, the event (at least two have the same birthday) is the complement of (everyone has a different birthday). This is fairly easy to work out. Person 1 has a birthday. $P(\text{person 2 has a different birthday}) = 364/365$. $P(\text{person 3 has a different birthday from both 1 and 2}) = 363/365$. The joint probability is the product, so what we are looking for is

$$P(\text{all different birthdays}) = (365/365)(364/365)(363/365)\dots(316/365) = .029626.$$

It follows that the probability that at least two people have the same birthday is .970374.

Example 9. Sampling and Probability. A researcher surveys individuals randomly at a supermarket. Each sampled individual is asked two questions:

- a. Do you use coupons at this store?
- b. Do you read PennSaver? (This is a local advertising newspaper that is mailed to every address in the town.)

The tabulated results are as follows, where event C denotes uses coupons, $\sim C$ means does not use coupons, P denotes reads PennySaver and $\sim P$ denotes does not read PennySaver:

		PennySaver?		
		P	$\sim P$	TOTAL
Coupons?	C	30	90	120
	$\sim C$	10	70	80
TOTAL		40	160	200.

Assuming that these represent a random sample from the population, estimate the following:

1. The probability that a person who reads PennySaver uses coupons. Since the person is known to read PennySaver, this is a conditional probability, $P(C|P) = P(C,P) / P(P) = (30/200) / (40/200) = 3/4$
2. The probability that a person who does not use coupons reads PennySaver. This is also a conditional probability, since it is given that the person does not use coupons. $P(P|\sim C)$ is computed the same way as in part 1. $P(P|\sim C) = P(P,\sim C) / P(\sim C) = (10/200) / (80/200) = 1/8$.
3. The probability that a person does not use coupons and does not read PennySaver. This says nothing about conditioning, so the answer is just $P(\sim P,\sim C) = 70/200 = .35$.
4. The probability that a person uses coupons. This is a marginal probability. $P(C) = P(C,P) + P(C,\sim P) = 30/200 + 90/200 = 120/200 = .6$. You could also read this directly off the margin of the table.
5. The probability that a person reads PennySaver. This is computed the same as the result in the previous problem. It is a marginal probability, $P(P) = 40/200$.

This problem describes a situation of cross section (naturalistic) sampling. In the medical research setting, this is known as “community based” sampling. This is as opposed to the type of sampling described in Example 10.

Example 10. Retrospective Sampling. A medical researcher is interested in whether the consumption of fish oil is related to chronic fatigue syndrome (CFS). He obtains a sample of 60 suffers of CFS from the records of local doctors and determines whether these 60 people consume high or low amounts of fish oil. He then obtains a control sample of 140 people who do not suffer from CFS and determines their rates of consumption of fish oil. The sample results are as follows:

row

	Chronic Fatigue Syndrome		Total
	Yes	No	
Fish oil high? Yes	12	56	68
Fish oil high? No	48	84	132
Total	60	140	200

For convenience, let C denote has CFS, $\sim C$ denote does not have CFS, H denotes high fish oil, $\sim H$ denotes not high (low) fish oil.

a. What is the estimated probability that a person with CFS has a high fish oil consumption? We are looking for $P(H|C)$. The table gives conditional values. That is, since the sample included 60 people specifically with CFS, then $P(H|C)$ is given directly as $12/60$.

b. What is the probability that a person with high consumption of fish oil has CFS? This is $P(C|H)$. Unfortunately, we cannot determine this from the table. The values we have are $P(H|C) = 12/60$ and $P(\sim H|C) = 48/60$. The 60 is not, in fact, useful, once we know these probabilities. $60/200$ is not $P(C)$ because the 60 observations were not randomly sampled – they were 60 people known to have CFS. Without $P(C)$, we cannot determine the probability $P(H,C) = P(H|C)P(C)$.

c. What is the probability that a person does not have CFS and also has a low consumption of fish oil. This would be $P(\sim C, \sim H)$. But, here, again, we have the same problem as in the previous question. We do not know the marginal probabilities.

d. What is the probability that a person has CFS. This is the problem we encountered in questions 3 and 4. Note that the researcher deliberately selected 60 people out of 200. We have no idea if the 60 out of 200 is representative of the full population.

e. What is the probability that a person has a high consumption of fish oil. This is $P(H)$. Same problem as before.

By this type of sampling, the researcher has limited the information to $P(H|C) = 12/60$, $P(\sim H|C) = 48/60$, $P(H|\sim C) = 56/140$, $P(\sim H|\sim C) = 84/140$. (The similar computations in the opposite direction are meaningless. Thus, $12/68$ does not estimate $P(C|H)$. Unfortunately, because of the way the sample was constructed, it does not estimate anything.

f. Suppose you knew that $P(C) = .02$. With this information, you can deduce the full set of values in the table.

We know that $P(H|C) = 12/60$. So, $P(H,C) = P(H|C)P(C) = (12/60).02 = .004$.

Also, $P(\sim H,C) = P(\sim H|C)P(C) = (48/60).02 = .016$

And, if $P(C) = .02$, then $P(\sim C) = .98$. So, $P(H,\sim C) = P(H|\sim C)P(\sim C) = (56/140).98 = .392$

Finally, $P(\sim H,\sim C) = P(\sim H|\sim C)P(\sim C) = (84/140).98 = .588$. This gives you most of the missing values in the probability distribution:

	Chronic Fatigue Syndrome		Total
	Yes	No	
Fish oil high? Yes	.004	.392	?
Fish oil high? No	.016	.588	?
Total	.02	.98	1.00

The two question marks in the table are $P(H)$ and $P(\sim H)$. But, these can be found now just by adding across. That is, $P(H) = P(H,C) + P(H,\sim C) = .004 + .392 = .396$. $P(\sim H)$ is just $1 - P(H) = 1 - .396 = .604$. So, the full set of results is

	Chronic Fatigue Syndrome		Total
	Yes	No	
Fish oil high? Yes	.004	.392	.396
Fish oil high? No	.016	.588	.604
Total	.02	.98	1.00

Note that you might be able to obtain data such as $P(C)$, the marginal proportion of the population that has CFS, through medical records. It's not likely that you could obtain the marginal proportion of the population that consumes a high amount of fish oil, not at least through public information. You would have to take a better designed sample.

g. What is the estimated probability that a person with high consumption of fish oil has CFS. This would be $P(C|H) = P(C,H)/P(H) = .004/.396 = .0101$.

h. What is the estimated probability that a person who does not have CFS also has low consumption of fish oil. This would be $P(\sim H|\sim C) = P(\sim H,\sim C)/P(\sim C) = .588/.98 = .6$.

i. What is the probability that a person does not have CFS and has low consumption of fish oil? This is the .588 that is in the table. Note the subtle difference in the language. Problem 8 asks for a conditional probability while problem 9 asks for a joint probability.

Example 9 dealt with a situation of naturalistic sampling. Example 10 was based on a retrospective sample. The sample was based on individuals who had the effect (CFS). A third possibility would be prospective sampling – select a sample of individuals who have high and low fish oil consumption, and determine (now or later) whether they contract CFS.