Contents

Chapter 1

More Probability

1.1 More Conditional Probability

This chapter continues the discussion of conditional chances from Chapter 13 of **FPP**.

Example 1. In a particular city 48% of the voters are Republicans and 52% are Democrats. And 78% of Democrats are in favor a particular referendum while only 44% of Republicans are in favor.

- (a). What percentage of voters is in favor of the referendum?
- (b). Among supporters of the referendum, what percentage are Democrats?

Solution.

- (a). There are four kinds of voters: (Dem, Favor), (Dem, Against), (Rep, Favor) and (Rep, Against). The percentage of voters that is (Dem, Favor) is 78% of 48% $\approx$ 37%. Likewise, the percentage of voters that is:
  - (Dem, Against) = 22% of 48% $\approx$ 11%,
  - (Rep, Favor) = 44% of 52% $\approx$ 23%,
  - (Rep, Against) = 56% of 52% $\approx$ 29%.

  So the percentage of voters in favor of the referendum is about 37% + 23% $\approx$ 60%.
- (b). About 60% of voters support the referendum: 37% Democrats and 23% Republicans. 37% is what percent of 60%? $37/60 = x/100$; $x = 37 \times 100/60 \approx 62\%$. That's the answer: about 62% of the Favor's are Democrats.

Example 1, part b can be recast as a question about a conditional chance. (See **FPP** pages 226-229.) Randomly select a voter. What is the conditional chance that he or she is a Democrat given that he or she is in favor of the referendum? In other words, what is P(Dem|Favor)? The numerical calculations will be just the same as in the previous solution but we will think of them as chances for random sampling instead of percentages of voters.

The *Multiplication Rule* (**FPP**, pg. 229) says

$$\text{P(Dem, Favor)} = \text{P(Favor)} \times \text{P(Dem|Favor)}.$$

So

$$\text{P(Dem | Favor)} = \text{P(Dem, Favor)/P(Favor)}$$

The numerator is

$$\text{P(Dem, Favor)} = \text{P(Dem)} \times \text{P(Favor|Dem)} \approx 48\% \times 78\% \approx 37\%.$$

And the denominator is

$$\text{P(Favor)} = \text{P(Dem,Favor)} + \text{P(Rep,Favor)} \approx 37\% + 23\% \approx 60\%.$$

So P(Dem|Favor) is about 37% / 60%, or about 62%.

Example 1 illustrates a common statistical problem. We know an unconditional chance for event A — party affiliation — and a conditional chance for event B — referendum support — given A. We want to find the conditional chance of A given B — party affiliation given referendum support. In symbols, we know P(A) and P(B|A); we want to find P(A|B). In symbols the solution is

$$\begin{aligned} P(A|B) &= P(A \text{ and } B)/P(B) \\ &= P(A \text{ and } B)/P(A \text{ and } B) + P(\text{not } A \text{ and } B) \\ &= P(A) \times P(B|A)/(P(A) \times P(B|A) + P(\text{not } A) \times P(B|\text{not } A)) \end{aligned}$$

And the final line is in terms of things we know.

Because this sort of problem is so important we examine it in more detail.

The first equality is the *Multiplication Rule*. The chance that both A and B happen equals the chance of B times the conditional chance of A given B. The chance that a randomly selected voter is (Dem, Favor) equals the chance that the voter is a Democrat times the conditional chance that she favors the referendum given that she is a Democrat.

The second equality is just listing all the ways B can happen — B can happen either with A or with its opposite — and adding the chances. The chance that the voter is for the referendum is the chance that she is (Dem, Favor) plus the chance that she is (Rep, Favor).

The third equality is the *Multiplication Rule* again. P(A and B) equals P(A) times P(B) given A; and P(not A and B) equals P(not A) times P(B) given not A. The chance that the voter is (Dem, Favor) equals the chance she is a Democrat times the chance she is in favor given that she is a Democrat; and the chance she is (Rep, Favor) equals the chance she is a Republican times the chance she is in favor given that she is a Republican.

And the result is in terms of things we know:

- P(A) (chance of *Democrat*, 48%),
- P(not A) (chance of *Republican*, 52%),
- P(B|A) (chance of *Favor* given *Democrat*, 78%), and
- P(B| not A) (chance of *Favor* given *Republican*, 44%).

Bayes' Rule.

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\text{not } A)P(B|\text{not } A)}$$

If you want to know $P(A|B)$, check to see whether you know $P(A)$, $P(B|A)$ and $P(B|\text{not } A)$. If you do, use Bayes' rule.

Example 2. (Hypothetical) Suppose the incidence (rate in the population at a point in time) of AIDS is 0.2%, or 2 in 1000 people and that a test to diagnose AIDS will correctly diagnose it in 97% of the people who have AIDS and correctly diagnose its absence in 99% of the people who do not have AIDS. If a randomly chosen person tests positive, what is the chance he has AIDS?

Solution. We want the chance that a randomly selected person has AIDS, given that he tests positive. Bayes' rule says the chance is

$$\frac{P(\text{AIDS}) \times P(\text{test positive}|\text{AIDS})}{P(\text{AIDS}) \times P(\text{test positive}|\text{AIDS}) + P(\text{not AIDS}) \times P(\text{test positive}|\text{not AIDS})},$$

which is

$$\frac{0.2\% \times 97\%}{0.2\% \times 97\% + 99.8\% \times 1\%},$$

or about 16%.

It may seem surprising that with such an accurate test the chance is only about 16% that a person who tests positive actually has AIDS, but

it's correct. Among 100,000 people, about 200 will have AIDS and about 194 of them will test positive. The other approximately 99,800 people will not have AIDS, and about 998 will test positive. That's about 194 + 998 = 1192 people who test positive. Among the positives, only 194, or about 16%, really have AIDS.

1.2 Probabilities for Degrees of Belief

If it is not known, it has probability.

FPP presents the *frequency theory* of probability. (See page 221 for explanation.) In addition to the frequency theory, this supplement also uses probability to represent *degrees of belief*. (See **Statistics: A Bayesian Perspective** by Don Berry for an introductory book based solely on the degrees-of-belief interpretation of probability.) It is often necessary for us to assess a degree of belief about an event which is not repeatable, and which therefore would not be covered by the frequency theory of probability.

For example, consider the space shuttle Challenger. The shuttle's booster rockets were constructed in four sections. The joints between the sections were sealed with O-rings. On the night of January 27, 1986 the Challenger was on the launch pad, scheduled for a morning flight. The temperature was 31°F, much colder than any previous launch, and the NASA engineers did not know whether the O-rings would function properly in such cold weather. With only a bit of simplification, we may say there were two possible decisions — either launch now or wait for warmer weather. And, if they launched now, there were two possible outcomes — either success or disaster. In order to decide whether to launch now, the NASA decision makers had to estimate the chances of success and disaster. That is, they had to assess their degree of belief in whether the O-rings would work properly.

There is no public record of their assessment of the chances. We believe they made the assessment informally and approximately. Later, several statisticians tried to make the assessment more formally and accurately. Their work is described in *Chance* magazine. (Dalal, Siddhartha, R., Fowlkes, Edward B., and Hoadley, Bruce, "Lesson Learned from Challenger: A Statistical Perspective." Chance, 3(2), 1989.). They estimated the chance of failure at 30° F, as about 16%.

Here are two simpler examples that illustrates the main features.

Example 1. Omar tosses a coin, catches it and looks at it; but does not show it to Raquel. For Raquel, the chance of Tails on the coin is 50%. But Omar has seen the coin and knows that it landed Tails. For Omar the chance of Tails is 100%.

There are two things to notice. First, Raquel assesses a 50% chance for an event that is not random. Once the coin has been tossed, its outcome is no longer a chance event. But for Raquel, a 50% chance of tails still makes sense because she doesn't know the outcome. The chance doesn't represent the event. It represents Raquel's belief about the event.

Second, two people can assess different chances for the same event. Omar and Raquel assess different chances for the coin toss because Omar has seen the result and Raquel hasn't. Her chance represents her belief and Omar's represents his. Omar and Raquel assess different chances because they have different information and different beliefs. And their different chances represent those differences.

Example 2. "Is the Defense Department's share of the national budget greater than or less than 30%?" Joe doesn't know the answer but thinks the DOD budget share is right around 30%. He would be willing to bet even money either way. For Joe the chance is about 50% for each alternative. But Flavia recalls reading that the DOD budget share is under 20% and is fairly certain of her recall. She assess the chance for greater than 30% as about 3% and for less than 30% as about 97%.

Of course, the truth could be ascertained by looking it up. And then Joe and Flavia would want to change their box models. But, until they do look it up, Joe and Flavia have different information and different box models.

And here's an example where the chances aren't so clear.

Example 3. Bettor Z is offered a bet on contestant A in a tennis match at even odds. Z will accept the bet if he believes that A has at least a 50% chance of winning. The tennis match is not a chance event. It will be determined by skill, strength and state of mind, among other things. And it is not an event that can be repeated many times to see how often A wins. It is a one shot, non-random occurence. Yet the "chance that A wins" is still a useful concept, at least for Z who must assess whether it is greater than 50%. Of course Bettor Y, who offered the bet at even odds, may have different beliefs and assess the chances differently.

1.3 Assessing Probabilities for Beliefs

It is sometimes useful to use a box model to represent a person's beliefs. How is the box model constructed? One way is to think carefully in a structured way about one's beliefs. In **Statistics: A Bayesian Perspective** Don Berry describes this process.

> To measure one's degree of belief requires a scale, just like any other measurement. For degrees of belief the scale is a

calibration experiment. The assessor — let's take it to be you — must be able to imagine an experiment with equally likely outcomes. To decide whether outcomes are equally likely, suppose you get to choose any one of the possible outcomes. I promise to pay you \$100 should the experiment result in the outcome you choose. Outcomes in the set are *equally likely for you* if you are indifferent among them. ....

For example, suppose I offer to pay you \$100 if you call the roll of a six-sided die correctly. If you are indifferent as to which one you call, then the six sides are equally likely for you.

There are many candidates for calibration experiments. .... One possibility is ... selecting a chip from a bowl that contains chips of the same size and shape.

... I will use *chip-from-bowl experiments* to calibrate. ... the chips are equally likely, each having probability 1 divided by the number of chips in the bowl.

Consider a specific setting. I would like to know your probability that average adult male emperor penguins weigh more than 50 lb — call this event A. Since I cannot obtain answers to my questions directly from you, I will guess them. My guesses will be wrong for many readers. Some of you know more about penguins than do others, and some of you may even be penguin experts. You should follow along in any case, but use your actual answers to modify what I say in a way that I hope will be clear to you.

A bowl contains a green chip and a red chip .... I offer you the choice of getting \$100 if a chip selected from the bowl is green and \$100 if A is true (that is, if average adult male emperor penguins do weight more than 50 lb.). If you choose to select from the bowl and the chip is red or if you choose A and it turns out that A is not true, then you receive nothing. You say you prefer A. Since you chose A over an event of probability $\frac{1}{2}$, I interpret this as saying that your probability of A is at least $\frac{1}{2}$.

Consider a new bowl, ...: three green chips and one red chip. Again, you may choose between \$100 if a chip selected from the bowl is green and \$100 if A is true. Now you prefer the chip. Taken together, your two answers mean that $\frac{1}{2} \leq P(A) \leq \frac{3}{4}$. .... We proceed in this way, ..., until we know $P(A)$ sufficiently accurately.

1.4 Exercises

1. Suppose 43% of a population favors a referendum. Of those that favor the referendum, 74% are Democrats. Of those that do not favor the referendum, 27% are Democrats. What is the chance that a randomly selected voter is a Democrat? What is the chance that a randomly selected voter is a Democrat who favors the referendum? Given that a person is Democrat, what is the chance he or she favors the referendum?

2. (Hypothetical) The advertisers of a home pregnancy test say that based on their clinical trials on a large random sample of women, the test is 99% accurate. Assume the following: a) if a women is pregnant the test accurately reports she is pregnant 99.5% of the time b) if a woman is not pregnant the test accurately reports that she is not pregnant 98.5% of the time, c) 5% of all women that buy a home pregnancy test are pregnant. If a woman buys a test kit, and the test is positive, what are the chances she is pregnant? If a woman buys a test kit, and the test is negative, what are the chances she is pregnant?

3. Design a chip-from-bowl experiment to assess several friend's beliefs that the earth has been visited by extraterrestrial life.

Chapter 2

Which Box Model?

2.1 Statistical Hypotheses

Data are arriving according to a chance process. We don't know what the correct box model is. To learn about the chance process we look for box models that do a good job of explaining the data. There are two steps:

- Make a list of possible box models;
- See how well each box model explains the data.

A *statistical hypothesis* is a box model. The hypothesis says two things. It says the data are like draws from a box. And it says what tickets are in the box.

Example 1. **Thumbtack.** Problem 1 of Exercise Set B on page 449 of **FPP** says "A thumbtack is thrown in the air. It lands either point up or point down." One person proposes a box model with two tickets — one U and one D. That's a simple hypothesis. Another person proposes a box model with three tickets — one U and two D's. That's another simple hypothesis. The first hypothesis is correct if the chance of U is equal to 50%; the second is correct if the chance of U is equal to 33%. We don't know whether either hypothesis is correct.

Example 2. **Psychic.** A card is dealt from a standard deck. A psychic claims to be able to guess the suit with 80% accuracy. The psychic's box model has four 1's and one 0. A skeptic's box model has one 1 and three 0's. These are two simple hypotheses.

Where do statistical hypotheses come from? Good hypotheses represent sensible statements about the chance process generating the data. In Examples 1 and 2 the data are classifying and counting either Ups and Downs

or Corrects and Incorrects. **FPP**, page 301 says "If you have to classify and count the draws, put 0's and 1's on the tickets." What's unknown is the chance of success. So statistical hypotheses are boxes with different percentages of 0's and 1's, representing different chances of success.

Example 3. **The Slater School.** An article written by Paul Brodeur (*The New Yorker*, ????) describes the Slater School, an elementary school in California where 8 out of 145 female teachers, about 5.5%, had contracted cancer of the reproductive system. The school was near some high tension power lines and there was concern among the teachers that the power lines were contributing to the high cancer rate.

Is contracting cancer a chance process? From a scientific perspective the answer is not clear. We don't know why some women get cancer and others don't. It might be chance or it might not. But from a statistical perspective we can imagine randomly selecting a woman from the population and observing whether she contracts cancer. What population? The population of women similar to the Slater teachers. Can we actually perform this experiment? No. The Slater teachers are not selected randomly.

A national cancer registry indicates that about 3% of women nationally, of an age typical of the Slater teachers, develop cancer of the reproductive system.

One statistical hypothesis is that female teachers at the Slater school develop cancer at the same rate as women nationally. According to this hypothesis, whether a particular female teacher develops cancer is like drawing from a box with three 1's and ninety-seven 0's. If this hypothesis were true, we would expect about $0.03 \times 145 \approx 4$ reproductive cancers to have developed among the 145 female teachers at Slater. The four "extra" cancers would just be due to chance variation.

Three other statistical hypotheses say that a particular female teacher at Slater developing a reproductive cancer is like drawing from a box with

- four 1's and ninety-six 0's or
- five 1's and ninety-five 0's or
- six 1's and ninety-four 0's.

Why these three hypotheses and not others? Because (a), learning the cancer rate to the nearest percentage point is accurate enough and (b), a cancer rate of 6% is twice the national average and it's unlikely that the underlying rate at Slater is higher than that. So these hypotheses will suffice, unless further investigation suggests otherwise.

Later in this supplement we will see how strongly the data support each of the four hypotheses.

In the three examples so far, the data have been like draws from a 0-1 box. They were classifying and counting the number of Ups and Downs for the thumbtacks, the number of rights and wrongs for the psychics, or the number of cancers for the Slater teachers. But not all data are like draws from a 0-1 box. Another important case is when the data follow the normal curve. In fact, **FPP** say "It is a remarkable fact that many histograms follow the normal curve." This can happen when either

- the data are like draws from Gauss' error box and the numbers in the error box follow the normal curve, or

- the data are a random sample from a population that follows the normal curve.

Example 4. **Weight Loss Drug.** A study evaluated the effectiveness of a weight loss drug. The quantity of interest is the average weight loss among people who use the drug. The weight of 1000 participants was taken before and after taking the drug for 8 weeks. Changes in weight followed the normal curve and ranged from +2 to -22. The average change was -10 and the SD of the sample was 3. Suppose that prior to the study, a previous 'pilot' study reported an average of -12 with an SE of 1. Based on these results your prior belief could be represented by a normal curve with an average of -12 and a SD of 1. This means that apriori, 2/3 of your probability was on values of the average between -17 to -7, 1/6 on values less than -17 and 1/6 on values greater than -7. Only about 1% of your apriori probability was on values of the average that reflects no change or weight gain.

What is your belief about the average weight loss after the study? Combining your prior belief with the data requires calculus, but calculus shows that after observing the data your belief will look like a normal curve. The new distribution will be somewhere between your priori belief and the average you observed in the sample. The variability in your prior beliefs and the variability in the sample will determine precisely where your new beliefs will fall. After you observe the data and apply Bayes Rule the distribution of the average weight loss will follow a normal curve. When the sample size is large, the normal curve will be centered around

$$\frac{\frac{1}{\text{prior SD}^2}}{\frac{1}{\text{prior SD}^2}+\frac{\text{n}}{\text{sample SD}^2}}\cdot\text{prior avg}+\frac{\frac{\text{n}}{\text{sample SD}^2}}{\frac{1}{\text{prior SD}^2}+\frac{\text{n}}{\text{sample SD}^2}}\cdot\text{sample avg.}$$

The normal curve will have an SD

$$\frac{1}{\sqrt{\frac{1}{\text{prior SD}^2} + \frac{\text{n}}{\text{sample SD}^2}}}.$$

If the sample size is small, adjustments must be made. In the weight loss example, the normal curve will be centerd around

$$\frac{\frac{1}{1^2}}{\frac{1}{1^2} + \frac{1000}{3^2}} \cdot \text{-}12 + \frac{\frac{1000}{3^2}}{\frac{1}{1^2} + \frac{1000}{3^2}} \cdot \text{-}10 = -10.02,$$

and the normal curve will have an SD of

$$\frac{1}{\sqrt{\frac{1}{1^2} + \frac{1000}{3^2}}} = .09.$$

After updating prior beliefs with the information provided by the data, the distribution is called the posterior distribution. Figure 2 shows prior and posterior distributions for the weight loss example.

Figure 2.1: **2. Prior and Posterior Distributions for Average Weight Loss**

2.2 Statistical Evidence

To learn about a chance process we see how well each statistical hypothesis explains the data. Each statistical hypothesis represents a different theory about the chance process and we evaluate theories by seeing how well they explain the data.

Example 1. **Thumbtack, continued.** Recall Example 1 of Chapter 2.1. There are two hypotheses about the chance a thumbtack lands Up when tossed: one says it's 50%, the other says 33.3%. We toss the thumbtack four times and get 2 Ups and 2 Downs. That seems to support the first hypothesis. But there are only four tosses so the evidence isn't very strong. Or is it? Can we say how strong the evidence is?

Use the binomial formula (**FPP** Chapter 15, Section 2) to calculate the chance of 2 Ups and 2 Downs. According to the first hypothesis it's

$$\frac{4!}{2!2!}0.5^2 0.5^2 = 0.375.$$

According to the second it's 0.296. So the first hypothesis does explain the data better. It's the ratio $0.375/0.296 \approx 1.3$ that matters. The first hypothesis explains the data about 1.3 times better than the second. The data support the first hypothesis over the second by a ratio of about 1.3 to 1. That's not much.

Example 2. **Psychic, continued.** Recall Example 2 of Chapter 2.1. There are two hypotheses about the chance a psyhic can guess the suit of a card randomly chosen from a standard deck: one says it's 80%, the other says 25%. The psychic makes 100 trials and guesses 30 correctly. How strong is the evidence for deciding between the two hypotheses?

The data are like the sum of 100 draws from a 0-1 box. The first hypothesis says the expected number correct is 80 and the SE is 4. 100 draws are a lot so the Normal approximation applies (**FPP**, Chap. 18). The chance of getting exactly 30 correct is about the area under the Normal curve from $(29.5 - 80)/4 \approx -12.6$ to $(30.5 - 80)/4 \approx -12.4$. That area is miniscule.

The second hypothesis says the expected number is 25 and the SE is about 4.33. So the chance of exactly 30 correct is like the area under the Normal curve from $(29.5 - 25)/4.33 \approx 1.04$ to $(30.5 - 25)/4.33 \approx 1.27$. The area is about 5%. It's the ratio that's important: 5% compared to almost 0%. The second hypothesis explains the data much better than the first. The data support the skeptic, not the psychic, by extremely large odds.

Example 3. **Slater school, continued.** Recall Example 3 of Chapter 2.1. 8 out of 145 women developed reproductive cancer. Four statistical hypotheses say the cancer rate is:

- (a) 3%,
- (b) 4%,
- (c) 5% and
- (d) 6%.

What does the evidence say? The Normal approximation applies.

- Hypothesis (a) says the expected number of cancers is 3% of 145 = 4.35 and the SE is $\sqrt{145 \times 0.03 \times 0.97} \approx 2$. So the chance of exactly 8 cancers is like the area under the Normal curve from $(7.5 - 4.35)/2 \approx 1.57$ to $(8.5 - 4.35)/2 \approx 2.07$. The area is about 4%.
- Hypothesis (b) says the expected number of cancers is 4% of 145 = 5.80 and the SE is $\sqrt{145 \times 0.04 \times 0.96} \approx 2.4$. So the chance of exactly 8 cancers is like the area under the Normal curve from $(7.5 - 5.80)/2.4 \approx 0.71$ to $(8.5 - 5.80)/2.4 \approx 1.13$. The area is about 11%.
- Hypothesis (c) says the expected number of cancers is 5% of 145 = 7.25 and the SE is $\sqrt{145 \times 0.05 \times 0.95} \approx 2.6$. So the chance of exactly 8 cancers is like the area under the Normal curve from $(7.5 - 7.25)/2.6 \approx 0.10$ to $(8.5 - 7.25)/2.6 \approx 0.48$. The area is about 14%.
- Hypothesis (d) says the expected number of cancers is 6% of 145 = 8.70 and the SE is $\sqrt{145 \times 0.06 \times 0.94} \approx 2.9$. So the chance of exactly 8 cancers is like the area under the Normal curve from $(7.5 - 8.70)/2.9 \approx -0.41$ to $(8.5 - 8.70)/2.9 \approx -0.07$. The area is about 13%.

The data support the four hypotheses in the ratio of about 4 to 11 to 14 to 13. That's moderate, but not overwhelming evidence against hypothesis (a).

Does this mean that high tension power lines are causing excess cancers among Slater teachers? No. The statistical analysis can only shed light on whether the cancer rate at Slater is higher than the national average. It does not tell us about possible causes of the excess cancer.

Does the analysis provide evidence that there is some hidden factor, possibly power lines, leading to a higher cancer rate at Slater? Not necessarily. The rate at Slater is higher than the national average, and the analysis says the discrepancy is moderately difficult to explain by chance. But there are plenty of other possible explanations besides chance. The Slater teachers are not selected randomly; they are probably different in many ways from the women who go into the national average. A more thorough analysis

would look for differences that might explain the increased cancer rate. And it would take account of other studies of the relationship between power lines and cancer.

2.3 Reassessing the Chances

Evidence or data causes us to revise our beliefs. And therefore it causes us to reassess the chances. Most people are used to revising their beliefs or opinions casually and might even think they do a good job of it. But research has shown time and time again that most people revise their beliefs irrationally. This section shows how to revise beliefs and reassess chances correctly, according to the laws of probability. The tool is Bayes' Theorem.

For example, a patient thinks she might have a particular disease and consults a doctor. Based on symptoms, history, prevalence, etc., the doctor forms an opinion and assesses the chance that the patient has the disease. Next, the doctor orders a test. But tests aren't perfect. Usually, patients with the disease test positive; but, sometimes they test negative. And usually, patients without the disease test negative; but sometimes they test positive. When the patient tests positive that is evidence, but not conclusive, that she has the disease. The doctor forms an updated opinion and reassesses the chance.

Because the test was positive, the new chance will be more than the old one. But how much more? That's the point where peoples' judgement tends to be faulty and where we need probability to help us. The doctor knows $P(\text{Disease})$, the chance this patient has Disease; that's what he assessed initially. He also knows $P(+|\text{Disease})$, the conditional chance of a positive test among people who have the disease and $P(+|\text{no Disease})$, the conditional chance of a positive test among people who don't have the disease. He wants $P(\text{Disease}|+)$, the conditional chance this patient has the disease given her positive test. This is a case for Bayes' theorem.

We've already seen just such an example in the AIDS example of Section 1.1. There, the incidence of AIDS was 0.2 of 1%. Let's take that as the doctor's initial assessment of the chance that this particular patient has AIDS. When she tested positive we saw that the revised assessment should be only about 16%.

Example. **The Slater School, continued.** Recall Example 1 of Chapter 1.3 about a high rate of reproductive cancer among female teachers at the Slater school. There were four statistical hypotheses: that the reproductive cancer rate among female teachers was (a), 3%, (b), 4%, (c), 5% or (d), 6%. I wanted to assess my chances for each of these four hypotheses accounting both for what I already knew about power lines and cancer

and for the evidence from Slater school itself. My plan was to assess the chances without considering the evidence from Slater, and then revise the assessment using Bayes' Theorem.

To begin, I considered the chance that high tension power lines contributed to the reproductive cancer rate. I had read some articles about epidemiological studies showing that proximity to high tension lines was correlated with cancer. But I had also read the statements of highly respected physicists and biologists that the electric and magnetic fields produced by high tension lines were so small that they could not have any biological effect. I wasn't sure which side was correct and I found the epidemiological studies and the statements of the physicists and botanists about equally compelling. So I assessed the chance of hypothesis (a), the Slater rate is 3%, as about 50%.

Now, supposing high tension lines do increase cancer, how much can they increase it? A rate of 6% meant that about 50% of all reproductive cancers would be due to the high tension lines. I thought the effect of high tension power lines was unlikely to be that high. So my chance for hypothesis (d) was fairly small. In fact, I assessed it as about 10%. I also thought that hypothesis (b) was slightly more plausible than hypothesis (c). After a bit more thought I assessed the chance of (b) as about 22% and the chance of (c) as about 18%.

These were my initial assessments:

$$P(A) \approx 50\%, \quad P(B) \approx 22\%, \quad P(C) \approx 18\%, and \quad P(D) \approx 10\%.$$

In Section 2.2 we computed

$$\begin{aligned} P(\text{8 cancers}|A) &\approx 4\% \\ P(\text{8 cancers}|B) &\approx 11\% \\ P(\text{8 cancers}|C) &\approx 14\% \\ P(\text{8 cancers}|D) &\approx 13\% \end{aligned}$$

I wanted my revised assessments, i.e. $P(A|\text{8 cancers})$, $P(B|\text{8 cancers})$, $P(C|\text{8 cancers})$, and $P(D|\text{8 cancers})$. Bayes' theorem gives the answer.

$$\begin{aligned} P(A|\text{8 cancers}) &\approx 25\% \\ P(B|\text{8 cancers}) &\approx 27\% \\ P(C|\text{8 cancers}) &\approx 31\% \\ P(D|\text{8 cancers}) &\approx 17\% \end{aligned}$$

2.4 Inference

In this chapter we've been using data to help distinguish between different box models. The process is called *inference*. Here's a summary of inference when we're distinguishing between 0-1 boxes. It applies when

- the data are like draws from a box,
- the box is a 0-1 box, and
- we don't know the proportions of 0's and 1's that belong in the box.

We want to learn about the proportions of 0's and 1's. The procedure is:

1. Propose some hypotheses; that is, propose box models with different proportions of 0's and 1's. In our examples, this is the number of U's and D's for the thumbtack, the number of 0's and 1's for the psychic, or the number of 0's and 1's for the Slater School. How many should you propose; and which ones? There's no set answer. In the Slater example we proposed four hypotheses with different proportions that
   (a) were different enough from each other to represent scientifically important differences, and
   (b) spanned the range of reasonable proportions,
2. Calculate how well each hypothesis explains the data. Because the data are like classifying and counting the draws from a 0-1 box, the binomial formula applies. That's what we used in the thumbtack example. When the sample size is large, the normal approximation is easier and accurate. We used it in the psychic and Slater examples.

   This step says how well each hypothesis explains the data and, what is the same thing, how strongly the data support each hypothesis. Sometimes that's all that's necessary. But sometimes we also want to reevaluate the chances. That's where the remaining steps come in.
3. Assess the *a priori* chances of the different hypotheses, that is, the chances you would assign each box model without considering the data at hand. We did that for the AIDS and Slater examples.
4. Use Bayes' Theorem to reassess the chances. The reassessed chances tell us how likely each hypothesis is in the light of both the data and everything else we know.

The previous procedure applies when the data are like draws from a 0-1 box and we are trying to learn about the proportion of 1's. But sometimes the data are not like draws from a 0-1 box. Another important case is when the data follow the normal curve. In this case we usually want to learn about the expected value of the chance process governing the data.

[examples here]

2.5 Odds And The Strength of Evidence

The *odds of A*, for some event A, means the ratio of two chances: the chance that A happens divided by the chance that A doesn't happen.

$$\textit{Odds of } A = \frac{\text{chance of A}}{\text{chance of not A}}.$$

Odds and chances are two ways of saying the same thing. If we know the odds we can compute the chances and if we know the chances we can compute the odds.

$$\begin{aligned} \text{If odds} &= \frac{3}{4}, \\ \text{then} & \\ \text{chance} &= \frac{3}{3+4}. \\ \text{If chance} &= \frac{7}{8}, \\ \text{then} & \\ \text{odds} &= \frac{7}{1}. \end{aligned}$$

Let's look more closely at how to reassess chances using the AIDS test (Sections 1.1, 2.2) as an example. What was the strength of evidence? The conditional chance of a positive test, given that the patient has AIDS, is 97%. And the conditional chance of a positive test, given that the patient does not have AIDS, is 1%. It's the ratio that matters: $97\%/1\% = 97$. The hypothesis that the patient has AIDS explains the test result about 97 times better than the hypothesis that the patient does not have AIDS.

The doctor originally assessed the chance of AIDS as 0.2 of 1%. The odds were $0.2/99.8 \approx 0.002$. The reassesed chance accounting for the diagnostic test was about 16%. The new odds are $16/84 \approx 0.19$. Look at the ratio.

$$\frac{\text{new odds}}{\text{old odds}} \approx \frac{0.19}{0.002} \approx 95 \approx \text{strength of evidence.}$$

It's not just a coincidence that the ratio of new odds to old odds is approximately equal to the strength of evidence. It's a mathematical necessity. For any event A and data D,

$$\begin{aligned}\frac{\text{new odds of } A}{\text{old odds of } A} &= \frac{\text{new chance of } A \times \text{old chance of not } A}{\text{new chance of not } A \times \text{old chance of } A} \\ &= \frac{\text{old chance of } A \times P(D|A) \times \text{old chance of not } A}{\text{old chance of not } A \times P(D|\text{not } A) \times \text{old chance of } A} \\ &= \frac{P(D|A)}{P(D|\text{not } A)},\end{aligned}$$

which is the strength of the evidence. The first line follows from the definition of odds; the second from Bayes' Theorem.

We've just shown that the ratio of new odds to old odds is the strength of evidence. Another way to put it is new odds of A = old odds of $A \times$ strength of evidence. If we had carried out the computations more accurately we would have this equality hold exactly in the AIDS example.

Statisticians have another name for the strength of evidence. They call it the *Bayes' factor*. Odds depend on how the chances are assessed. Two people might assess the chances differently, and have different odds. But the Bayes' factor is the same for both of them. And they both revise their odds the same way: by multiplying the old odds times the Bayes' factor. The Bayes' factor summarizes the evidence in a way that depends only on the data, not on how different people assess the chances.

If the conditional chances $P(D|A)$ and $P(D|\text{not } A)$ are about the same then the Bayes' factor is about 1, the two hypotheses explain the data about equally well, and the new odds are about the same as the old odds. On the other hand, if $P(D|A)$ is much bigger than $P(D|\text{not } A)$ then the Bayes' factor is very large, A explains the data much better than not A and the new odds are much bigger than the old odds.

2.6 Exercises

1. Suppose the evidence from the Slater School had been 27 cancers out of 450 teachers. What would my revised chances be?

2. Suppose the evidence from the Slater School had been 8 cancers out of 145 teachers, but my original chances had been .8, .1, .05, .05 for hypotheses A, B, C, and D, respectively. What would my revised chances be?

3. You have been told by one student that a particular professor always grades in a way that 70% of the students get As and 30% get lower

grades. Another student insists that the professor gives 30% As and 70% lower grades. You trust the opinion of both students equally. The professor tells you one of the two students is correct but will not tell you which one. Hence you give prior probability of 50% to each student being correct. You then get a random sample of 10 students that have taken this professor's class and you find out their grades. In your sample, 6 received an A grade. What is your posterior probability that the professor gives 70% As? What is your posterior probability that the professor gives 30% As?

4. In the weight loss example of section 2.1, what is the prior probability that the average weight loss is less than 10 pounds? (i.e. average ¿ -10) What is the posterior probability that the average weight loss is less than 10 pounds?

Chapter 3

Making Decisions

3.1 Actions and Utilities

We are often faced with the problem of making a decision. There are at least two possible *choices* or *actions*; if there were only one, there would be no decision. And we don't know what the outcome of the decision will be; if we did, there would be no problem.

Recall the space shuttle Challenger from Section 1.2. The shuttle was on the launch pad. NASA had two possible actions: *launch now* and *postpone*. And the outcome of *launch now* was unknown; either the O-ring would work, or it wouldn't.

The outcomes were unknown because the state of the world was unknown. To simplify, we can picture the O-ring as either supple (flexible enough to work properly) or stiff (not sufficiently flexible). It is useful to make a table such as Table 3.1. The possible states are listed across the top; the possible actions are listed down the left. The entries in the table are *utilities*. If NASA decides to *postpone*, the outcome is moderately positive because the O-ring works (for the postponed launch) but there is a cost of delay. If NASA decides to *launch now*, and the O-ring is *supple*, the result is very positive. That's the best possible outcome. On the other hand, if NASA decides to *launch now*, and the O-ring is *stiff*, the result is very negative. That's the best possible outcome. And that's what actually happened.

L. J. Savage was .... Here's what he says. .... The classic example of utilities is the rain-umbrella example, shown in Table 3.1. A person going out for a walk is deciding whether to carry an umbrella. There are two possible *decisions*, or *actions*. The person could either carry the umbrella,

	O-ring supple	*O-ring stiff*
Launch Now	very positive	very negative
Postpone	moderately positive	moderately positive

Table 3.1: The Challenger

	Rain	No Rain
Carry Umbrella	slightly positive	moderately positive
Leave Umbrella	very negative	very positive

Table 3.2: Umbrella Problem

or leave it home. The actions are shown down the left side of the table. There are also two possible *truths*, or *states of nature*. Either it will rain, or it won't. The states are shown along the top of the table.

The entries in the table are the *utilities* of all possible combinations of actions and states of nature. Utility means value, or worth. The utility of the combination (Carry Umbrella, Rain) is slightly positive because, although the person will enjoy the walk, the rain will keep him from enjoying it very much. The utility of (Carry Umbrella, No Rain) is moderately positive because the person will enjoy the walk but will have the inconvenience of carrying the umbrella. The utility of (Leave Umbrella, Rain) is very negative because the person gets wet and has a miserable time. And the utility of (Leave Umbrella, No Rain) is very high, because the person will enjoy the walk without any inconvenience.

The ultimate purpose of writing down the utilities is to help make decisions. In many cases, the utilities only need to be assessed approximately; and a table such as Table 3.1 is sufficient. In other cases it is necessary to assess utilities more accurately. Table 3.1 is an example of how one person might assess the utilites more accurately. These utilities are specific to the person making the decision. You and I may have different utilities for the four possible combinations.

A statistics instructor announces at the beginning of the term that there will be six pop quizzes during the term. One student taking the class is deciding how to spend the evening — either studying or attending a classical music concert. (In our experience, these are the only two activities that

	Rain	No Rain
Carry Umbrella	5	15
Leave Umbrella	-30	20

	Quiz Tomorrow	No Quiz
Study	very positive	0
Concert	very negative	small positive

students report doing in the evenings.) If she studies, she will miss the concert. But, if she goes to the concert and there is a quiz tomorrow, disaster! Table 3.1 shows her utilities. Everyone would agree that Statistics is **much** more important than music. So the utility of (Study, Quiz) is very positive while the utility of (Concert, Quiz) is very negative. We must concede that some people find pleasure in music; so the utility of (Concert, No Quiz) is slightly positive. Her decision will depend on exactly how she assesses her utilities, and on how she assesses the chance of a quiz tomorrow.

[examples with real numbers?] [mastectomy, or other cancer treatment example?] [public policy example?] [example with different utilities for different people?]

3.2 Decisions

Consider again the rain-umbrella example for the walker who filled in Table 3.1. Also, the walker has a box model for the occurence of rain. The box model might come from a professional weather forecaster or it might be made subjectively by the walker. Let's say he estimates the chances of rain as 30%. How should he decide whether to carry the umbrella?

If he has accurately assessed his utilities then carrying an umbrella is like drawing from a box in which 30% of the tickets are marked 5 and 70% are marked 15. The average of the box is $30\% \times 5 + 70\% \times 15 = 12$. And leaving the umbrella at home is like drawing from a box in which 30% of the tickets are marked -30 and 70% are marked 20. The average of the box is $30\% \times -30 + 70\% \times 20 = 5$. Because it has a higher average, he should prefer the first box. He should carry the umbrella.

The rain-umbrella decision is an example of the formal way decisions can be made in any situation. The ingredients are

- a list of actions
- a list of states of nature
- a box model for the states of nature
- utilities for each combination of action and state of nature.

With these ingredients, the procedure is

	Quiz Tomorrow	No Quiz
Study	10	0
Concert	-10	2

1. For each action, make a box model. The numbers on the tickets are the utilities for all combinations of that action with any state of nature. These are the numbers in one row of the utility table — the row corresponding to that action. The percentages of the different kinds of tickets are the chances of the different states of nature.

2. For each action, calculate the average of the box model.

3. Select the action with the largest average.

Here's how it would work for the student deciding between studying Statistics and attending a concert. (See Chapter **??**). The ingredients are

- A list of actions. In this case the actions are *Study* and *Attend Concert*.
- A list of states of nature. These are *Quiz Tomorrow* and *No Quiz*.
- A box model for the states. Let's say that the student judges the chance of *Quiz Tomorrow* to be 30%.
- Utilities. Suppose the student agrees in spirit with Table 3.1 and specifies the specific numbers in Table 3.2.

She would follow this procedure.

1. She makes two box models — one for *Study* and one for *Attend Concert*. The box model for *Study* has two kinds of tickets. 30% of the tickets are 10's; 70% are 0's. The box model for *Attend Concert* has 30% -10's and 70% 2's.

2. The average of the *Study* box is $30\% \times 10 + 70\% \times 0 = 3$. The average of the *Attend Concert* box is $30\% \times -10 + 70\% \times 2 = -1.6$.

3. She chooses the action with the highest average. Naturally, because we're constructing the examples, she chooses to study.

Here's how it would work in the space shuttle example. Just to get started, we've filled in some numbers in Table 3.1 to get the utilities in Table 3.2. The two numbers in the Postpone row are equal because what happens at the postponed launch does not depend on whether the O-ring is

	O-ring works	O-ring fails
Launch Now	1,000	-100,000
Postpone	900	900

working today. In other words, it does not depend on which state of nature (of the two in the utility table) is true. The utility for (Postpone, O-ring works) is a little less than the utility for (Launch Now, O-ring works) to reflect the cost and inconvenience of postponement. And the cost of (Launch Now, O-ring fails) reflects the disasterous result of a crash.

The one missing ingredient so far is the box model for the states of nature. Let's suppose for the moment that the chance of *O-ring fails* is about 10%.

The procedure says to make two box models — one for *Launch Now* and one for *Postpone*. The one for *Launch Now* has two types of tickets. Some are marked 1,000; the rest are marked -100,000. Those are the utilities on the *Launch Now* row. And the procedure also gives the percentages of each kind of ticket. 10% of the tickets are 1,000's, because the chance of *O-ring works* is 10%. And 90% of the tickets are -100,000, because the chance of *O-ring fails* is 90%.

On the other hand, the box for *Postpone* has only one kind of ticket. They are all marked 900.

The average of the *Launch Now* box is $10\% \times 1,000 + 90\% \times -100,000 = -89,900$. And the average of the *Postpone* box is 900. The correct decision is to choose the box with the larger average. And that means, assuming we have correctly specified the chances and utilities, the correct decision is to postpone the launch.

3.2.1 Sensitivity

(In this section we explain sensitivity to utilities, chances and affine transformations of utilities.)

In formal decision making there is always the question of how accurately the utilities and chances have been assessed and whether the decision would change if they are assessed slightly differently. In the shuttle example it is easy to calculate how different the chances and utilities would have to be in order to change the decision. For example, how low would the chance of *O-ring fails* have to be before the correct decision is *Launch Now*?

Let $x\%$ stand for the chance of *O-ring works*. Then the average of the *Launch Now* box is $x\% \times 1,000 + (100 - x)\% \times -100,000 = x\% \times 101,000 - 100,000$. And the average of the *Postpone* box is 900. The decision is borderline if the two averages are equal. That means $x\% \times 101,000 -$

$100,000 = 900$, or $x\% = 100,900/101,000 \approx 0.999$, or about 99.9%. In other words, the correct decision is *Postpone*, unless we believe that the chance of *O-ring fails* is less that 0.1%.

We can make a similar calculation for the utilities. Let u be the utility of (Launch Now, O-ring fails), and leave the other utilities as they are. Then the average of the *Launch Now* box is $10\% \times 1,000 + 90\% \times u$. And the average of the *Postpone* box is still 900. The decision is borderline if the two averages are equal. That means $10\% \times 1,000 + 90\% \times u = 900$ or $u = (900 - 1,000)/90\% \approx -111$. In other words, the decision is borderline if the cost of (Launch Now, O-ring fails) is just a little bit more than the cost of postponement.

We see that the decision to postpone the launch is correct over a wide range of utilities and chances of failure. As long as Table 3.2 is a reasonable approximation of the true utilities, and 10% is a reasonable estimate of the chance of failure, then the correct decision is clearly to postpone the launch.

3.3 Exercises

1.

2.

3.

Chapter 4

Predictions

Chapter 5

Case Study

5.1 GUSTO

The following case study was adapted from "The Mathematics of Making Up Your Mind" by Will Hively, *Discover*, May, 1996. The data was taken from "Placing Trials in Context Using Bayesian Analysis: GUSTO Revisited by Reverand Bayes" by James Brophy and Lawrence Joseph, *Journal of the American Medical Association*, Vol. 273, 1996.

A clinical trial was conducted to compare two treatments for heart attacks–streptokinase and tissue plasminogen activator (t-PA). Cardiologists agree that both drugs work well: more than 90% of all patients who receive either medication survive. Where they disagree is on which of the drugs they should used. In one large study of about 20,000 patients, streptokinase did slightly (less than 1%) better. In another study of about 30,000 patients, t-PA did slightly (less than 1%) better. The cost of t-PA is $1,530 per use while the cost of streptokinase is $220. In Canada and Europe, most doctors give streptokinase. In the US, most doctors give t-PA.

A few years ago, Genentech, t-PA's manufacturer, joined with 4 other companies in sponsoring a third clinical trial with over 40,000 patients–called GUSTO (Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Arteries). GUSTO organizers said that if they could show that the t-PA survival rate was 1than the streptokinase survival rate, then t-PA should be considered clinically superior. A 1% difference may seem small, but in cardiology it can mean a lot - perhaps as many as 5,000 lives per year. The results published in 1993 found 92.7% survival for streptokinase and 93.7% survival for t-PA. Published with the results was a test of the null hypothesis that the survival rate on streptokinase was

equal to the survival rate on t-PA. The observed significance level (p-value) showed that there was .001, or 1 in 1,000 chance that t-PA would perform at least this much better if it was merely as good as streptokinase.

Lawrence Joseph (statistician) and cardiologist James Brody did not buy the results. They conducted an analysis of the data using Bayes theorem to combine the results of previous trials with the results from the GUSTO study. They also calculate results starting with beliefs that placed equal weight on the superiority of each drug. They calculated two probabilities. The first was the probability the survival rate on t-PA was greater than the survival rate on streptokinase given the data. This probability was close to 1. Even starting from beliefs that gave some advantage to streptokinase, Bayes rule showed that the probability that the survival rate for t-PA was greater than the survival rate for streptokinase was near 1. There is no controversy on that point. However, a second calculation is necessary. The organizers of the trial said that differences of at least 1% showed clinical superiority, differences less than 1% did not. This is like saying that if t-PA reduces the death rate by at least 1% then it is worth the extra $1,310 per use. Joseph and Brody calculated the probability that the difference in the survival rates of the two drugs was greater than or equal to 1%. They made this calculation even with beliefs that ignored the previous study which found streptokinase more effective. That is they stacked the deck in favor of t-PA. Even so, given the GUSTO data, the probability that the difference in survival rates was greater than or equal to 1% was only 50%. As a comparison, Joseph and Brody used prior beliefs that weighted the two previous studies equally with GUSTO. Under this prior Bayes Rule gives probability for t-PA's clinical superiority as nearly 0actually be weighted less than the two earlier studies as it was not a blind trial. Physicians knew which drug they were giving, and patients who got t-PA 'apparently' were 1coronary bypass operation as well.

At the time of publication Brophy rated the chances of t-PA's being clinically superior to streptokinase as "no better than 5 or 10%." At that rate, t-PA would save about one more life among every 250 heart attack victims. To justify using t-PA, that person's life must be worth $327,500, the extra cost of giving t-PA to all 250 patients.

5.2 Exercises

1. Using normal approximations calculate the p-value for the null hypothesis of no treatment difference for each of the three studies separately. Then calculate this p-value for all three studies combined. Write one paragraph summarizing the results. What might explain

Trial	Agent	Sample Size	No. (%) of Deaths
Trial 1	SK	13780	1455 (10.6%)
	t-PA	13746	1418 (10.3%)
Trial 2	SK	10396	929 (8.9%)
	t-PA	10372	993 (9.6%)
Trial 3	SK	20173	1473 (7.3%)
	t-PA	10343	652 (6.3%)

the conflicting results between the studies?

2. If apriori our beliefs about the treatment difference are represented by a 'flat prior' (i.e. all possible values are equally believable), then after observing the data, the probability distribution of the treatment difference is approximately a normal curve centered at the observed treatment difference with a SD equal to the standard error of the difference. Using such a prior, calculate the probability that t-PA has at least a 1% lower mortality rate than streptokinase in each study separately. Then calculate the same probability using the data from all three trials combined. Write one paragraph summarizing the results. What do you conclude about the GUSTO investigators' claim that t-PA is superior?

3. You are the head of Genentech. You want a market for your drug t-PA. Using the statistical evidence above, write one paragraph justifying why clinicians should treat with t-PA, third-party payers (i.e. insurance companies, Medicare, ...) should cover the cost, and consumers should demand its use.

4. You are the head of an insurance agency. You must decide whether you are going to cover the cost of t-PA rather than streptokinase in your insurance policy. Make a decision, and using the statistical evidence above, write one paragraph justifying your decision.

5. You are the mother/father of a child (age 12) with leukemia. Your child has gone through all standard therapies and the only hope is for a new experimental treatment that costs $327,500 and has a survival rate of 1 in 250 among young children with leukemia. You know insurance companies have to make choices. They can't afford covering both t-PA and your child's treatment. Using the statistical evidence above, write an argument to your insurance company as to why they should opt to pay your child 's medical bill rather than cover t-PA treatment for the next 250 heart attack victims (average age=70).