Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Lecture 17: What is Statistical Inference?

Relevant textbook passages:
Larsen–Marx [7]: Section 5.1, [5.2]

17.1 Probability versus statistics

Probability theory is a branch of pure mathematics, and could be considered to be a subfield of positive operator theory, but that would be wrong.[1] The concepts of conditioning and independence are what separate it from pure measure theory. While it is, in one sense, just the study of the consequences of a few axioms and definitions, the questions addressed are motivated by applied concerns.

Statistics, especially "mathematical statistics," uses the tools of probability theory to study data from experiments (both laboratory experiments and "natural" experiments) and the information the data reveal. Probability theory investigates the properties of a particular probability measure, while the goal of statistics is to figure which probability measure is involved in generating the data. To a statistician, the "state of the world" is the measure, not the state in the sense that we used it earlier. Indeed, "Statistics means never having to say you're certain."

17.2 The subject matter of statistics

Larsen–Marx [7]: Chapter 5

Description. Descriptive statistics include such things as sample mean, sample medium, sample variance, interquartile range. These provide a handle to think about your data. This is the material that is often taught in "business statistics" courses, and is perhaps the reason my colleague David Politzer dismisses statistics as mere "counting."

One aspect of descriptive statistics is data "exploration" or "data mining." The ubiquity of machines that thirty years ago would have been called supercomputers has led to an entirely new discipline of "data science," much of which comes under the heading of descriptive statistics.

Many of the methods of data science have been neglected by traditional statisticians. Leo Breiman, whose credentials as a probabilist and mathematical statistician are impeccable, describes "two cultures" [1] in statistics, and says in his abstract:

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is

[1] Wrong as in Richard Nixon's comment to Bob Haldeman and John Dean regarding raising $1 million to pay off the Watergate defendants. Nixon's response was, according to Dean, "Still, there's no problem in raising a million dollars. We can do that, [pause] but it would be wrong." (This quote may be an urban legend but see, e.g., this article.)

> to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Visualization. "Data visualization" is a hot topic these days. The idea of using diagrams to represent data is surprisingly recent. According to Wikipedia, in 1765 Joseph Priestley (of oxygen fame) created the first timeline charts. These inspired the Scottish engineer and economist William Playfair to invent the line graph and bar chart in 1786. He also invented the pie chart in 1806. Today these tools are familiar and taught to elementary schoolchildren, but at the time they were controversial. According to the *The Economist*, Dec. 19, 2007,

> Playfair was already making a leap of abstraction that few of his contemporaries could follow. Using the horizontal and vertical axes to represent time and money was such a novelty that he had to explain it painstakingly in accompanying text. "This method has struck several persons as being fallacious", he wrote, "because geometrical measurement has not any relation to money or to time; yet here it is made to represent both."

Another early adopter of charts and graphs was Florence Nightingale (of nursing fame) in her analysis of disease and its relation to sanitation. In 1861, William Farr, the Compiler of Abstracts in the General Registry Office (who compiled the first mortality tables), wrote to her complaining about her use of charts and graphs, "We do not want impressions, we want facts. You complain that your report would be dry. The dryer the better. Statistics should be the dryest of all reading." (*The Economist*, op. cit.)

In the dark ages of data science (before 2000), John Tukey [13] invented a diagram for exploring data, called the **Box Plot** or the **Box and Whisker Plot**.[2] Box plots are still used in almost every presentation I have seen in neuroscience.

There are several kinds of box plots: whiskers at max and min; whiskers at quartiles $\pm$ 1.5$\times$ interquartile range. See [8] or the Wikipedia article for descriptions of other kinds of Box Plots.

Herman Chernoff [2] introduced **face diagrams** as a way to visualize data and identify outliers or subgroups. Each observation consists of a vector of measurements that are then used to determine the characteristics of a human face. Since humans are generally adept at facial recognition (unless they suffer from prosopagnosia), this is a potentially useful technique for data exploration. See Figure 17.1 for an example.

Advances in computer graphics have led to entirely new tools for data visualization. There is a course, **Ay 119, Methods of Computational Science** that deals with data visualization and management. Caltech hosted a conference on data visualization in 2013 (the program is here) and may do so again.

There are number of excellent books on ideas for presenting data including Tufte [10, 11, 12] and Cook [3]. The Caltech course **BEM/Ec 150: Business Analytics** devotes a session to data visualization and the cognitive neuroscience underlying effective presentation.

It is also possible to use sound to "audibilize" data. My colleague Charlie Plott has turned data on double oral auctions into sounds, and you can actually hear price "bubbles" form and then collapse. Here is a link to a QuickTime video. The horizontal access denotes time, and the vertical axis denotes price. Buyers and Sellers are bidding on securities with a random payout. The bidders know the distribution. The sounds represent bids, asks, and transactions. The pitch represents the price level. There are two sloping lines. The lower line represents the expected value, and the upper line represents the maximum possible value. Once transaction take place above the upper line, buyers are paying more than the security could possibly be worth. That is, there is a price bubble. You can hear it crash. The crash is foreshadowed by some low rumbling, caused by sellers hoping to unload their overvalued inventory.

[2] Tukey is also the co-inventor of the Fast Fourier Transform [4], which was selected as one of the Top Ten Algorithms of the 20$^{\text{th}}$ Century by *Computing in Science & Engineering* [5], a joint publication of the American Institute of Physics and the IEEE Computer Society.

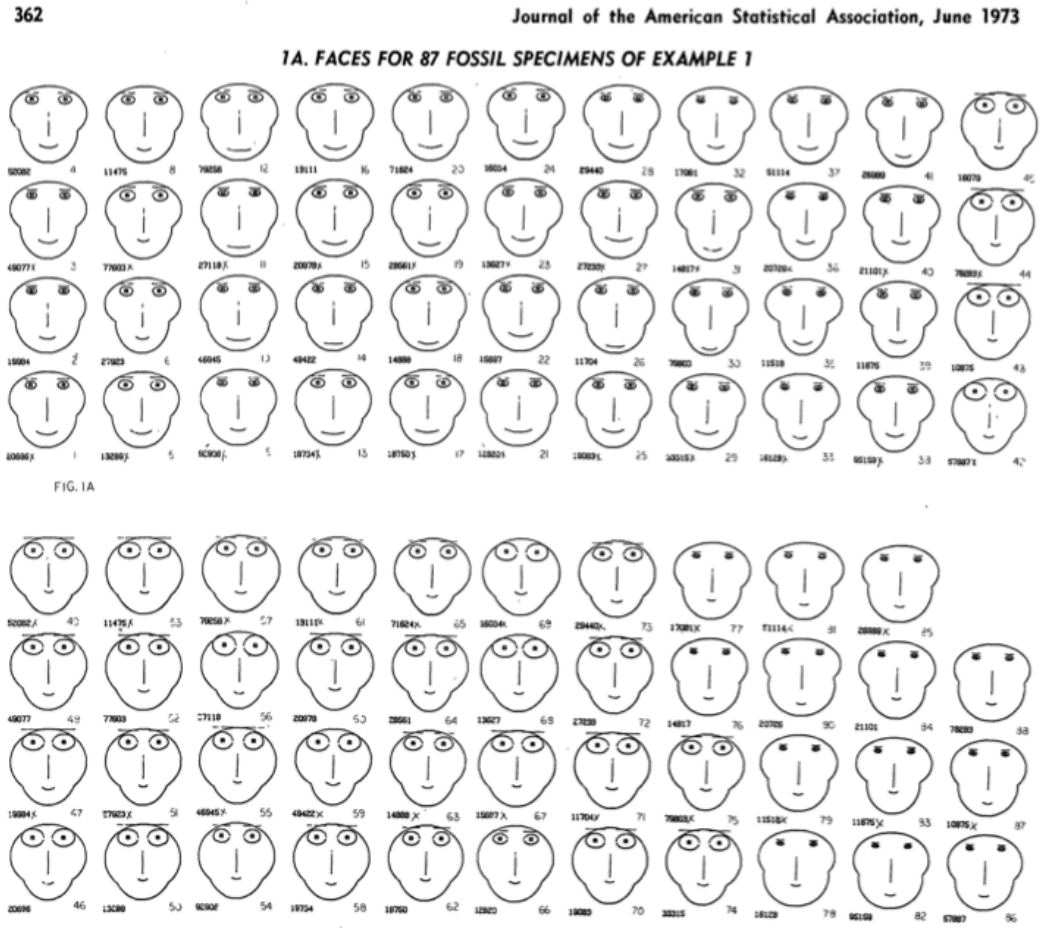

Figure 17.1. Chernoff faces. They represent data on fossils from the Eocene Yellow Limestone Formation in northwestern Jamaica. See Chernoff [2, § 2.1] for more details.

Estimation. Statisticians usually assume there is a **data generating process** (**dgp**) that stochastically generates the data, and typically is governed by a probability distribution governed by a small number of **parameters**. The goal is to **identify** or **estimate** the parameters from the information in the data.

Sometimes the number of parameters may not be small, and **nonparametric methods** may be used.

A nice discussion of estimation and its role in data analysis can be found in Brad Efron's [6] 1981 Wald Memorial Lecture.

Hypothesis testing. Once the parameters of the dgp have been estimated, we might ask how much confidence should we put in these estimates. This is the object of **hypothesis testing**, which may address such questions as, How confident are we that the parameter really nonzero?

Hypothesis testing also addresses the choice of model for the data generating process. Breiman [1] complains that most of the dgps considered by traditional statistics are too simplistic, and that it is arrogant of statisticians to think that they can sit in their armchairs and imagine the form of the dgp that generates real data sets.

Prediction. Once we have the parameters of the dgp, we can use it to make predictions about future behavior of the dgp. We also care about how reliable these predictions can be expected to be.

17.3 Statistics and Estimation

We start with the idea that we have data generated by some dgp, which has unknown parameters. A **statistic** is function of the observable data, and not of the unknown parameters.

Examples:

- The number T of Tails observed in N coin tosses. The pair (N, T) is a statistic, since it something we can observe, measure, and know. The probability that a Tails will occurs is not observable, so is not a statistic.

- The list of how many World Series lasted 4, 5, 6, and 7 games is a statistic. The probability that a given team wins is not observable.

- The number of observed arrivals in a time of a given length is a statistic, the arrival rate λ in Poisson process is not observable.

17.3.1 Estimation in the abstract

An **estimator** is a statistic that takes on values in the set of parameter values. That is, if $\mathcal{X}$ is the set of possible values of the observed outcomes of a random experiment, that is, the **sample space**,[3] and Θ is the set of possible parameter values for the dgp modeling the experiment, then

I need to find better terminology.

an estimator is a function

$$T\colon \mathcal{X} \to \Theta.$$

An **estimate** is the value of an estimator at a particular datum.

[3] There is an unfortunate ambiguity in the terminology here. A random variable X has been defined as a function on an underlying sample space, and for statistical purposes the sample space of an experiment is actually the set of values of the random variable X.

17.4 The Likelihood Function

To be a little more concrete, suppose we want to estimate the value of the mean of a distribution, when we know that it is Normal$(\mu, 1)$ where $\mu = 0$ or $\mu = 1$. If x is the outcome of the experiment, how do we decide whether

$$T(x) = 0 \quad \text{or} \quad T(x) = 1?$$

Consider Figure 17.2, which shows the probability densities of the two normals. For $x = -1.5$

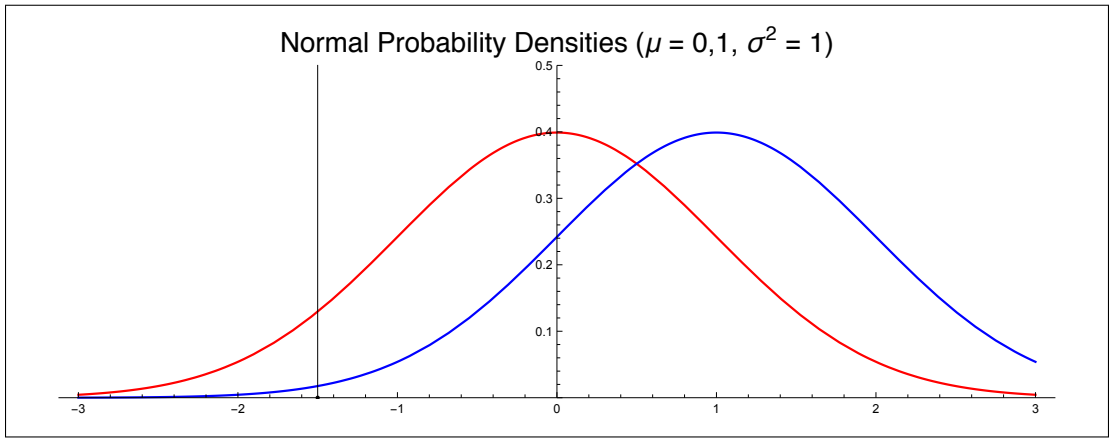

Figure 17.2. A two point parameter set.

which μ would you choose? For $x = 3$? Intuitively, it is more believable or more likely that when $x = -1.5$ that we should estimate μ to be zero, and when $x = 3$ we should estimate μ to be one. R. A. Fisher formalized this intuition by introducing the **likelihood function**.

Larsen–Marx [7]: § 5.2

More generally, there is a data generating process that stochastically selects an element of $\mathcal{X}$, the set of possible observations or experimental outcomes, and Θ is the set of possible parameter values for the dgp.

The function $f(x;\theta)$ is the probability that x is the observation when the parameter θ is generating the data, or possible the density of x.

17.4.1 Example • If the experiment is to toss a coin independently N times, x might be the number of Tails. If θ is the probability of a Tail, then

$$f(x;\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x}.$$

• If the experiment is to select a real number from a Normal distribution with mean θ and variance 1, then

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2}.$$

□

These are just familiar probability mass functions and densities.

Now we turn everything on its head and define the **likelihood function** by

$$L(\theta;x) = f(x;\theta).$$

What? Why?

When we have used the term "more likely" in the past in this course, we usually meant "more probable." Is that what we mean when discussing likelihood? Most statisticians would say no, we are not talking about the probability that the θ has one value or another. Most would say that θ is fixed but unknown. Then what interpretation are we to give to "likelihood?" R. A. Fisher developed his ideas about statistics based on the notion of likelihood, which he insisted was not probability. This led to a feud with Jerzy Neyman and Egon Pearson over the proper interpretation of a number of statistical tests and methods. If it seems odd to you that a mathematics course there would be such foundational disputes, you are right. It is odd. But statistics is not solely mathematics, it has elements of philosophy science embedded in it.

To make things more controversial, there is a camp of statisticians, usually referred to as Bayesians, who are quite willing to talk about θ as if it were random. That is they will talk about the probability distribution of θ. But typically, they do not believe that the value of θ is the outcome of a random experiment. Instead they take the position that the only way to sensible talk about unknown values is probabilistically. They view the probabilities as representing degrees of belief about the unknown value of θ. But the calculations they do are exactly like those that you have done in the various two-stage urn problems you have seen, where an urn is selected at random and ball is randomly drawn from the urn. It's just that in the real world, we never find out from which urn the ball has been drawn.

To make matters, more obscure, your textbook, Larsen–Marx [7, Comment, p. 284], tells you not to think of L as a function of x (even though it is).

One reason to justify thinking about the likelihood function this way is that it gives a general method for constructing estimators that may be a good method. That is, **maximum likelihood estimators** often have desirable properties. I'll get more into the properties later on, but frequently they include the properties of consistency, unbiasedness, efficiency, and asymptotic normality.

But it is a bit too early to get into such abstract ideas without some grounding in a real (but very simple) example.

17.5 An Example of Maximum Likelihood Estimation

17.5.1 Example (Binomial) Suppose we observe k successes in n independent trials. What is the maximum likelihood estimator of p? The likelihood function is just

$$P\,(k \text{ successes in } n \text{ trials}) = L(p; n, k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

If $k = 0$, this reduces to $\binom{n}{k}(1-p)^n$, which is clearly maximized when $p = 0$. When $k = n$, it reduces to $\binom{n}{k}p^n$, which is maximized at $p = 1$. When $0 < k < n$, the **first order condition for a maximum** of this is that $d/dp = 0$, or

$$\binom{n}{k}\left(kp^{k-1}(1-p)^{n-k} - (n-k)p^k(1-p)^{n-k-1}\right) = 0.$$

For $0 < p < 1$, we may divide both sides by $\binom{n}{k}p^{k-1}(1-p)^{n-k-1}$ to get

$$k(1-p) - (n-k)p = 0 \implies k - pk - np - kp = 0 \implies k - np = 0 \implies p = \frac{k}{n}.$$

Thus the maximum likelihood estimator of p when the data indicate k success in n trials is simply

$$\hat{p} = \frac{k}{n}.$$

Now one of the tricks that statisticians employ is that they will maximize the logarithm of the likelihood function rather than the likelihood function itself. There are a few reasons for

this. The first is that theoretically it doesn't make any difference. For if x^* maximizes $f(x)$, then it also maximizes $\log\big(f(x)\big)$. Also, likelihoods are often very small positive numbers with lots of leading zeroes. Taking logs puts them into more manageable numerical range. Finally, likelihood functions often involve products, and taking logs can make expressions simpler.

For instance, in our Binomial example,

$$\log\big(L(p)\big) = \log \binom{n}{k} + k \log p + (n-k)\log(1-p).$$

Differentiating with respect to p gives

$$\frac{d}{dp}\log\big(L(p)\big) = \frac{k}{p} - \frac{n-k}{1-p}$$

and setting this derivative to zero gives

$$\frac{k}{p} - \frac{n-k}{1-p} = 0 \implies k(1-p) - (n-k)p = 0 \implies \hat{p} = \frac{k}{n}.$$

□

17.6 Application to the Coin Tossing Experiment

Here are the data from 2016[4] and all years combined:

	Number			Percent	
Year	Sample size	Heads	Tails	Heads	Tails
2016	23,808	11,932	11,876	50.118%	49.882%
All	109,184	54,647	54,537	50.050%	49.950%

Figure 17.3 shows the graph of the likelihood function for x = Probability of Tails for the pooled sample, as produced by the R command

```
curve(dbinom(54537,109184,x))
```

That's not very informative, so let's replot it. See Figure 17.4.

```
curve( dbinom(54537,109184,x, log=TRUE),xlim=c(.499,.501),
          ylab="Log likelihood",xlab="p" )
```

Notice that I made several changes to the R code. The first was that I added `log=TRUE` to the `dbinom` function. This is option that plots the logarithm of the function instead of it actual value. The reason for this that the likelihood function varies tremendously, so it is often easier to deal with the log-likelihood, both numerically and visually. Just remember that since the likelihood is $\leqslant 1$, that its logarithm is negative.

I also changed the axes labels (`xlab`, `ylab`), and the range (`xlim`) over which to plot.

It is clear from the pictures that the maximum occurs a little above 0.5, but where? Let's try numerical optimization. (In this case it's silly, since we have the formula, but it's instructive nevertheless.)

In R, the `optimize` command will minimize a function. To get it to maximize, use the `maximize=TRUE` option.

```
L = function (x) dbinom(54537,109184,x)
optimize(L, interval=0:1, maximum=TRUE)
```

which produces the output

[4] These numbers differ slightly from those in Lecture 3, because I accepted some late submissions.

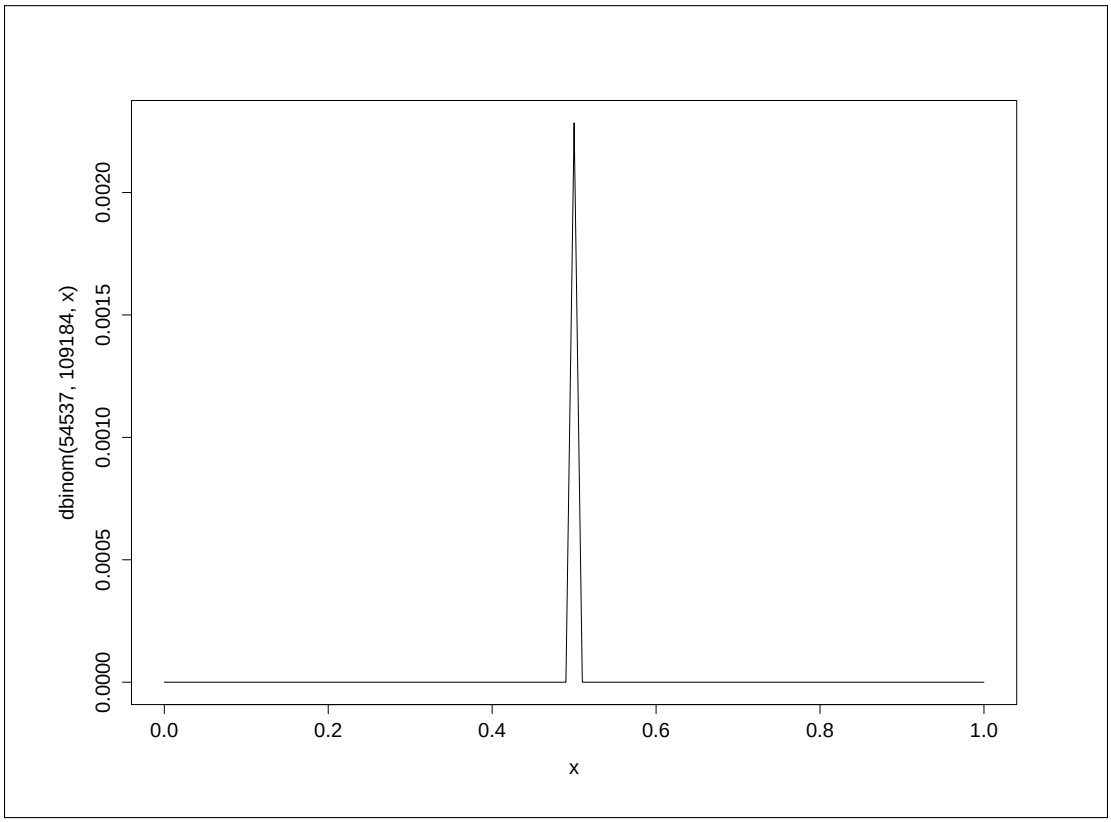

Figure 17.3. The coin tossing likelihood.

```
$maximum
[1] 0.9999339

$objective
[1] 0
```

which is not at all correct. **Welcome to the world of numerical computation. You can't take the word of a computer as the truth.** Lets' try to figure out the problem. It may help to know that the function `optimize` really does. The help says, "The method used is a combination of golden section search and successive parabolic interpolation, and was designed for use with continuous functions." Not very useful. But look at Figure 17.3. The likelihood function is pretty flat almost everywhere except very near 0.50. Perhaps the algorithm is getting "stuck" in a flat spot. Let's try searching where we think the answer might be.

[The economist streetlight joke.]

```
optimize(L, interval=c(0.4,0.6), maximum=TRUE)
```

produces the output

```
$maximum
[1] 0.4994949

$objective
[1] 0.00241468
```

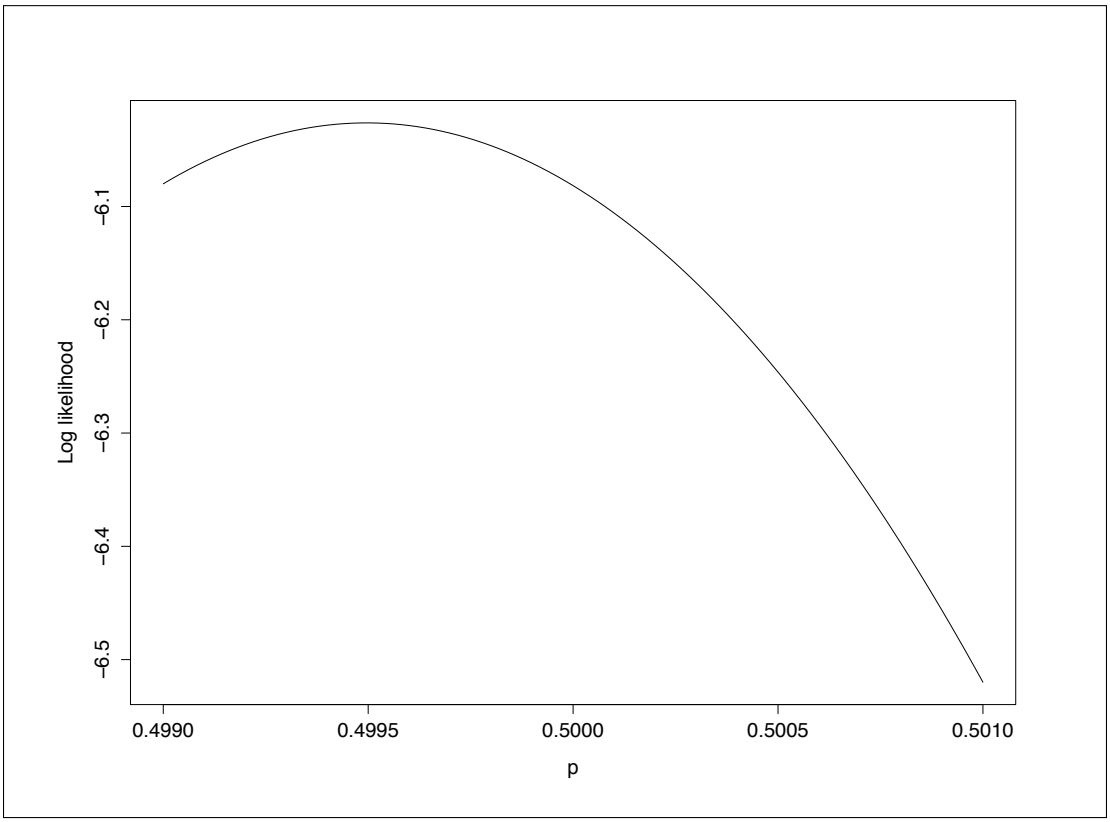

Figure 17.4. The coin tossing log likelihood, zoomed in.

which looks pretty good.

Another tactic worth trying is what I mentioned earlier: Take the log-likelihood function.

```
L = function (x) dbinom(29979,59904,x, log=TRUE)
optimize(L, interval=0:1, maximum=TRUE)
```

which produces the output

```
$maximum
[1] 0.499501

$objective
[1] -6.026193
```

In this case it hardly matters.

If this seems ad hoc and unscientific, I apologize, but numerical methods are the subject of an entire course here, **ACM 106 abc, Introductory Methods of Computational Mathematics**.

The moral of this overkilled numerical analysis of a trivial problem is that you cannot blindly accept what the computer tells you. You have to look at the output and see if it makes sense.

With any numeric results from reputable software, you should follow the Russian proverb, adopted by Ronald Reagan, *Доверяй, но проверяй* [Trust, but verify]. [a]

[a] See, e.g., http://en.wikipedia.org/wiki/Trust,_but_verify.

17.7 The World Series, Once Again

I am not a fanatical baseball fan(atic) (I never even played Little League, only intramural softball), but Frederick Mosteller's analysis of the World Series [9] is a wonderful introduction to parameter estimation. So much so that your homework assignment this week will be to redo his analysis with another 60+ years of data.

Mosteller uses three methods for estimating the average probability that the better team wins any given game in the World Series. They are the method of moments, maximum likelihood, and minimum chi-square estimation. Naturally, in order to apply any of these methods, one must make certain assumptions about the nature of the process that generates the data, and these assumptions may or may not be true. But that is true of any scientific endeavor. We are always making assumptions about what may be neglected, and what matters.

Mosteller [9, p. 370] puts it this way (emphasis mine):

> We have emphasized the binomial aspects of the model. The twin assumptions needed by a binomial model are that throughout a World Series a given team has a fixed chance to win each game, and that the chance is not influenced by the outcome of other games. It seems worthwhile to examine these assumptions a little more carefully, because any fan can readily think of good reasons why they might be invalid. Of course, strictly speaking, *all such mathematical assumptions are invalid when we deal with data from the real world. The question of interest is the degree of invalidity and its consequences.*

This is one of my favorite quotations about applied science.

Be sure to update this each year.

The first "World Series" was played in 1903. Since then there has been a World Series every year except 1904 (when the NL champ refused to play the AL champ) and 1994 (the strike year). That makes 111 Series. In 1903, 1919, 1920, and 1921 the Series had a best-of-9 games format, and in 1907, 1912, and 1922 there was a tie game (?!) in each Series. That leaves 107 (as of January, 2016) "normal" best-of-7 Series. (The 1919 "Black Sox" scandal was a best-of-9 Series.)

So how do we get a handle on p, the average probability that the better team wins a game?

The answer lies in the length of the series, or equivalently, the number of games that the series winner loses. If the better team always won, then all best-of-7 game series would last only four games. As the probability gets closer to 1/2, one would expect more 7 game Series. The likelihood function depends on p and on N_k where N_k is the number of Series that last $4 + k$ games, $k = 0, \ldots, 3$.

Bibliography

[1] L. Breiman. 2001. Statistical modeling: The two cultures. *Statistical Science* 16(3):199–215. http://www.jstor.org/stable/2676681

[2] H. Chernoff. 1973. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association* 68(342):361–368. http://www.jstor.org/stable/2284077

[3] G. Cook, ed. 2013. *The best American infographics 2013*. New York: Houghton Mifflin Harcourt Publishing.

[4] J. W. Cooley and J. W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation* 19(90):297–301. http://www.jstor.org/stable/2003354

[5] J. Dongarra and F. Sullivan. 2000. Guest editors' introduction: The top 10 algorithms. *Computing in Science & Engineering* 2(1):22–23. DOI: 10.1109/MCISE.2000.814652

[6] B. Efron. 1982. Maximum likelihood and decision theory. *Annals of Statistics* 10(2):340–356. http://www.jstor.org/stable/2240671

[7] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[8] R. McGill, J. W. Tukey, and W. A. Larsen. 1978. Variations of box plots. *American Statistician* 32(1):12–16. http://www.jstor.org/stable/2683468

[9] F. Mosteller. 1952. The world series competition. *Journal of the American Statistical Association* 47(259):355–380. http://www.jstor.org/stable/2281309

[10] E. R. Tufte. 1983. *The visual display of quantitative information*. Cheshire Connecticut: Graphics Press.

[11] ——— . 1990. *Envisioning information*. Cheshire Connecticut: Graphics Press.

[12] ——— . 2006. *Beautiful evidence*. Cheshire Connecticut: Graphics Press.

[13] J. W. Tukey. 1977. *Exploratory data analysis*. Addison-Wesley.