Chapter 5

Probability

In Chapter 2, I said that a probability is a frequency expressed as a fraction of the sample size. That's one definition of probability, but it's not the only one. In fact, the meaning of probability is a topic of some controversy.

We'll start with the uncontroversial parts and work our way up. There is general agreement that a probability is a real value between 0 and 1 that is intended to be a quantitative measure corresponding to the qualitative notion that some things are more likely than others.

The "things" we assign probabilities to are called **events**. If E represents an event, then $P(E)$ represents the probability that E will occur. A situation where E might or might not happen is called a **trial**.

As an example, suppose you have a standard six-sided die[1] and want to know the probability of rolling a 6. Each roll is a trial. Each time a 6 appears is considered a **success**; other trials are considered **failures**. These terms are used even in scenarios where "success" is bad and "failure" is good.

If we have a finite sample of n trials and we observe s successes, the probability of success is s/n. If the set of trials is infinite, defining probabilities is a little trickier, but most people are willing to accept probabilistic claims about a hypothetical series of identical trials, like tossing a coin or rolling a die.

We start to run into trouble when we talk about probabilities of unique events. For example, we might like to know the probability that a candidate will win an election. But every election is unique, so there is no series of identical trials to consider.

[1] "Die" is the singular of "dice".

In cases like this some people would say that the notion of probability does not apply. This position is sometimes called **frequentism** because it defines probability in terms of frequencies. If there is no set of identical trials, there is no probability.

Frequentism is philosophically safe, but frustrating because it limits the scope of probability to physical systems that are either random (like atomic decay) or so unpredictable that we model them as random (like a tumbling die). Anything involving people is pretty much off the table.

An alternative is **Bayesianism**, which defines probability as a degree of belief that an event will occur. By this definition, the notion of probability can be applied in almost any circumstance. One difficulty with Bayesian probability is that it depends on a person's state of knowledge; people with different information might have different degrees of belief about the same event. For this reason, many people think that Bayesian probabilities are more subjective than frequency probabilities.

As an example, what is the probability that Thaksin Shinawatra is the Prime Minister of Thailand? A frequentist would say that there is no probability for this event because there is no set of trials. Thaksin either is, or is not, the PM; it's not a question of probability.

In contrast, a Bayesian would be willing to assign a probability to this event based on his or her state of knowledge. For example, if you remember that there was a coup in Thailand in 2006, and you are pretty sure Thaksin was the PM who was ousted, you might assign a probability like 0.1, which acknowledges the possibility that your recollection is incorrect, or that Thaksin has been reinstated.

If you consult Wikipedia, you will learn that Thaksin is not the PM of Thailand (at the time I am writing). Based on this information, you might revise your probability estimate to 0.01, reflecting the possibility that Wikipedia is wrong.

5.1 Rules of probability

For frequency probabilities, we can derive rules that relate probabilities of different events. Probably the best known of these rules is

$P(A \text{ and } B) = P(A)\ P(B)$ Warning: not always true!

where $P(A \text{ and } B)$ is the probability that events A and B both occur. This formula is easy to remember; the only problem is that it is *not always true*.

This formula only applies if A and B are **independent**, which means that if I know A occurred, that doesn't change the probability of B, and vice versa.

For example, if A is tossing a coin and getting heads, and B is rolling a die and getting 1, A and B are independent, because the coin toss doesn't tell me anything about the die roll.

But if I roll two dice, and A is getting at least one six, and B is getting two sixes, A and B are not independent, because if I know that A occurred, the probability of B is higher, and if I know B occurred, the probability of A is 1.

When A and B are not independent, it is often useful to compute the conditional probability, $P(A \mid B)$, which is the probability of A given that we know B occurred:

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

From that we can derive the general relation

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

This might not be as easy to remember, but if you translate it into English it should make sense: "The chance of both things happening is the chance that the first one happens, and then the second one given the first."

There is nothing special about the order of events, so we could also write

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

These relationships hold whether A and B are independent or not. If they are independent, then $P(A \mid B) = P(A)$, which gets us back where we started.

Because all probabilities are in the range 0 to 1, it is easy to show that

$$P(A \text{ and } B) \leq P(A)$$

To picture this, imagine a club that only admits people who satisfy some requirement, A. Now suppose they add a new requirement for membership, B. It seems obvious that the club will get smaller, or stay the same if it happens that all the members satisfy B. But there are some scenarios where people are surprisingly bad at this kind of analysis. For examples and discussion of this phenomenon, see `http://wikipedia.org/wiki/Conjunction_fallacy`.

Exercise 5.1 If I roll two dice and the total is 8, what is the chance that one of the dice is a 6?

Exercise 5.2 If I roll 100 dice, what is the chance of getting all sixes? What is the chance of getting no sixes?

Exercise 5.3 The following questions are adapted from Mlodinow, *The Drunkard's Walk*.

1. If a family has two children, what is the chance that they have two girls?

2. If a family has two children and we know that at least one of them is a girl, what is the chance that they have two girls?

3. If a family has two children and we know that the older one is a girl, what is the chance that they have two girls?

4. If a family has two children and we know that at least one of them is a girl named Florida, what is the chance that they have two girls?

You can assume that the probability that any child is a girl is 1/2, and that the children in a family are independent trials (in more ways than one). You can also assume that the percentage of girls named Florida is small.

5.2 Monty Hall

The Monty Hall problem might be the most contentious question in the history of probability. The scenario is simple, but the correct answer is so counter-intuitive that many people just can't accept it, and many smart people have embarrassed themselves not just by getting it wrong but by arguing the wrong side, aggressively, in public.

Monty Hall was the original host of the game show *Let's Make a Deal*. The Monty Hall problem is based on one of the regular games on the show. If you are on the show, here's what happens:

- Monty shows you three closed doors and tells you that there is a prize behind each door: one prize is a car, the other two are less valuable prizes like peanut butter and fake finger nails. The prizes are arranged at random.

- The object of the game is to guess which door has the car. If you guess right, you get to keep the car.

- So you pick a door, which we will call Door A. We'll call the other doors B and C.
- Before opening the door you chose, Monty likes to increase the suspense by opening either Door B or C, whichever does not have the car. (If the car is actually behind Door A, Monty can safely open B or C, so he chooses one at random).
- Then Monty offers you the option to stick with your original choice or switch to the one remaining unopened door.

The question is, should you "stick" or "switch" or does it make no difference?

Most people have the strong intuition that it makes no difference. There are two doors left, they reason, so the chance that the car is behind Door A is 50%.

But that is wrong. In fact, the chance of winning if you stick with Door A is only 1/3; if you switch, your chances are 2/3. I will explain why, but I don't expect you to believe me.

The key is to realize that there are three possible scenarios: the car is behind Door A, B or C. Since the prizes are arranged at random, the probability of each scenario is 1/3.

If your strategy is to stick with Door A, then you will win only in Scenario A, which has probability 1/3.

If your strategy is to switch, you will win in either Scenario B or Scenario C, so the total probability of winning is 2/3.

If you are not completely convinced by this argument, you are in good company. When a friend presented this solution to Paul Erdős, he replied, "No, that is impossible. It should make no difference.[2]"

No amount of argument could convince him. In the end, it took a computer simulation to bring him around.

Exercise 5.4 Write a program that simulates the Monty Hall problem and use it to estimate the probability of winning if you stick and if you switch.

Then read the discussion of the problem at `http://wikipedia.org/wiki/Monty_Hall_problem`.

[2] See Hoffman, *The Man Who Loved Only Numbers*, page 83.

Which do you find more convincing, the simulation or the arguments, and why?

Exercise 5.5 To understand the Monty Hall problem, it is important to realize that by deciding which door to open, Monty is giving you information. To see why this matters, imagine the case where Monty doesn't know where the prizes are, so he chooses Door B or C at random.

If he opens the door with the car, the game is over, you lose, and you don't get to choose whether to switch or stick.

Otherwise, are you better off switching or sticking?

5.3 Poincaré

Henri Poincaré was a French mathematician who taught at the Sorbonne around 1900. The following anecdote about him is probably fabricated, but it makes an interesting probability problem.

Supposedly Poincaré suspected that his local bakery was selling loaves of bread that were lighter than the advertised weight of 1 kg, so every day for a year he bought a loaf of bread, brought it home and weighed it. At the end of the year, he plotted the distribution of his measurements and showed that it fit a normal distribution with mean 950 g and standard deviation 50 g. He brought this evidence to the bread police, who gave the baker a warning.

For the next year, Poincaré continued the practice of weighing his bread every day. At the end of the year, he found that the average weight was 1000 g, just as it should be, but again he complained to the bread police, and this time they fined the baker.

Why? Because the shape of the distribution was asymmetric. Unlike the normal distribution, it was skewed to the right, which is consistent with the hypothesis that the baker was still making 950 g loaves, but deliberately giving Poincaré the heavier ones.

Exercise 5.6 Write a program that simulates a baker who chooses n loaves from a distribution with mean 950 g and standard deviation 50 g, and gives the heaviest one to Poincaré. What value of n yields a distribution with mean 1000 g? What is the standard deviation?

Compare this distribution to a normal distribution with the same mean and the same standard deviation. Is the difference in the shape of the distribution big enough to convince the bread police?

Exercise 5.7 If you go to a dance where partners are paired up randomly, what percentage of opposite sex couples will you see where the woman is taller than the man?

In the BRFSS (see Section 4.5), the distribution of heights is roughly normal with parameters $\mu = 178$ cm and $\sigma^2 = 59.4$ cm for men, and $\mu = 163$ cm and $\sigma^2 = 52.8$ cm for women.

As an aside, you might notice that the standard deviation for men is higher and wonder whether men's heights are more variable. To compare variability between groups, it is useful to compute the **coefficient of variation**, which is the standard deviation as a fraction of the mean, σ/μ. By this measure, women's heights are slightly more variable.

5.4 Another rule of probability

If two events are **mutually exclusive**, that means that only one of them can happen, so the conditional probabilities are 0:

$$P(A \mid B) = P(B \mid A) = 0$$

In this case it is easy to compute the probability of either event:

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{Warning: not always true.}$$

But remember that this only applies if the events are mutually exclusive. In general the probability of A or B or both is:

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

The reason we have to subtract off $P(A \text{ and } B)$ is that otherwise it gets counted twice. For example, if I flip two coins, the chance of getting at least one tails is $1/2 + 1/2 - 1/4$. I have to subtract $1/4$ because otherwise I am counting heads-heads twice. The problem becomes even clearer if I toss three coins.

Exercise 5.8 If I roll two dice, what is the chance of rolling at least one 6?

Exercise 5.9 What is the general formula for the probability of A or B but not both?

5.5 Binomial distribution

If I roll 100 dice, the chance of getting all sixes is $(1/6)^{100}$. And the chance of getting no sixes is $(5/6)^{100}$.

Those cases are easy, but more generally, we might like to know the chance of getting k sixes, for all values of k from 0 to 100. The answer is the **binomial distribution**, which has this PMF:

$$\text{PMF}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where n is the number of trials, p is the probability of success, and k is the number of successes.

The **binomial coefficient** is pronounced "n choose k", and it can be computed directly like this:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Or recursively like this

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

with two base cases: if $n = 0$ the result is 0; if $k = 0$ the result is 1. If you download `http://thinkstats.com/thinkstats.py` you will see a function named `Binom` that computes the binomial coefficient with reasonable efficiency.

Exercise 5.10 If you flip a coin 100 times, you expect about 50 heads, but what is the probability of getting exactly 50 heads?

5.6 Streaks and hot spots

People do not have very good intuition for random processes. If you ask people to generate "random" numbers, they tend to generate sequences that are random-looking, but actually more ordered than real random sequences. Conversely, if you show them a real random sequence, they tend to see patterns where there are none.

An example of the second phenomenon is that many people believe in "streaks" in sports: a player that has been successful recently is said to have a "hot hand;" a player that has been unsuccessful is "in a slump."

Statisticians have tested these hypotheses in a number of sports, and the consistent result is that there is no such thing as a streak[3]. If you assume that each attempt is independent of previous attempts, you will see occasional long strings of successes or failures. These apparent streaks are not sufficient evidence that there is any relationship between successive attempts.

A related phenomenon is the clustering illusion, which is the tendency to see clusters in spatial patterns that are actually random (see `http://wikipedia.org/wiki/Clustering_illusion`).

To test whether an apparent cluster is likely to be meaningful, we can simulate the behavior of a random system to see whether it is likely to produce a similar cluster. This process is called **Monte Carlo** simulation because generating random numbers is reminiscent of casino games (and Monte Carlo is famous for its casino).

Exercise 5.11 If there are 10 players in a basketball game and each one takes 15 shots during the course of the game, and each shot has a 50% probability of going in, what is the probability that you will see, in a given game, at least one player who hits 10 shots in a row? If you watch a season of 82 games, what are the chances you will see at least one streak of 10 hits or misses?

This problem demonstrates some strengths and weaknesses of Monte Carlo simulation. A strength is that it is often easy and fast to write a simulation, and no great knowledge of probability is required. A weakness is that estimating the probability of rare events can take a long time! A little bit of analysis can save a lot of computing.

Exercise 5.12 In 1941 Joe DiMaggio got at least one hit in 56 consecutive games[4]. Many baseball fans consider this streak the greatest achievement in any sport in history, because it was so unlikely.

Use a Monte Carlo simulation to estimate the probability that any player in major league baseball will have a hitting streak of 57 or more games in the next century.

Exercise 5.13 A cancer cluster is defined by the Centers for Disease Control (CDC) as "greater-than-expected number of cancer cases that occurs within a group of people in a geographic area over a period of time.[5]"

[3]For example, see Gilovich, Vallone and Tversky, "The hot hand in basketball: On the misperception of random sequences," 1985.

[4]See `http://wikipedia.org/wiki/Hitting_streak`.

[5]From `http://cdc.gov/nceh/clusters/about.htm`.

Many people interpret a cancer cluster as evidence of an environmental hazard, but many scientists and statisticians think that investigating cancer clusters is a waste of time[6]. Why? One reason (among several) is that identifying cancer clusters is a classic case of the Sharpshooter Fallacy (see `http://wikipedia.org/wiki/Texas_sharpshooter_fallacy`).

Nevertheless, when someone reports a cancer cluster, the CDC is obligated to investigate. According to their web page:

> "Investigators develop a 'case' definition, a time period of concern, and the population at risk. They then calculate the expected number of cases and compare them to the observed number. A cluster is confirmed when the observed/expected ratio is greater than 1.0, and the difference is statistically significant."

1. Suppose that a particular cancer has an incidence of 1 case per thousand people per year. If you follow a particular cohort of 100 people for 10 years, you would expect to see about 1 case. If you saw two cases, that would not be very surprising, but more than than two would be rare.

   Write a program that simulates a large number of cohorts over a 10 year period and estimates the distribution of total cases.

2. An observation is considered statistically significant if its probability by chance alone, called a p-value, is less than 5%. In a cohort of 100 people over 10 years, how many cases would you have to see to meet this criterion?

3. Now imagine that you divide a population of 10000 people into 100 cohorts and follow them for 10 years. What is the chance that at least one of the cohorts will have a "statistically significant" cluster? What if we require a p-value of 1%?

4. Now imagine that you arrange 10000 people in a 100 $\times$100 grid and follow them for 10 years. What is the chance that there will be at least one 10 $\times$10 block anywhere in the grid with a statistically significant cluster?

5. Finally, imagine that you follow a grid of 10000 people for 30 years. What is the chance that there will be a 10-year interval at some point with a 10 $\times$10 block anywhere in the grid with a statistically significant cluster?

[6] See Gawande, "The Cancer Cluster Myth," *New Yorker*, Feb 8, 1997.

5.7 Bayes's theorem

Bayes's theorem is a relationship between the conditional probabilities of two events. A conditional probability, often written $P(A \mid B)$ is the probability that Event Awill occur given that we know that Event Bhas occurred. Bayes's theorem states:

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

To see that this is true, it helps to write $P(A \text{ and } B)$, which is the probability that A and B occur

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

But it is also true that

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

So

$$P(B)\ P(A \mid B) = P(A)\ P(B \mid A)$$

Dividing through by $P(B)$ yields Bayes's theorem[7].

Bayes's theorem is often interpreted as a statement about how a body of evidence, E, affects the probability of a hypothesis, H:

$$P(H|E) = P(H)\frac{P(E|H)}{P(E)}$$

In words, this equation says that the probability of H after you have seen E is the product of $P(H)$, which is the probability of H before you saw the evidence, and the ratio of $P(E \mid H)$, the probability of seeing the evidence assuming that H is true, and $P(E)$, the probability of seeing the evidence under any circumstances (H true or not).

This way of reading Bayes's theorem is called the "diachronic" interpretation because it describes how the probability of a hypothesis gets **updated** over time, usually in light of new evidence. In this context, $P(H)$ is called the **prior** probability and $P(H \mid E)$ is called the **posterior**. $P(E \mid H)$ is the **likelihood** of the evidence, and $P(E)$ is the **normalizing constant**.

[7] See `http://wikipedia.org/wiki/Q.E.D.`!

A classic use of Bayes's theorem is the interpretation of clinical tests. For example, routine testing for illegal drug use is increasingly common in workplaces and schools (See `http://aclu.org/drugpolicy/testing`.). The companies that perform these tests maintain that the tests are sensitive, which means that they are likely to produce a positive result if there are drugs (or metabolites) in a sample, and specific, which means that they are likely to yield a negative result if there are no drugs.

Studies from the Journal of the American Medical Association[8] estimate that the sensitivity of common drug tests is about 60% and the specificity is about 99%.

Now suppose these tests are applied to a workforce where the actual rate of drug use is 5%. Of the employees who test positive, how many of them actually use drugs?

In Bayesian terms, we want to compute the probability of drug use given a positive test, $P(D \mid E)$. By Bayes's theorem:

$$P(D|E) = P(D)\frac{P(E|D)}{P(E)}$$

The prior, $P(D)$ is the probability of drug use before we see the outcome of the test, which is 5%. The likelihood, $P(E \mid D)$, is the probability of a positive test assuming drug use, which is the sensitivity.

The normalizing constant, $P(E)$ is a little harder to evaluate. We have to consider two possibilities, $P(E \mid D)$ and $P(E \mid N)$, where N is the hypothesis that the subject of the test does not use drugs:

$$P(E) = P(D)\ P(E \mid D) + P(N)\ P(E \mid N)$$

The probability of a false positive, $P(E \mid N)$, is the complement of the specificity, or 1%.

Putting it together, we have

$$P(D|E) = \frac{P(D)P(E|D)}{P(D)P(E|D) + P(N)P(E|N)}$$

Plugging in the given values yields $P(D \mid E) = 0.76$, which means that of the people who test positive, about 1 in 4 is innocent.

[8] I got these numbers from Gleason and Barnum, "Predictive Probabilities In Employee Drug-Testing," at `http://piercelaw.edu/risk/vol2/winter/gleason.htm`.

Exercise 5.14 Write a program that takes the actual rate of drug use, and the sensitivity and specificity of the test, and uses Bayes's theorem to compute $P(D \mid E)$.

Suppose the same test is applied to a population where the actual rate of drug use is 1%. What is the probability that someone who tests positive is actually a drug user?

Exercise 5.15 This exercise is from `http://wikipedia.org/wiki/Bayesian_inference`.

> "Suppose there are two full bowls of cookies. Bowl 1 has 10 chocolate chip and 30 plain cookies, while Bowl 2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. The cookie turns out to be a plain one. How probable is it that Fred picked it out of Bowl 1?"

Exercise 5.16 The blue M&M was introduced in 1995. Before then, the color mix in a bag of plain M&Ms was (30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan). Afterward it was (24% Blue , 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown).

A friend of mine has two bags of M&Ms, and he tells me that one is from 1994 and one from 1996. He won't tell me which is which, but he gives me one M&M from each bag. One is yellow and one is green. What is the probability that the yellow M&M came from the 1994 bag?

Exercise 5.17 This exercise is adapted from MacKay, *Information Theory, Inference, and Learning Algorithms*:

Elvis Presley had a twin brother who died at birth. According to the Wikipedia article on twins:

> "Twins are estimated to be approximately 1.9% of the world population, with monozygotic twins making up 0.2% of the total—and 8% of all twins."

What is the probability that Elvis was an identical twin?

5.8 Glossary

event: Something that may or may not occur, with some probability.

trial: One in a series of occasions when an event might occur.

success: A trial in which an event occurs.

failure: A trail in which no event occurs.

frequentism: A strict interpretation of probability that only applies to a series of identical trials.

Bayesianism: A more general interpretation that uses probability to represent a subjective degree of belief.

independent: Two events are independent if the occurrence of one does has no effect on the probability of another.

coefficient of variation: A statistic that measures spread, normalized by central tendency, for comparison between distributions with different means.

Monte Carlo simulation: A method of computing probabilities by simulating random processes (see `http://wikipedia.org/wiki/Monte_Carlo_method`).

update: The process of using data to revise a probability.

prior: A probability before a Bayesian update.

posterior: A probability computed by a Bayesian update.

likelihood of the evidence: One of the terms in Bayes's theorem, the probability of the evidence conditioned on a hypothesis.

normalizing constant: The denominator of Bayes's Theorem, used to normalize the result to be a probability.