Chapter 11

Hypothesis Testing

11.1 Back to the Euro problem

In Section 4.1 I presented a problem from MacKay's *Information Theory, Inference, and Learning Algorithms*:

> A statistical statement appeared in "The Guardian" on Friday January 4, 2002:
>
> > When spun on edge 250 times, a Belgian one-euro coin came up heads 140 times and tails 110. 'It looks very suspicious to me,' said Barry Blight, a statistics lecturer at the London School of Economics. 'If the coin were unbiased, the chance of getting a result as extreme as that would be less than 7%.'
>
> But do these data give evidence that the coin is biased rather than fair?

We estimated the probability that the coin would land face up, but we didn't really answer MacKay's question: Do the data give evidence that the coin is biased?

In Chapter 4 I proposed that data are in favor of a hypothesis if the data are more likely under the hypothesis than under the alternative or, equivalently, if the Bayes factor is greater than 1.

In the Euro example, we have two hypotheses to consider: I'll use F for the hypothesis that the coin is fair and B for the hypothesis that it is biased.

If the coin is fair, it is easy to compute the likelihood of the data, $\mathrm{p}(D|F)$. In fact, we already wrote the function that does it.

```
def Likelihood(self, data, hypo):
    x = hypo / 100.0
    head, tails = data
    like = x**heads * (1-x)**tails
    return like
```

To use it we can create a `Euro` suite and invoke `Likelihood`:

```
suite = Euro()
likelihood = suite.Likelihood(data, 50)
```

$p(D|F)$ is $5.5 \cdot 10^{-76}$, which doesn't tell us much except that the probability of seeing any particular dataset is very small. It takes two likelihoods to make a ratio, so we also have to compute $p(D|B)$.

It is not obvious how to compute the likelihood of B, because it's not obvious what "biased" means.

One possibility is to cheat and look at the data before we define the hypothesis. In that case we would say that "biased" means that the probability of heads is 140/250.

```
actual_percent = 100.0 * 140 / 250
likelihood = suite.Likelihood(data, actual_percent)
```

This version of B I call `B_cheat`; the likelihood of `b_cheat` is $34 \cdot 10^{-76}$ and the likelihood ratio is 6.1. So we would say that the data are evidence in favor of this version of B.

But using the data to formulate the hypothesis is obviously bogus. By that definition, any dataset would be evidence in favor of B, unless the observed percentage of heads is exactly 50%.

11.2 Making a fair comparison

To make a legitimate comparison, we have to define B without looking at the data. So let's try a different definition. If you inspect a Belgian Euro coin, you might notice that the "heads" side is more prominent than the "tails" side. You might expect the shape to have some effect on x, but be unsure whether it makes heads more or less likely. So you might say "I think the coin is biased so that x is either 0.6 or 0.4, but I am not sure which."

We can think of this version, which I'll call `B_two` as a hypothesis made up of two sub-hypotheses. We can compute the likelihood for each sub-hypothesis and then compute the average likelihood.

```
like40 = suite.Likelihood(data, 40)
like60 = suite.Likelihood(data, 60)
likelihood = 0.5 * like40 + 0.5 * like60
```

The likelihood ratio (or Bayes factor) for `b_two` is 1.3, which means the data provide weak evidence in favor of `b_two`.

More generally, suppose you suspect that the coin is biased, but you have no clue about the value of x. In that case you might build a Suite, which I call `b_uniform`, to represent sub-hypotheses from 0 to 100.

```
b_uniform = Euro(xrange(0, 101))
b_uniform.Remove(50)
b_uniform.Normalize()
```

I initialize `b_uniform` with values from 0 to 100. I removed the sub-hypothesis that x is 50%, because if x is 50% the coin is fair, but it has almost no effect on the result whether you remove it or not.

To compute the likelihood of `b_uniform` we compute the likelihood of each sub-hypothesis and accumulate a weighted average.

```
def SuiteLikelihood(suite, data):
    total = 0
    for hypo, prob in suite.Items():
        like = suite.Likelihood(data, hypo)
        total += prob * like
    return total
```

The likelihood ratio for `b_uniform` is 0.47, which means that the data are weak evidence against `b_uniform`, compared to F.

If you think about the computation performed by `SuiteLikelihood`, you might notice that it is similar to an update. To refresh your memory, here's the `Update` function:

```
def Update(self, data):
    for hypo in self.Values():
        like = self.Likelihood(data, hypo)
        self.Mult(hypo, like)
    return self.Normalize()
```

And here's `Normalize`:

```
def Normalize(self):
    total = self.Total()

    factor = 1.0 / total
```

```
        for x in self.d:
            self.d[x] *= factor

        return total
```

The return value from `Normalize` is the total of the probabilities in the Suite, which is the average of the likelihoods for the sub-hypotheses, weighted by the prior probabilities. And `Update` passes this value along, so instead of using `SuiteLikelihood`, we could compute the likelihood of `b_uniform` like this:

```
likelihood = b_uniform.Update(data)
```

11.3 The triangle prior

In Chapter 4 we also considered a triangle-shaped prior that gives higher probability to values of x near 50%. If we think of this prior as a suite of sub-hypotheses, we can compute its likelihood like this:

```
b_triangle = TrianglePrior()
likelihood = b_triangle.Update(data)
```

The likelihood ratio for `b_triangle` is 0.84, compared to F, so again we would say that the data are weak evidence against B.

The following table shows the priors we have considered, the likelihood of each, and the likelihood ratio (or Bayes factor) relative to F.

Hypothesis	Likelihood $\times 10^{-76}$	Bayes Factor
F	5.5	–
`B_cheat`	34	6.1
`B_two`	7.4	1.3
`B_uniform`	2.6	0.47
`B_triangle`	4.6	0.84

Depending on which definition we choose, the data might provide evidence for or against the hypothesis that the coin is biased, but in either case it is relatively weak evidence.

In summary, we can use Bayesian hypothesis testing to compare the likelihood of F and B, but we have to do some work to specify precisely what B means. This specification depends on background information about coins and their behavior when spun, so people could reasonably disagree about the right definition.

My presentation of this example follows David MacKay's discussion, and comes to the same conclusion. You can download the code I used in this chapter from `http://thinkbayes.com/euro3.py`. For more information see Section 0.3.

11.4 Discussion

The Bayes factor for `B_uniform` is 0.47, which means that the data provide evidence against this hypothesis, compared to F. In the previous section I characterized this evidence as "weak," but didn't say why.

Part of the answer is historical. Harold Jeffreys, an early proponent of Bayesian statistics, suggested a scale for interpreting Bayes factors:

Bayes Factor	Strength
1 – 3	Barely worth mentioning
3 – 10	Substantial
10 – 30	Strong
30 – 100	Very strong
> 100	Decisive

In the example, the Bayes factor is 0.47 in favor of `B_uniform`, so it is 2.1 in favor of F, which Jeffreys would consider "barely worth mentioning." Other authors have suggested variations on the wording. To avoid arguing about adjectives, we could think about odds instead.

If your prior odds are 1:1, and you see evidence with Bayes factor 2, your posterior odds are 2:1. In terms of probability, the data changed your degree of belief from 50% to 66%. For most real world problems, that change would be small relative to modeling errors and other sources of uncertainty.

On the other hand, if you had seen evidence with Bayes factor 100, your posterior odds would be 100:1 or more than 99%. Whether or not you agree that such evidence is "decisive," it is certainly strong.

11.5 Exercises

Exercise 11.1. *Some people believe in the existence of extra-sensory perception (ESP); for example, the ability of some people to guess the value of an unseen playing card with probability better than chance.*

What is your prior degree of belief in this kind of ESP? Do you think it is as likely to exist as not? Or are you more skeptical about it? Write down your prior odds.

Now compute the strength of the evidence it would take to convince you that ESP is at least 50% likely to exist. What Bayes factor would be needed to make you 90% sure that ESP exists?

Exercise 11.2. *Suppose that your answer to the previous question is 1000; that is, evidence with Bayes factor 1000 in favor of ESP would be sufficient to change your mind.*

Now suppose that you read a paper in a respectable peer-reviewed scientific journal that presents evidence with Bayes factor 1000 in favor of ESP. Would that change your mind?

If not, how do you resolve the apparent contradiction? You might find it helpful to read about David Hume's article, "Of Miracles," at `http://en.wikipedia.org/wiki/Of_Miracles`.