Module 7.11: A Worksheet on Independence, Repetition, and Bernoulli's Formula

Gregory V. Bard

January 24, 2017

1 Theory

- Technically, the definition of independence is

  The events A and B are independent if and only if $Pr\{A \cap B\} = Pr\{A\}Pr\{B\}$

- However, we use this definition in two distinct ways.
  - Sometimes, you want to test if A and B are independent. Then check to see if $Pr\{A \cap B\} = Pr\{A\}Pr\{B\}$ is true.
    - If they are exactly equal, then you know A and B are independent for sure.
    - If they are far apart, then you know A and B are definitely not independent.
    - If they are close, then there are tests from advanced courses in statistics that will cover how to handle this situation.
  - Sometimes, you know A and B are independent. In that situation, $Pr\{A \cap B\} = Pr\{A\}Pr\{B\}$ is a handy formula.
  - Another handy fact is that if A and B are independent, then for sure, A^c and B^c are independent.

- Of course, if event A occurs with probability p, and you make two independent attempts at event A, you get event A both times with probability $(p)(p) = p^2$.

- That last bullet generalizes to more than 2 repetitions. This gives us the two repetition formulas:
  - If you have n independent repetitions of an attempt of event A, and if each attempt succeeds with probability p, then the probability of all n attempts succeeding is p^n.
  - If you have n independent repetitions of an attempt of event A, and if each attempt succeeds with probability p, then the probability of all n attempts failing is $(1-p)^n$.

Now if I have n repetitions of an event, the last two bullets give me the probability of n success and 0 failures, or 0 successes and n failures. What about situations in between these two extremes?

Bernoulli's Formula If you have n independent repetitions of an attempt of event A, and if each attempt succeeds with probability p, the probability that A actually happens x times, (and therefore fails to happen $n-x$ times), is given by

$$C_{n,x}(p^x)(1-p)^{n-x}$$

Note, if $x = 0$ we have

$$C_{n,x}(p^x)(1-p)^{n-x} = C_{n,0}(p^0)(1-p)^{n-0} = (1)(1)(1-p)^n = (1-p)^n$$

and if $x = n$ we have

$$C_{n,x}(p^x)(1-p)^{n-x} = C_{n,n}(p^n)(1-p)^0 = (1)(p^n)(1) = p^n$$

which means that the repetition formulas are just "special cases" of Bernoulli's formula. On the other hand, using the repetition formulas can be vastly easier than using Bernoulli's formula.

2 Questions

1. Suppose that on a given college campus, 95% of the students have consumed alcohol, 60% of the students have smoked marijuana, and 57% of the students have used both. Is the probability of a randomly selected student (from that campus) having smoked marijuana independent of a randomly selected student (from that campus) having consumed alcohol?

2. Let us imagine that among a certain cheaply-made smartphone, the screens have flaws with probability 3%, and the batteries have flaws with probability 5%. Bob is going to assume that these flaws are independent, because there are two different companies supplying the parts. If a smartphone has no flaws, it is shipped out for sale. If it has one flaw, it is kept for refurbishing. If it has both flaws, it is tossed out. Compute the probability distribution for smartphones being shipped, refurbished, and tossed.

3. Continuing with the previous problem, suppose that 1% of the phones are found to have both flaws. Was Bob's assumption of independence correct? What is the new probability distribution?

Note: Do you remember, back in the module "A Formal Introduction to Probability," when I told you that there were five common errors? I mentioned that one of them is hard to explain, deals with independence, and would be explained in this module. The previous two problems illustrate this common pitfall. It is a common mistake to assume independence when the events might not necessarily be independent.

Background: Some of us have heard of "cancer alley," the portion of the Mississippi river between Baton Rogue and New Orleans. This area had (at least for a time) an unusually high rate of cancer—and some estimated it to be nine times the national average. The movie *Erin Brockovich* also told the story of another area with a high density of cancer, but in California. The Centers for Disease Control use all sorts of mathematical techniques to detect these cancer "hotspots." In turn, identifying these locations can help uncover the underlying causes (such as industrial pollution). Then the problems can be remedied, preventing much unnecessary loss of life. It is interesting for us to explore the mathematics that underlies these investigative techniques. We will do that in the following problem.

(For the more human aspects of investigating cancer hotspots, it is worth noting that the movie *Erin Brockovich* is only 2 hours long and has won many awards.)

4. Imagine a small town, with a population of 9000 people. There is a petrochemical factory there, and 900 of the town's residents work there. There are 234 people in the town who have cancer, or who have had cancer, and 72 of them work at the factory. The CDC has dispatched an epidemiologist to this town to investigate. The factory's public relations team notes that *more than two thirds* of the cancer patients in the town are *not employed at the factory in any capacity whatsoever*. Moreover, 92% of the employees of the factory are cancer free. This might make a naïve person think that there is no link between the factory and cancer. The epidemiologist, of course, can use the mathematics of probability to see through this metaphorical cloud of smoke.

   Note: As is standard, we will use the phrase "has cancer" as an abbreviation for "has cancer, or has had cancer." We will use the phrase "is cancer free" as an abbreviation for "does not have cancer, nor has ever had cancer."

   (a) What is the probability that a random person in the town has cancer?

   (b) What is the probability that a random person in the town works for the factory?

   (c) What is the probability that a random person in the town both works for the factory and has cancer?

   (d) Are the events "this person works for the factory" and "this person has cancer" independent?

   (e) What is the probability that a random factory worker has cancer?

   (f) What is the probability that a random *resident who is not a factory worker* has cancer?

   (g) Last but not least, were either of the claims from the factory's public relations office false?

Hint: it might be helpful to make a Venn Diagram to organize the data, with one circle being the people who have cancer, and one circle being the people who work for the factory.

5. Let us suppose that the probability of Alice being in class today is independent of the probability of Bob being in class today. (It is worth thinking about the reliability of this assumption.) If Alice attends with probability 95% and Bob attends with probability 85%, then what is the probability of both of them being present? both absent? one of them present and the other absent?

6. Imagine that you have an internship at a factory that manufactures smart watches. The suppliers of the components are fairly good, as it turns out. There are three components that might be defective: the accelerometer (to measure movement), the screen, and the battery. These fail with probability 0.005, 0.02, and 0.01. All other defects are extremely rare. Since the components come from completely different suppliers, scattered around the world, the probabilities of defects are definitely independent. What is the probability that a random watch will have no defective components? What is the probability of a random watch having at least one defect?

7. As we learned in the module "A Formal Introduction to Probability," there is an approximately 2% chance that a rocket launch of a satellite will end in a complete disaster, such as the rocket exploding on the launch pad or the stages failing to separate and the rocket plummeting into the ocean. Of course, that also means there is a high probability of a disaster not happening at all. In any case, a new manager at a commercial satellite-launching firm has recently arrived, and has chided the launch operations department for the fact that they have never had 40 consecutive launches without an accident—their record for consecutive launches without an accident 38.

   To help the new manager understand matters, compute the probability of 39 consecutive launches without a single failure.

8. Suppose your friend Speedy is a very bad driver, and gets into a car accident with probability 1% on any given day. If he drives every day for one year, what is the probability that he gets into one or more car accidents during that year? Use a 365-day year.

9. Suppose a new website is booming. The company decides to get 8 servers, scattered around the world. Since the servers are around the world, independence is a fair assumption. Each server has a 99% up time. What is the probability that, at any given moment, there is at least one server that is down?

10. Suppose that in a particular Alaskan oil field, the chance of striking oil after drilling is 5%. If the drilling sites are sufficiently far apart, then the

attempts are independent. The boss wants to report a successful oil strike this month, because the COO is coming to visit. He thinks that 20 drillings should guarantee a success. He thinks this because $(5\%)(20) = 100\%$.

Of course, he's wrong. If there are 20 drillings, sufficiently separated geographically to guarantee independence, what is the probability of striking oil at least once?

11. Re-examine the previous problem. Using the repetition formulas, compute the number of drillings required, so that the chance of striking oil at least once is 95% or greater.

12. This problem was suggested by Prof. Seth Dutter. A study by condom manufacturers noted that condoms are 99.5% effective. Another study tracked heterosexual couples, and noted that those who rely on condoms alone for one year had a 50% chance of becoming pregnant. Of course, the former study is *per episode* and the latter study is *per year*. However, using the repetition formulas, we can solve for n, and determine the average number of episodes per year of the couples in the second study.

Note: None of the problems above this point require Bernoulli's formula, but all the problems after it do.

13. A student has a multiple choice test of ten questions with five options each. Because the student has not studied, he has to guess. However, he is not entirely clueless, so he has a 40% chance of getting each question correct, instead of the 20% represented by blind guessing. What is the probability distribution of his performance? (i.e. the probability of getting x correct and $10 - x$ wrong, for $x \in \{0, 1, 2, \ldots, 10\}$.)

14. We return to the online company (from Problem 9) with 8 servers. Each server has 99% up-time. The servers are scattered in hosting centers around the planet, so independence is a very fair assumption. What does Bernoulli's formula give as the probability distribution for how many servers are up versus down at any given moment? (For example, we might want to know the probability of 6-up and 2-down.)

15. Now we return to the oil-drilling problems (Problems 10 and 11). Find the probability distribution for the number of oil strikes, for x strikes with $x \in \{0, 1, 2, \ldots, 15\}$. Assume that the boss went with your suggestion of 59 drillings. Alternatively, just try $x \in \{0, 1, 2, 3, 4\}$ if you prefer, because it is a lot of work to use the formula sixteen times.

16. Suppose that it is known that an altimeter on a particular model of commercial aircraft breaks with probability 1 in 800 on any given flight. An aircraft design firm has a decision to make: should they have five altimeters, taking the common output of 3 out of 5 to be the trustworthy altitude? Or should they have seven altimeters, taking the common output

of 4 out of 7? Or should they have three altimeters, taking the common output of 2 out of 3?

To help make this decision, make the Bernoulli Table for 3 altimeters. Then, compute the probability of 2 or 3 altimeters being broken using that table.

17. Continuing with the previous problem, suppose that an intern gives you the Bernoulli Table for 5 altimeters. However, he spilled coffee on it, and a coffee stain has wiped out two entries, as shown. Find the two missing entries.

```
Probability 0 up and 5 down: 3.05175781250000e-15
Probability 1 up and 4 down: 1.21917724609375e-11
Probability 2 up and 3 down: 1.94824523925781e-8
Probability 3 up and 2 down: coffee stain
Probability 4 up and 1 down: coffee stain
Probability 5 up and 0 down: 0.993765605480954
```

18. Tell me why it was necessary, in the previous problem, that the coffee stain knock out two entries? Why would the math problem be broken (that is to say, far too easy) if I only had the coffee stain wipe out one entry of the Bernoulli Table?

3 Answers

1. $Pr\{M\}Pr\{A\} = (0.6)(0.95) = 0.57$ and $Pr\{M \cap A\} = 0.57$ so, yes. The events are independent.

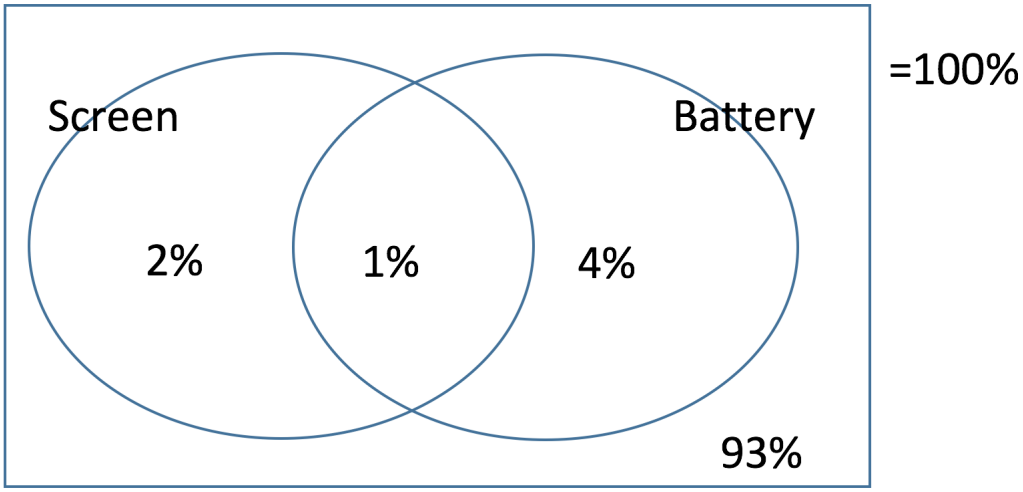
2. We will proceed by steps:

   - The probability of both flaws is $Pr\{S \cap B\} = Pr\{S\}Pr\{B\} = (0.03)(0.05) = 0.0015$ or 0.15%.
   - The probability of neither flaw is $Pr\{S^c \cap B^c\} = Pr\{S^c\}Pr\{B^c\} = (0.97)(0.95) = 0.9215$, or 92.15%.
   - The probability of refurbishment is $1 - 0.0015 - 0.9215 = 0.077$, or 7.70%.

3. The independence assumption was wrong, because

$$Pr\{S \cap B\} = Pr\{S\}Pr\{B\} = (0.03)(0.05) = 0.0015 \neq 0.01$$

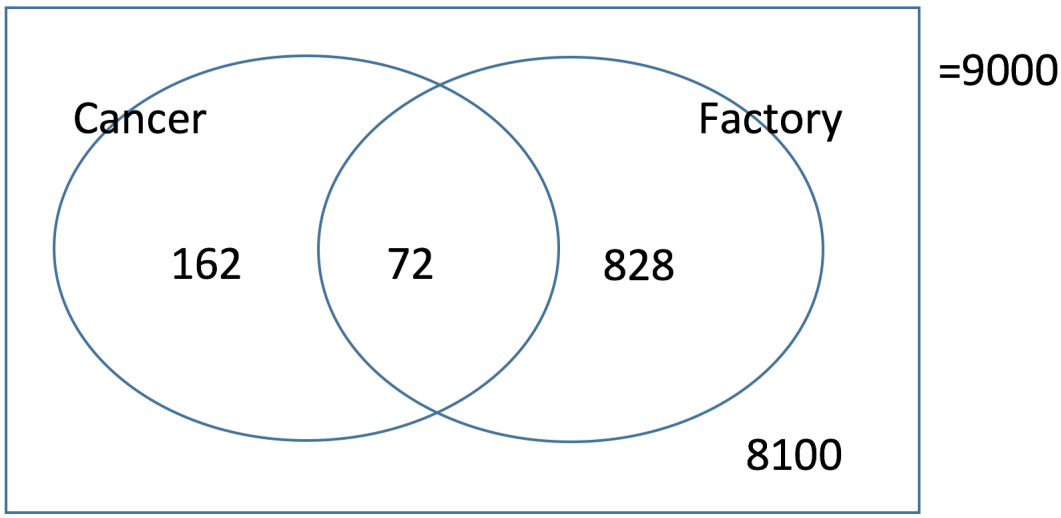
   but then we can still compute the probability distribution by making a Venn Diagram.



We conclude that 93% will be shipped out for sale, 6% will be refurbished, and 1% will be thrown out.

4. Here are the answers to the question about the town with a cancer problem.

Hint: We begin by making the Venn Diagram.



(a) $Pr\{\text{Cancer}\} = \frac{234}{9000} = 0.026$.
(b) $Pr\{\text{Factory}\} = \frac{900}{9000} = 0.1$.
(c) $Pr\{\text{Cancer} \cap \text{Factory}\} = \frac{72}{9000} = 0.008$.

(d) First we compute,

$$Pr\{\text{Cancer}\}Pr\{\text{Factory}\} = (0.026)(0.1) = 0.0026$$

and of course

$$0.0026 \neq 0.008$$

which means that clearly, the events are not independent.

(e) The probability that a random factory worker has cancer is $\frac{72}{900} = 0.08$.

(f) The probability that a random person who is not a factory worker has cancer is $\frac{162}{8100} = 0.02$.

Summary: As you can see, the probability of a person who works for the factory having cancer is *four times higher* than the probability of a person who does not work for the factory having cancer. This is obviously a serious problem. Moreover, the overall rate of cancer in the town is elevated, but still within the realm of possibility.

(g) Let's look now at those two claims from the public relations department:

- There are 234 people who have cancer. Two thirds of 234 is 156. There are 162 people who have cancer but who do not work at the factory. Since $162 > 156$, it is true that more than two-thirds of the cancer patients in the town are not employed at the factory. Adding the phrase "in any capacity whatsoever" is a common smoke screen. You bring undue emphasis on a minor irrelevant detail to distract from other aspects which might be embarrassing.
- Since 72 out of 900 employees of the factory have cancer, then 828 employees do not. We compute $828 \div 900 = 0.92 = 92\%$. This statistic is exactly true. Of course, for something potentially fatal like cancer, 92% is far too low. You'd like 98% or 99% of the town to be cancer-free, or better.

5. We have three things to compute.

   - Let A be the probability that Alice is present, and let B be the probability that Bob is present.
   - The probability of both present is $Pr\{A \cap B\} = Pr\{A\}Pr\{B\} = (0.95)(0.85) = 0.8075$.
   - The probability of both absent is $Pr\{A^c \cap B^c\} = Pr\{A^c\}Pr\{B^c\} = (0.05)(0.15) = 0.0075$.
   - The probability of one-absent/one-present is $1 - 0.8075 - 0.0075 = 0.185$.

6. A lot of students, without thinking, will simply compute

$$(0.005)(0.02)(0.01) = 10^{-6}$$

which is one-in-a-million. That's actually the probability that, by coincidence, a smartwatch has all three defects.

We do not want defective components—we want working components! Therefore, we compute the complements

$$1-0.005 = 0.995 \qquad 1-0.02 = 0.98 \qquad 1-0.01 = 0.99$$

and then multiply to obtain

$$(0.995)(0.98)(0.99) = 0.965349$$

which means that there is 96.5349% chance that a watch has no defective components. The complement of that

$$1 - 0.965349 = 0.034651 = 3.4651\%$$

is the probability of a random watch having at least one defect.

Note that it is wrong, but a useful approximation, to compute

$$0.005 + 0.02 + 0.01 = 0.035 = 3.5\%$$

7. The key to this problem is to realize that we are repeating a successful launch, not a failed launch. If $\mathcal{D}$ is the probability of a disaster, then we know $Pr\{\mathcal{D}\} = 0.02$. While this is true, that's not the event that we're repeating. We want to repeat $\mathcal{D}^c$, and surely

$$Pr\{\mathcal{D}^c\} = 1 - 0.02 = 0.98$$

With that in mind, the probability of 39 consecutive launches without a single failure is

$$(0.98)^{39} = 0.454796\cdots$$

which is clearly less than half. While 45.4796% is not a small probability by any means, we should not be surprised that this event did not occur.

In contrast, many students will write

$$(0.02)^{39} = 5.49755\cdots \times 10^{-67}$$

which is phenomenally small. As it turns out, this is the probability that all 39 launches have an unfortunate accident (such as exploding on the launch pad), and none of them survive the launch process.

To put this in perspective, there are about 1.8×10^{57} protons, neutrons, and electrons in the entire solar system. For comparison, 10^{67} is ten billion times as large as 10^{57}. There is no easy way for us to wrap our mind around 10^{-67}.

8. The key here is to realize that he has a 1% chance of getting into an accident each day, but that means he has a 99% chance of not getting into an accident each day. We're actually interested in the latter case. We compute

$$(0.99)^{365} = 0.0255179\cdots$$

is the probability that he will get into 0 car accidents in 365 consecutive days. The complement principle gives us

$$1 - 0.0255179\cdots = 0.974482\cdots$$

for the probability that he will get into at least one car accident per 365 consecutive days. Don't use Bernoulli's formula, because you'd have to do a sum over

$$x \in \{1, 2, 3, 4, \ldots, 364, 365\}$$

and no one wants to evaluate Bernoulli's formula 365 times.

9. Each server is up with probability 99%, so the probability that all 8 of them are up is

$$(0.99)^8 = 0.922744\cdots$$

The complement principle tells us that the probability of having one or more down is

$$1 - 0.922744\cdots = 0.0772553\cdots$$

10. One drilling will fail to strike oil with probability 95%. Then all 20 of them will fail with probability

$$(0.95)^{20} = 0.358485\cdots$$

This means that striking oil at least once will occur with probability

$$1 - 0.358485\cdots = 0.641514\cdots$$

11. If the chance of striking oil at least once is 95% or greater, that means that the chance of missing oil each and every time is 5% or less. Each drilling has a 95% chance of failure. We have

$$\begin{aligned} 0.05 &> 0.95^n \\ \log 0.05 &> \log(0.95^n) \\ \log 0.05 &> n \log 0.95 \\ \frac{\log 0.05}{\log 0.95} &> n \\ 58.4039 &> n \end{aligned}$$

Therefore, we need 59 or more drillings. We should see that 58 will not be enough. Let's check

- With 58 drillings, we have $0.95^{58} = 0.0510468\cdots$.
- With 59 drillings, we have $0.95^{59} = 0.0484945\cdots$.
- This seems to be correct!

12. Using the repetition formula

$$\begin{aligned} 0.5 &= 0.995^n \\ \log 0.5 &= \log(0.995^n) \\ \log 0.5 &= n \log 0.995 \\ \frac{\log 0.5}{\log 0.995} &= n \\ 138.282\cdots &= n \end{aligned}$$

We conclude that the average couple in the study experienced roughly 138 coital episodes per year. Note, the above result is not unrealistic. If a request is made each Friday, Saturday, and Sunday, as well as twice during the interior of the week, then that is $(5)(52) = 260$ requests, with each request granted with probability approximately $138/260 = 0.530769\cdots \approx 53\%$.

13. According to Sage:

```
0 successes and 10 failures occurs with probability 0.00604661760000000
1 successes and 9 failures occurs with probability 0.0403107840000000
2 successes and 8 failures occurs with probability 0.120932352000000
3 successes and 7 failures occurs with probability 0.214990848000000
4 successes and 6 failures occurs with probability 0.250822656000000
5 successes and 5 failures occurs with probability 0.200658124800000
6 successes and 4 failures occurs with probability 0.111476736000000
7 successes and 3 failures occurs with probability 0.0424673280000000
8 successes and 2 failures occurs with probability 0.0106168320000000
9 successes and 1 failures occurs with probability 0.00157286400000000
10 successes and 0 failures occurs with probability 0.000104857600000000
Grand Total: 1.00000000000000
```

He's most likely to get 30%, 40%, or 50%, with a chance of 20% or 60%.

14. According to Sage,

```
0 successes and 8 failures occurs with probability 1.00000000000001e-16
1 successes and 7 failures occurs with probability 7.92000000000005e-14
2 successes and 6 failures occurs with probability 2.74428000000001e-11
3 successes and 5 failures occurs with probability 5.43367440000002e-9
4 successes and 4 failures occurs with probability 6.72417207000002e-7
5 successes and 3 failures occurs with probability 0.000053255442794400 1
6 successes and 2 failures occurs with probability 0.00263614441832280
```

```
7 successes and 1 failures occurs with probability 0.0745652278325593
8 successes and 0 failures occurs with probability 0.922744694427920
Grand Total: 1.00000000000000
```

Notice that the 8 success and 0 failures matches our previous result of $0.922744\cdots$.

15. According to Sage,

```
0 successes and 59 failures occurs with probability 0.0484945252494231
1 successes and 58 failures occurs with probability 0.150588262616630
2 successes and 57 failures occurs with probability 0.229845242941172
3 successes and 56 failures occurs with probability 0.229845242941172
4 successes and 55 failures occurs with probability 0.169359652693495
5 successes and 54 failures occurs with probability 0.0980503252436023
6 successes and 53 failures occurs with probability 0.0464448909048643
7 successes and 52 failures occurs with probability 0.0185081144207354
8 successes and 51 failures occurs with probability 0.00633172335446211
9 successes and 50 failures occurs with probability 0.00188840871975186
10 successes and 49 failures occurs with probability 0.000496949663092594
11 successes and 48 failures occurs with probability 0.000116509729624579
12 successes and 47 failures occurs with probability 0.0000245283641314904
13 successes and 46 failures occurs with probability 4.66734054323907e-6
14 successes and 45 failures occurs with probability 8.07134078906004e-7
15 successes and 44 failures occurs with probability 1.27442222985159e-7
```

As you can see, getting 15 strikes has probability around 1 in 8 million, so we don't really need to continue our tabulations past 15 strikes.

16. Here is the Bernoulli Table for three altimeters

```
Probability 0 successes and 3 failures: 1.95312500000000e-9
Probability 1 successes and 2 failures: 4.68164062500000e-6
Probability 2 successes and 1 failures: 0.00374063085937500
Probability 3 successes and 0 failures: 0.996254685546875
```

The probability of having an emergency (because 2 or 3 out of the 3 altimeters are broken) is given by

$$1.95312500000000 \times 10^{-9} + 4.68164062500000 \times 10^{-6} = 4.68359375 \times 10^{-6}$$

17. Here is the Bernoulli Table for five altimeters

```
Probability 0 successes and 5 failures: 3.05175781250000e-15
Probability 1 successes and 4 failures: 1.21917724609375e-11
Probability 2 successes and 3 failures: 1.94824523925781e-8
Probability 3 successes and 2 failures: 0.0000155664794616699
Probability 4 successes and 1 failures: 0.00621880854493713
Probability 5 successes and 0 failures: 0.993765605480954
```

Note, you should check the answer by seeing that it all adds to 1.

18. If I had given you a table with only one stain missing, the problem would be far too easy. That's because all the probabilities add to 1. If you were to add up all the given probabilities, then the missing entry must be one minus that sum. This would not test your knowledge of Bernoulli's formula.