Just Enough Likelihood*

Alan R. Rogers†

September 2, 2013

1. Introduction

Statisticians have developed several methods for comparing hypotheses and for estimating parameters from data. Of these, the method of maximum likelihood is remarkable both for its power and for its flexibility. Yet it is seldom covered in introductory statistics courses because it is hard to present either as a canned computer program or within a cookbook-style textbook. It is much too flexibile for that.

In this document, I will assume that you know a little about probability. If you need to brush up, consult *Just Enough Probability*, which you can find at `http://content.csbs.utah.edu/~rogers/pubs/index.html`. The method is easiest to describe by example, so I begin with the simplest experiment I can imagine.

2. Tossing a coin once

You are given a (possibly unfair) coin, and you toss it once. Since the probability of tossing heads is unknown, let us call it p. Let x represent the number of heads observed (either zero or one). If the coin comes up heads, then $x = 1$ and you have observed an event with probability p. On the other hand, if the coin comes up tails then $x = 0$ and you have observed an event with probability $1 - p$. In symbols,

$$\left.\begin{array}{rcl} \Pr[x=1] & = & p \\ \Pr[x=0] & = & 1-p \end{array}\right\} \quad (1)$$

†Dept. of Anthropology, 270S 1400E Rm 102, University of Utah.

2.1. The likelihood function

If we have just tossed a coin, then we will know the value of x, but we may not be sure about p. Consequently, we can think of formula 1 as a function of this unknown parameter value. This is the reverse of the usual situation in probability theory, where the parameters are taken as given and the outcomes are allowed to vary. To distinguish these two situations, equation 1 is called a *probability distribution* if taken as a function of the outcome variable x, but is called a *likelihood* if taken as a function of its parameter, p. For example, if we toss one coin and observe heads, then we have observed an event of probability p. The likelihood function in this case is

$$L(p) = p$$

On the other hand, if the coin had come up tails, the likelihood function would be

$$L(p) = 1 - p$$

The likelihood function is useful because it summarizes the information that the data provide about the parameters. To estimate a parameter, we make use of the principle of maximum likelihood:

Principle 1. *To estimate a parameter, the method of maximum likelihood chooses the parameter value that makes L as large as possible.*

• EXAMPLE **1**
If the coin came up heads, then $L(p) = p$. This function reaches its maximum value when $p = 1$. The maximum likelihood estimate of p is therefore $\hat{p} = 1$.
(Here, $\hat{p}$ is pronounced "pee-hat." It is the conventional symbol for a maximum likelihood estimate of p.)

• EXAMPLE **2**
If the coin came up tails, then $L(p) = 1 - p$ and $\hat{p} = 0$.

3. Tossing the coin twice

Had we tossed the coin twice, there would have been four possible ordered outcomes:

Table 1: Ordered outcomes from two tosses of a coin

Outcome		
toss 1	toss 2	Probability
heads	heads	p^2
heads	tails	$p(1-p)$
tails	heads	$p(1-p)$
tails	tails	$(1-p)^2$

It is often inconvenient to keep track of the outcome of each toss, so the table above is usually abbreviated as:

Table 2: Unordered outcomes from two tosses of a coin

x	Probability
2	p^2
1	$2p(1-p)$
0	$(1-p)^2$

• EXAMPLE **3**
If you observe one head in two tosses, then $L(p) = 2p(1-p)$ and $\hat{p} = 1/2$.
• EXAMPLE **4**
If you observe two heads in two tosses, then $L(p) = p^2$ and $\hat{p} = 1$.

4. Tossing the coin several times

In the general case, the coin is tossed N times yielding x heads and $N-x$ tails. The probability of observing x heads in N tosses is given by the binomial distribution function:

$$\Pr[x; N, p] = \binom{N}{x} p^x (1-p)^{N-x} \qquad (2)$$

In this expression, the notation $\binom{N}{x}$ is pronounced "N choose x" and represents the number of ways of choosing x heads out of N tosses. If you're unfamiliar with this formula, you can read more about it in *Just Enough Probability*.
• EXAMPLE **5**
In the case of two tosses, table 1 shows two ways of getting a single head. Consequently, $\binom{2}{1} = 2$, and equation 2 gives $\Pr[1; 2, p] = 2p(1-p)$, in agreement with the second row of table 2.

To understand how equation 2 works in the case of three tosses ($N = 3$), consider all the outcomes that can result from three tosses of a coin:

toss 1	toss 2	toss 3	x
H	H	H	3
H	H	T	2
H	T	H	2
H	T	T	1
T	H	H	2
T	H	T	1
T	T	H	1
T	T	T	0

Note that there is only one outcome in which H is absent—only one, in other words, in which $x = 0$. This implies that $\binom{3}{0} = 1$.

⋆EXERCISE **4–1** Use the table to figure out the values of $\binom{3}{1}$, $\binom{3}{2}$, and $\binom{3}{3}$.

Equation 2 becomes a likelihood function if we think of it as a function of p rather than x. For example, if we toss three coins and observe one head, the likelihood function is

$$L(p) = 3p(1-p)^2$$

To estimate p, the method of maximum likelihood chooses the value of p that maximizes the likelihood. The value of p that maximizes L will also maximize $\ln L$, so we can work with either function. It is often more convenient to work with $\ln L$ rather than with L itself. If $x = 1$ and $N = 3$, the log likelihood function is

$$\ln L(p) = \ln 3 + \ln p + 2\ln(1-p) \qquad (3)$$

Figure 1 graphs $\ln L$ for several coin tossing experiments in each of which 1/3 of the tosses come up heads. Each curve reaches a maximum at roughly $p = 1/3$. Thus, the maximum likelihood estimator must be close to 1/3 in each case. We can see this much by studying the graph.

The curves in figure 1 also provide information about the precision of the estimates: The likelihood function is flat when the sample size is

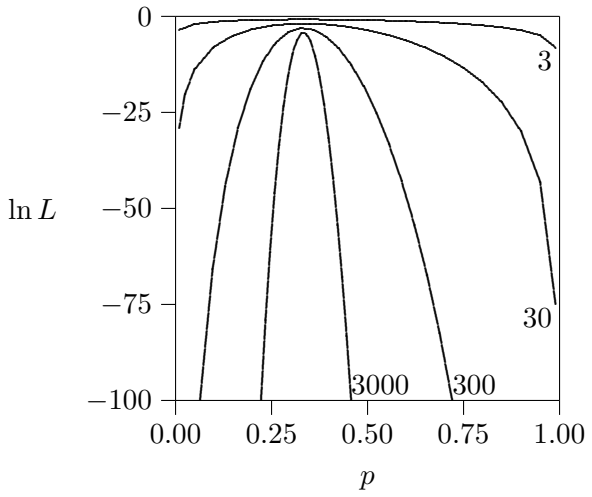

Figure 1: Log likelihood functions for binomial experiments in which $x/n = 1/3$. Each curve is for a different value of n. The numbers next to each curve show the value of n.

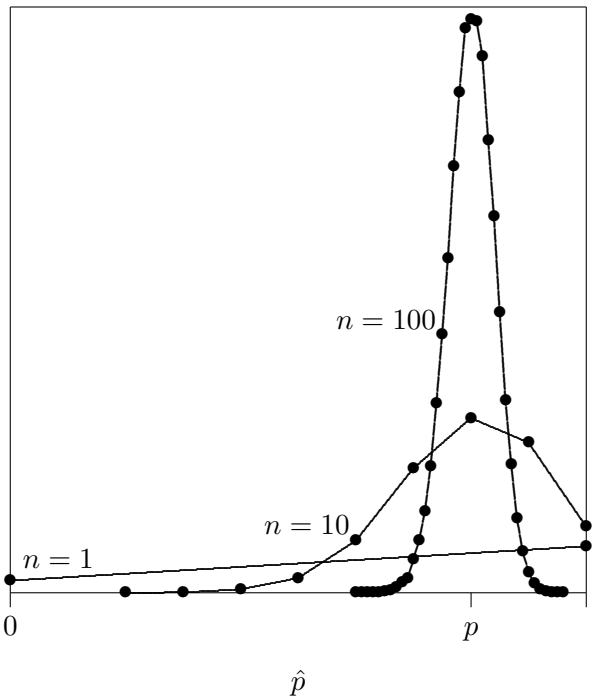

Figure 2: Frequency distributions of $\hat{p}$ when $p = 0.8$.

small, but is narrow and peaked when the sample size is large. This is a crucial point. It means that when N is large, our estimate of p is unlikely to be far from 1/3, the true value. The larger the data set, the stronger this claim becomes.

5. A maximum likelihood estimator for p

To obtain an estimator, it is convenient to work with the logarithm of L rather than with L itself. Taking the log of equation 2 gives

$$\ln L(p) = \ln \binom{N}{x} + x \ln p + (N-x) \ln(1-p) \quad (4)$$

The maximum likelihood estimator of p is the value of p that makes $\ln L(p)$ as large as possible. This estimator turns out to be

$$\hat{p} = x/N \quad (5)$$

in agreement with the examples above where $x/N = 1/3$.

⋆EXERCISE **5–1** Verify equation 5.

How well does this formula work? To find out, I analyzed data from computer simulations in which $p = 0.8$. The results are shown in figure 2 and cover three values of N. For each value of N, I generated thousands of simulated data sets and estimated p from each data sets. The distributions of these estimates are shown in figure 2. First look at the distribution for $N = 100$. In that case, the distribution is centered narrowly around the true parameter value, $p = 0.8$. Few of the simulated estimates are far from the true value, so we could have high confidence in an estimate from 100 real coin tosses. Now look at the distribution for $N = 10$. In that case, the distribution is spread much more widely—it would be easy to find estimates that differ from the true value by 0.2 or so. With a sample of 10 we get only a crude idea of the value of p. Finally, look at the distribution for $N = 1$. Here the distribution goes from wall to wall. A single toss of the coin would tell little about p.

Clearly, equation 5 gives estimates that vary in accuracy, and the accuracy depends heavily on the size of the sample. We need a way to measure this effect. The standard method involves something called the *sampling variance*.

Table 3: Results of 30 computer experiments, each simulating 20 tosses of an unfair coin. See text for details.

	Results	$\hat{p}$	$(\hat{p}-p)^2$
1	11011111101101111111	0.85	0.0025
2	11111001101111111111	0.85	0.0025
3	01100111111110111111	0.80	0.0000
4	10111110111101101111	0.80	0.0000
5	01101111111111101101	0.80	0.0000
6	10111111111100110111	0.80	0.0000
7	11111101101111001111	0.80	0.0000
8	01111111111111111110	0.90	0.0100
9	10111101111111010101	0.75	0.0025
10	11110111001111111111	0.85	0.0025
11	11111111101111110111	0.90	0.0100
12	11011101110011111110	0.75	0.0025
13	11111110111111110101	0.85	0.0025
14	01110111011101101001	0.65	0.0225
15	11001111011011110010	0.65	0.0225
16	11101111111111111111	0.95	0.0225
17	10101111101111110111	0.80	0.0000
18	01111110001110111011	0.70	0.0100
19	11110111101111101111	0.85	0.0025
20	11111111111111100011	0.85	0.0025
21	01111100111111111001	0.75	0.0025
22	11111111111111111111	1.00	0.0400
23	11111111111110101110	0.85	0.0025
24	01100111011111111111	0.80	0.0000
25	11111111111101111111	0.95	0.0225
26	10011110001111111111	0.75	0.0025
27	11111111111110111111	0.95	0.0225
28	10111111111110011101	0.80	0.0000
29	11101111101011111010	0.75	0.0025
30	11111011111111111111	0.95	0.0225

6. Sampling variance, standard error, and confidence intervals

Table 3 shows the results of 30 computer experiments, each simulating 20 tosses of an unfair coin. On each toss, we observe "1" with probability $p = 0.8$ and "0" with probability $1 - p$. For each experiment, $\hat{p}$ is the mean as estimated by equation 5. These estimates vary, reflecting the relatively small sample size in each experiment. The variance of the estimates about their expected value, p, is called the *sampling variance*. With these data, the sampling variance is estimated by

$$v = \frac{1}{30}\sum_{i=1}^{30}(\hat{p}_i - p)^2 = 0.0078 \qquad (6)$$

It is usually easier to interpret the square root of the sampling variance, which is called the *standard error*. With these data, the standard error is estimated by

$$S.E. = \sqrt{v} = 0.089$$

The standard error can be thought of as the size of a "typical" error.

Calculations such as these are easy when one has the luxury of repeating an experiment many times. Yet that is seldom the case. Fortunately, it is possible to estimate the sampling variance even when the experiment has been performed only once.

6.1. Using likelihood to estimate the sampling variance

On page 2, I pointed out that large samples generate narrow likelihood surfaces that are sharply curved at the peak. Small samples, on the other hand, generate broad flat ones with little curvature anywhere. This relationship underlies a remarkable formula, which makes it possible to estimate the sampling variance of a maximum likelihood estimator from a single data set. The sampling variance is approximately

$$v \approx -1 \Big/ E\left\{\frac{d^2 \ln L}{dp^2}\right\} \qquad (7)$$

For example, with equation 2 the first and second derivatives are

$$\begin{aligned}\frac{\partial \ln L}{\partial p} &= \frac{x}{p} - \frac{N-x}{1-p} \\ \frac{\partial^2 \ln L}{\partial p^2} &= -\frac{x}{p^2} - \frac{N-x}{(1-p)^2}\end{aligned}$$

The expected value of x is Np and that of $N-x$ is $N(1-p)$. Thus,

$$\begin{aligned}E\left\{\frac{\partial^2 \ln L}{\partial p^2}\right\} &= -\frac{Np}{p^2} - \frac{N(1-p)}{(1-p)^2} \\ &= -\frac{N}{p(1-p)}\end{aligned}$$

Plugging this into equation 7 gives the sampling variance for our estimate of p:

$$v = \frac{p(1-p)}{N}$$

This expresses the sampling variance of $\hat{p}$ in terms of the unknown parameter p. To use this answer with data, we would have to use $\hat{p}$ as an approximation for p. Thus, in practice the standard error is estimated as

$$v = \frac{\hat{p}(1-\hat{p})}{N} \tag{8}$$

• EXAMPLE **6**
For the data in table 3, $p = 0.8$ and $N = 20$. Thus, $v = 0.8 \times 0.2/20 = 0.008$, very close to the value (0.0078) estimated in equation 6.

It often turns out to be difficult to take the expectation of the second derivative. In such cases, the usual practice is to approximate equation 7 by

$$v \approx -1 \Big/ \left.\frac{d^2 \ln L}{dp^2}\right|_{p=\hat{p}} \tag{9}$$

Instead of taking the expectation of the second derivative, one simply evaluates it at the point where the parameter is equal to its estimate. With the example above,

$$\left.\frac{d^2 \ln L}{dp^2}\right|_{p=\hat{p}} = -\frac{N}{\hat{p}(1-\hat{p})}$$

so equations 9 and 7 both give the answer shown in equation 8.

• EXAMPLE **7**
When $\hat{p} = 0.8$ and $N = 100$, the estimated sampling variance is $v = 0.8 \times 0.2/100 = 0.0016$.

⋆EXERCISE **6–1** Use equation 9 to estimate the sampling variance for each of the first five experiments in table 3.

⋆EXERCISE **6–2** Use equations 7 and 9 to obtain a formula for the sampling variance of $\hat{p}$, which estimates the probabity of heads in a binomial experiment. Compare your answer to equation 8.

6.2. Standard error

The standard error, like the sampling variance, is an estimate of the error in an estimate. The larger the standard error of an estimate, the less acurrate that estimate is likely to be. The two estimates of error are closely related: the standard error is the square root of the sampling variance.

6.3. Confidence interval

What is a 95% confidence interval? Since a confidence interval is calculated from data, and the data themselves are random, the confidence interval is a random quantity too. If we repeat some experiment again and again and calculate a confidence interval from each fresh set of data, we will likely get a different confidence interval each time. The procedure for constructing confidence intervals is devised so that, on average, 95% of the intervals we construct will contain the true parameter value.

If a data set consists of a large number of independent observations, the maximum likelihood estimates that we obtain from it will have sampling distributions that are approximately normal. Consequently, normal distribution theory is usually used to place approximate confidence intervals around maximum likelihood estimates. In a normal distribution, 95% of the probability mass is within 1.96 standard deviations of the mean. Consequently, a 95% confidence interval for some parameter θ is the interval between

$$\hat{\theta} - 1.96 S.E. \quad \text{and} \quad \hat{\theta} + 1.96 S.E.$$

In words, the lower bound of the interval is 1.96 standard errors below the estimate and the upper bound is 1.96 standard errors above.

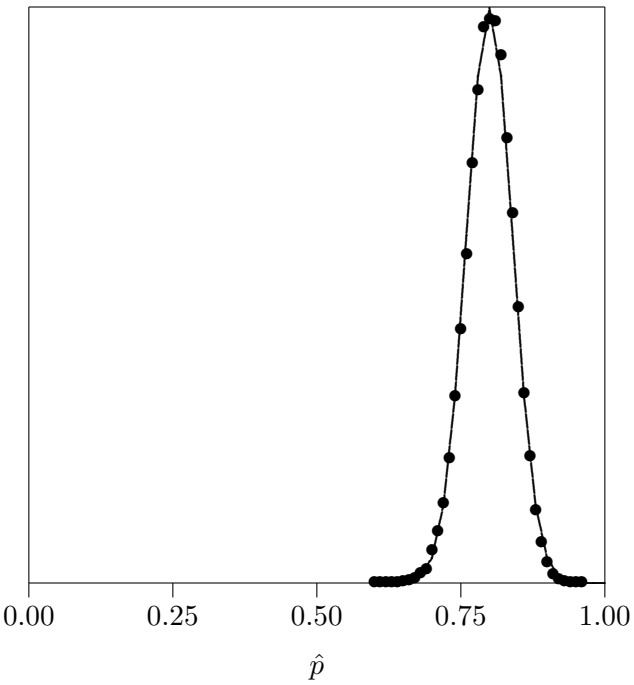

Figure 3: Normal approximation to sampling distribution. The solid line shows the normal approximation to the sampling distribution of $\hat{p}$ in the case where $p = 0.8$ and $N = 100$. The bullets are copied from the the corresponding curve in figure 2, and show the true sampling distribution of $\hat{p}$ as estimated by computer simulation.

• EXAMPLE **8**
Use the data from example 7 to calculate the standard error and the 95% confidence interval for $\hat{p}$.
◦ ANSWER
In that example $\hat{p} = 0.8$ and the sampling variance was 0.0016. The standard error is thus $\sqrt{0.0016} = 0.04$, and the 95% confidence interval is $[0.7216, 0.8784]$.

In this example, the sampling distribution of p should be normal with mean 0.8 and standard deviation 0.04. This distribution is shown as a solid line in figure 3. The bullets in the figure are remarkably close to the solid line, but they were not drawn using the normal approximation to the sampling distribution. They are simply copied from figure 2. They show that the normal distribution does a remarkable job of approximating the sampling distribution of $\hat{p}$.

⋆ EXERCISE **6–3** Use the first experiment in table 3 to calculate the standard error and the 95% confidence interval for $\hat{p}$.

The normal approximation is less useful when the likelihood function is asymmetric in the neighborhood of the estimate [2]. If it falls steeply on one side but only slowly on the other, the approximation is likely to be poor. This problem is particular severe in the case with which we started: that of tossing a coin a single time. In that case, the likilihood function reaches its maximum at $p = 0$ or $p = 1$. Since these are the smallest and largest permissible values, the likelihood function cannot be symmetric about either of them.

7. Likelihood-ratio tests

7.1. Nested models

Likelihood is also used to test one statistical model against another. The method method works for models that are "nested." In other words, it works if one model is a restricted version of another.

For example, suppose we suspect that two samples are drawn from the same normal distribution. This is a special case of a more general hypothesis, which assumes that the two distributions are normal but says nothing about whether they are the same. The first hypothesis differs from the second only in asserting that the parameters of the two normal distributions are equal. In other words, the first is a restricted version of the second, and the two are said to be nested. For this reason, we can test the first against the second using a likelihood-ratio test.

Or again, suppose we have data on migration among a set of villages, and we wish to know whether migration is symmetric in the sense that the rate from any village to another equals the opposing rate. One alternative might assume no relationship between the rates in opposing directions. The first hypothesis is a special case of the second, so the two are nested.

On the other hand, we might be interested in a different alternative hypothesis, which holds that the rate in one direction is the reciprocal of the opposing rate. If r_{ij} is the rate from village i to village j, then the original hypothesis (symmetric migration) holds that $r_{ij} = r_{ji}$ for

all village pairs. This new hypothesis holds that $r_{ij} = 1/r_{ji}$. Neither hypothesis is a restricted version of the other, so the two are not nested, and we cannot use a likelihood-ratio test. (We could however evaluate them using methods such as the *Akaike information criterion* [1].)

7.2. How to perform the test

Suppose H_0 and H_1 are hypotheses and that the second nested within the first. We wish to test H_1 against H_0. The procedure involves the following steps:

1. For each hypothesis, build a model that will allow you to calculate the likelihood given the data and specific assumptions about parameter values.

2. Calculate maximum-likelihood estimates of all parameters under each model.

3. Substitute these estimates into the likelihood functions to obtain L_0 and L_1, the maximal likelihood under each model.

4. Calculate $Z = -2\ln(L_1/L_0)$.

5. If the sample is large and H_0 is correct, Z is approximately a chi-squared random variable with degrees of freedom equal to the number of constraints imposed by H_1. Calculate the probability P that a random value drawn from this distribution would exceed Z.

6. Reject H_1 if $P < \alpha$, where α is the significance level chosen by the analyst (often 0.05 or 0.01).

7.3. Counting constraints

To perform this test, one needs to count the constraints involved in reducing H_1 to H_0. For example, in the first example above, H_1 assumed that two normal distributions (let's call them A and B) were the same. The normal distribution has two parameters (the mean μ and the variance σ^2), so there are two constraints: $\mu_A = \mu_B$, and $\sigma_A^2 = \sigma_B^2$. In general, constraints are described by equations, and the number of constraints equals the number of equations.

In the second example, the hypothesis of symmetric migration involved constraints of the form $r_{ij} = r_{ji}$, where r_{ij} is the rate of migration from village i to village j. We have a constraint of this form for each pair of villages, so the number of constraints equals the number of such pairs.

Sometimes these counts can be confusing. Consider a genetic locus with three alleles (three variants) labeled A_1, A_2, and A_3. If we have samples from two populations, we might want to know whether the two had different allele frequencies. Let H_1 represent the hypothesis that the two populations have equal frequencies and H_0 the hypothesis that makes no assumption about these frequencies. The first hypothesis is nested within the second, so we can use a likelihood-ratio test. How many constraints are there? Let p_{ij} represent the frequency of allele i in population j. H_1 differs from H_0 in assuming that

$$p_{11} = p_{12} \quad (10)$$

$$p_{21} = p_{22} \quad (11)$$

$$p_{31} = p_{32} \quad (12)$$

There are three equations here, so one might conclude that there were three constraints. Actually, there are only two. Within each population, the allele frequencies must sum to 1, so the frequency of the third allele can be written in terms of the other two: $p_{3i} = 1 - p_{1i} - p_{2i}$. Using this fact, we can rewrite the third equation as

$$1 - p_{11} - p_{21} = 1 - p_{12} - p_{22}$$

This is guaranteed to be true by virtue of equations 10 and 11, so equation 12 adds no additional information. There are only two constraints.

7.4. Example

Consider the following two samples of data.

$$x = [0, 1, 1, 1, 0, 0, 1, 0, 1, 1]$$

$$y = [1, 1, 0, 1, 1, 1, 1, 1, 0, 1]$$

Is it plausible to suppose that both were drawn from a binomial distribution with the same parameters? To find out, we consider two hypotheses: H_0 (which assumes that the two binomial distribution are the same) and H_1 (which makes no such assumption). We'll use a likelihood-ratio test to test H_0 against H_1. The binomial has two parameters, N and p. We'll need to estimate p, but we don't need to estimate N because its value was determined by the experiment that generated the data. Consequently, H_0 involves one constraint: it assumes that $p_x = p_y$, where p_x and p_y are the probabilities of observing a "1" in a single trial of the binomial experiment.

For H_0, we have $N = 20$ observations from the same distribution. To estimate p, we use equation 5, which gives $\hat{p} = 14/20 = 0.7$. Data set x had 6 "1"s, so equation 2 gives its likelihood as $L_{x0} = \Pr[6; 10, 0.7] = 0.2001209$. Data set y had 8 "1"s, so its likelihood is $L_{y0} = \Pr[8; 10, 0.7] = 0.2334744$. Our hypothesis assumes that the two data sets were drawn independently, so the likelihood of the whole data set is the product of L_x and L_y. In other words, the likelihood under H_0 is

$$L_0 = L_{x0}L_{y0} = 0.04672313.$$

For H_1, we estimate p_x and p_y separately for data sets x and y. This gives $\hat{p}_x = 0.6$ and $\hat{p}_y = 0.8$. The corresponding likelihoods are $L_{x1} = \Pr[6; 10, 0.6] = 0.2508227$ and $L_{y1} = \Pr[8; 10, 0.8] = 0.3019899$. The product of these,

$$L_1 = L_{x1}L_{y1} = 0.0757459,$$

is the likelihood under H_1. Note that $L_0 < L_1$. This is always the case, because H_0 is a restricted version of H_1, and those restrictions reduce likelihood. Given these values, we can calculate

$$Z = -2\log(L_0/L_1) = 0.9662903$$

If H_0 were true, Z would be a chi-squared random variable with one degree of freedom, because H_0 imposes one constraint.

Is this large enough to reject H_0? There are several ways to find out. In the back of most introductory statistics books, there are tables describing the cumulative values of the chi-squared distribution function. One can also get the answer from any number of computer packages. Here is the calculation in the R statistical language:

```
> 1-pchisq(0.9662903, 1)
[1] 0.3256071
```

My command is in the first line following R's prompt character (>). The second line contains the answer. Here is the same calculation in Maple:

```
> with(Statistics):
> X := RandomVariable(ChiSquare(1)):
> 1 - CDF(X, 0.9662903);
    0.3256071127
```

Both packages give the same answer: 0.326. This answer tells us that there is a substantial probability (nearly 1/3) of observing two data sets as different as ours even if both were drawn from the same binomial distribution. Consequently, we cannot reject hypothesis H_0.

Appendix A. Answers to Exercises

⋆EXERCISE **4–1** $\binom{3}{1} = 3$, $\binom{3}{2} = 3$, and $\binom{3}{3} = 1$.

⋆EXERCISE **5–1** We want the value of p that makes L as large as possible. This value will also maximize the value of $\ln L$ (since $\ln L$ increases motonically with L), so we can work either with L or with $\ln L$. We will get the same answer in either case. It is easiest to work with $\ln L$, so let us focus on that. Taking its derivative gives

$$d\ln L/dp = x/p - (N - x)/(1 - p)$$

Take a close look at this expression. If $x = 0$, this derivative is negative whatever the value of p. In other words, $\ln L$ is a decreasing function of p, and $\hat{p} = 0$. If $x = 1$, the derivative is always positive, and $\hat{p} = 1$. When $0 < x < N$, the $\ln L$ function reaches a maximal value at the value of p that makes the derivative equal 0. To find this maximal value, set $d\ln L/dp = 0$ and solve for p. This gives equation 5.

We also need to make sure that this is indeed a maximum point, rather than a minimum or a saddle point. There are several ways to do

this. For me, it is easiest simply to examine the expression for $d \ln L/dp$. The first term divides by p, so it will be very large when $p \approx 0$. Since this term is positive, $d \ln L/dp > 0$ when p is small. For similar reasons, the second term will be large in magnitude when $p \approx 1$. Since this term is negative, $d \ln L/dp < 0$ when p is large. This implies that as p increases from 0 to 1, $d \ln L$ first increases and then decreases. The point at which the derivative is zero can only be a local maximum. If you find this argument hard to follow, then use the second derivative test: if the second derivative of $\ln L$ is negative, then $\hat{p}$ is a local maximum. Here is some Maple code that does this calculation:

```
> # Define lnL
> lnL := Const + x*log(p)
> + (N-x)*log(1-p);
    lnL := Const + x ln(p)
    + (N - x) ln(1 - p)

> # Solve d lnL / dp = 0 to find the maximum.
> phat := solve(diff(lnL, p) = 0, p);
    phat := x/N

> # Calculate the second derivative.
> simplify(subs(p=phat, diff(lnL, p, p)));
          3
         N
    ----------
    (-N + x) x
```

If $x < N$, the entire expression is negative and $\hat{p}$ is at a local maximum. Thus, in this case $\hat{p}$ is a maximum likelihood estimator of p. When $x = N$, however, the second derivative is undefined, so the second-order condition provides no information.

⋆EXERCISE **6–1** The estimated sampling variance is 0.006375 for experiments 1 and 2, and 0.008 for experiments 3, 4, and 5.

⋆EXERCISE **6–2** Begin by defining `lnL` and `phat` as in example 5–1 above. Then calculate the second derivative of `lnL`.

```
> # d2 is the 2nd derivative of lnL with
> # respect to p
> d2 := diff(lnL, p, p);
              x      N - x
    d2 := - ---- - --------
             2           2
            p     (1 - p)
```

To take the expectation of this expression, replace x with its expected value Nx.

```
> # Ed2 is the expectation of d2
> Ed2 := simplify(subs(x = N*p, d2));
                N
    Ed2 := ----------
           p (-1 + p)
```

Students who have studied probability theory will recognize that this substitution is justified because `d2` is a linear function of the random variable x. The rest of you must take it on faith. Now, equation 7 gives

```
> # v1 is the sampling variance of phat
> v1 := -1/Ed2;
              p (-1 + p)
    v1 := - ----------
                  N
```

To use this formula with data, we would have to substitute $\hat{p}$ for p, and this would give equation 8.

The next step is to use equation 9, which gives

```
> # v2 is also the sampling variance of phat
> v2 := simplify(-1 / subs(p=phat, d2));
              (-N + x) x
    v2 := - ----------
                   3
                  N
```

But $x = N\hat{p}$, so this can be re-expressed as:

```
> phat := 'phat';
    phat := phat

> v2 := simplify(subs(x = N*phat, v2));
              (-1 + phat) phat
    v2 := - ----------------
                    N
```

This is also equivalent to equation 8.

⋆EXERCISE **6–3** In that example $\hat{p} = 0.85$ and the sampling variance was 0.006375. The standard error is thus $\sqrt{0.006375} \approx 0.08$, and the 95% confidence interval is $[0.69, 1.01]$.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Ziheng Yang. *Computational Molecular Evolution*. Oxford University Press, Oxford, 2006. ISBN 0198567022.