Caltech Department of Mathematics

Lecture 18: Introduction to Estimation

Relevant textbook passages:

Larsen–Marx [12]: Section 5.1, [5.2]

18.1 Probability versus statistics

Probability theory as a branch of pure mathematics could be considered to be a subfield of positive operator theory, but that would be misleading. The concepts of conditioning and independence give probability theory a separate identity. 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."

18.2 The subject matter of statistics

Larsen–Marx [12]: 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 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 1858 she introduced a type of chart now known as "Nightingale's Rose" or "Nightingale's Coxcomb," in her monograph, "Notes on matters affecting the health, efficiency and hospital administration of the British army." (See Figure 18.1.) In the same year she became the first female fellow of

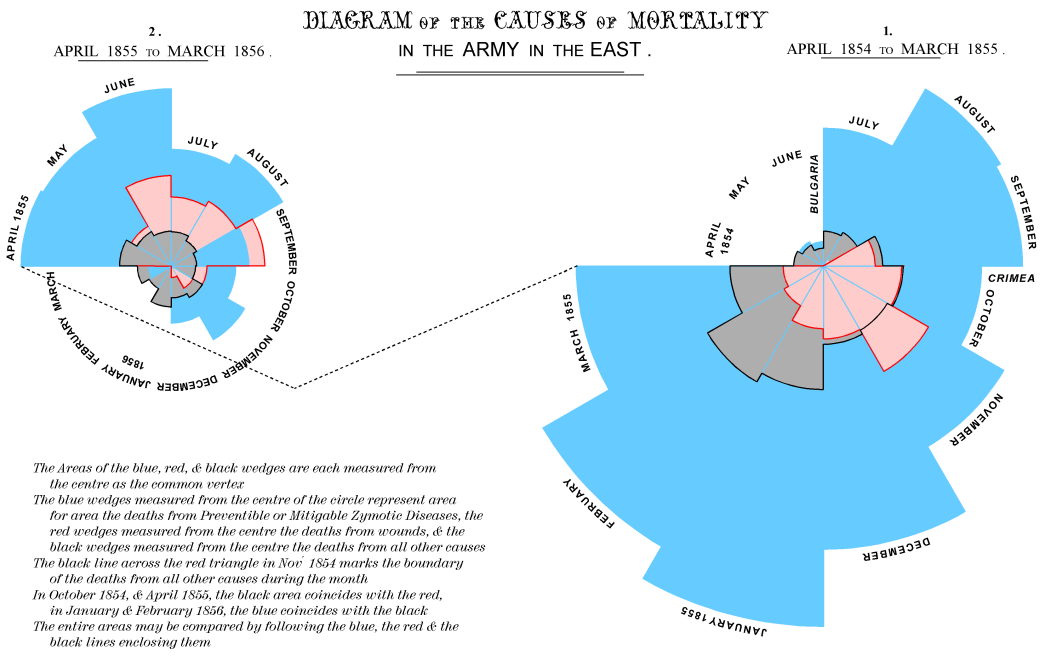


Figure 18.1. An example of Florence Nightingale's coxcomb chart.

the Statistical Society of London (now the Royal Statistical Society). 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.)

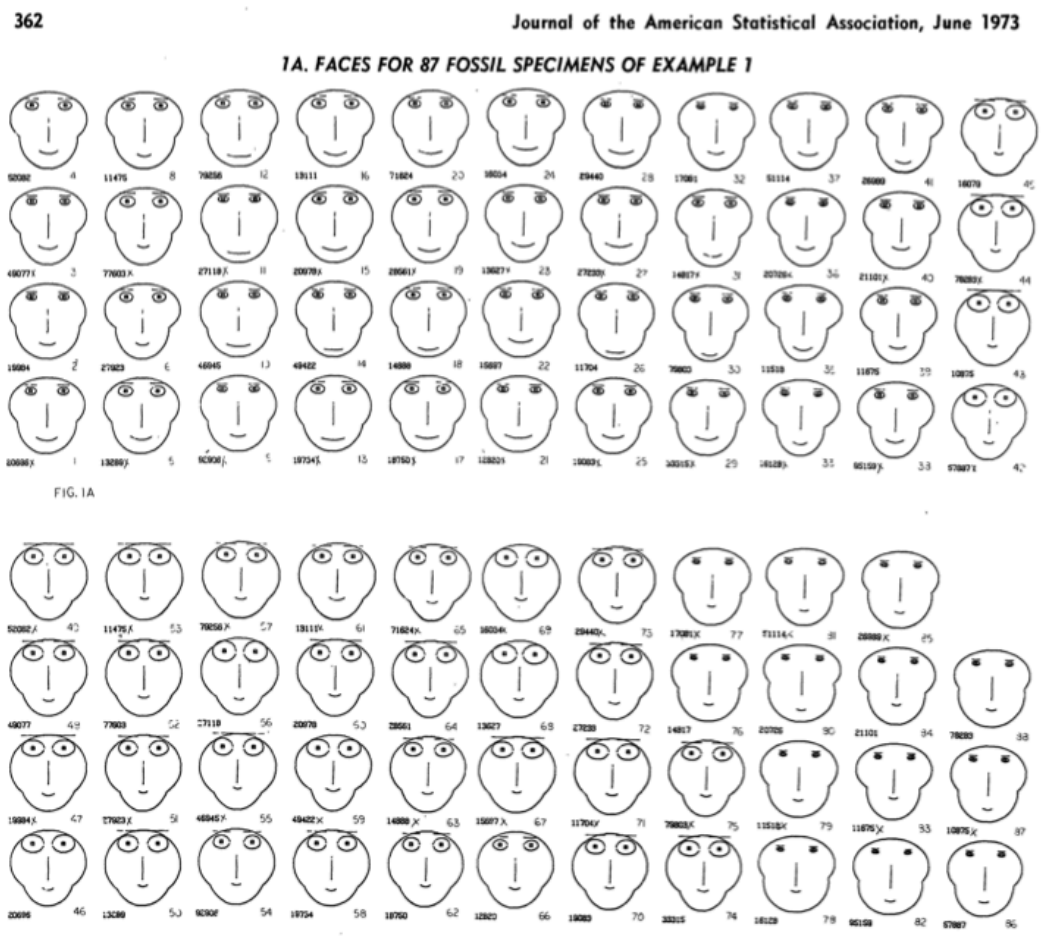
In the dark ages of data science (before 2000), John Tukey [19] invented a diagram for exploring data, called the **Box Plot** or the **Box and Whisker Plot**.[1] Box plots are still used in almost

[1] Tukey is also the co-inventor of the Fast Fourier Transform [5], which was selected as one of the Top Ten Algorithms of the 20th Century by *Computing in Science & Engineering* [8], a joint publication of the American Institute of Physics and the IEEE Computer Society.

every presentation I have seen in neuroscience.

There are several kinds of box plots: whiskers at max and min; whiskers at quartiles ± 1.5× interquartile range. See [14] or the Wikipedia article for descriptions of other kinds of Box Plots.

Herman Chernoff [3] 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 18.2 for an example.


362 Journal of the American Statistical Association, June 1973

1A. FACES FOR 87 FOSSIL SPECIMENS OF EXAMPLE 1

FIG. 1A

Figure 18.2. Chernoff faces. They represent data on fossils from the Eocene Yellow Limestone Formation in northwestern Jamaica. See Chernoff [3, § 2.1] for more details.

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 [16, 17, 18] and Cook [4]. 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 axis 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 transactions 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.

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.

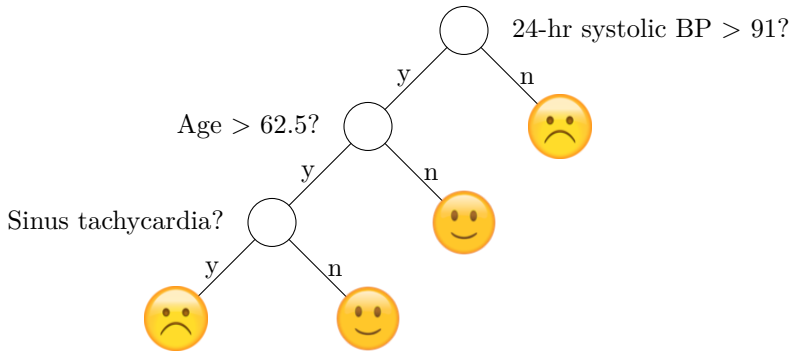
A nice discussion of estimation and its role in data analysis can be found in Brad Efron's [9] 1981 Wald Memorial Lecture.

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 is 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.

Classification Breiman et al. [2] start their book on regression trees by describing the following classification problem. Given a heart attack patient admitted to the UC San Diego Medical Center, classify them as those will not survive thirty days, and those who will. The best classifier in this case turned out be remarkably simple, and requires answers to three questions. Here is the classification tree.



A classic use of classification techniques has been the classification of fossils and bones, according to say species, or gender.

18.3 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 a Poisson process is not observable.

18.4 The Likelihood Function

Larsen–Marx [12]: § 5.2

A mathematical model of a **data generating process** has three components: $\mathcal{X}$, the set of possible observations or experimental outcomes, Θ the set of possible parameter values indexing probability measures on X, and a function $f\colon \mathcal{X} \times \Theta \to \boldsymbol{R}_+$.

The value $f(x;\theta)$ is either the probability mass function or the probability density that x is the observation when the parameter θ is generating the data.

18.4.1 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.

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**,[2] 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

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

[2] 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.

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

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

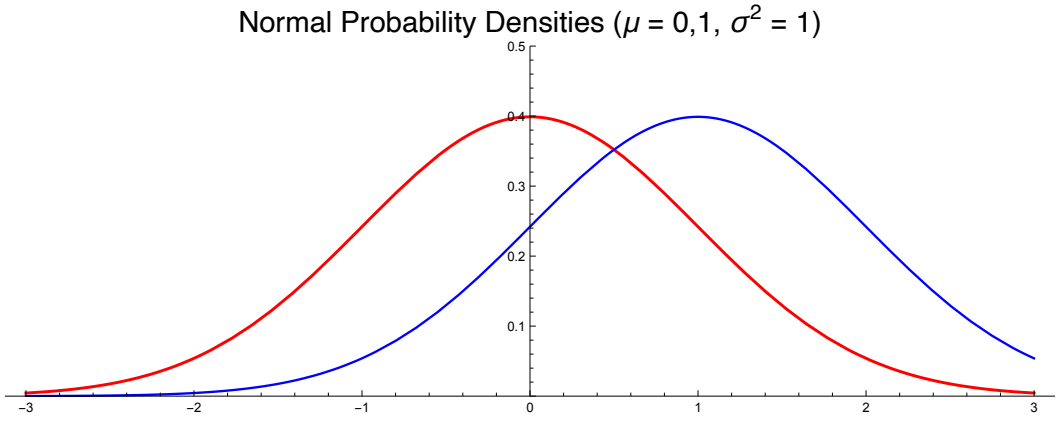


Figure 18.3. A two point parameter set.

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**.

The **likelihood function** for a dgp is defined by

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

The method of **Maximum Likelihood Estimation** gives a general method for estimating θ, namely, given a datum x, choose as the estimate

$$\hat{\theta}_{\text{MLE}} \text{ to maximize } L(\theta; x).$$

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 in 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 [12, 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.

18.5 An Example of Maximum Likelihood Estimation

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

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

Note that the leading term $\binom{n}{k}$ is positive and independent of p, and so it has no relevance to MLE, and it is often convenient to omit it, and just write

$$L(p; k) \propto p^k (1-p)^{n-k}.$$

where the symbol $\propto$ is read "is proportional to."

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

$$\hat{p}_{\text{MLE}} = \frac{k}{n}.$$

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,

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

Differentiating with respect to p gives

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

and setting this derivative to zero gives

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

□

18.6 Application to the Coin Tossing Experiment

Here are the data from 2020 and all previous years combined:

	Number			Percent	
Year	Sample size	Heads	Tails	Heads	Tails
2020	25,600	12,982	12,618	50.711%	49.289%
All	212,480	106,432	106,048	50.09%	49.91%

So the likelihood function for p given the datum 106,048 Tails in 212,480 tosses is

$$L(p) = \binom{212480}{106048} p^{106048}(1-p)^{106432}.$$

Figure 18.4 shows the graph of the likelihood function for x = Probability of Tails for the pooled sample, as produced by the R command

```
curve(dbinom(106048,212480, x))
```

Actually, to save it to a `.pdf` file, you want to do something like this:

```
pdf("CoinTossTailsLikelihoodAll2020.pdf")
curve(dbinom(106048,212480, x))
dev.off()
```

That's not very informative, so let's replot it. See Figure 18.5.

```
pdf("CoinTossTailsLogLikelihoodAll2020.pdf")
curve( dbinom(106048,212480, x, log = TRUE), xlim = c(.498,.502),
ylab = "Log likelihood", xlab = "p" )
dev.off()
```

Notice that I made several changes to the R code. The first was that I added `log = TRUE` to the `dbinom` function. This option 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 $\leqslant 1$, its logarithm is negative.

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

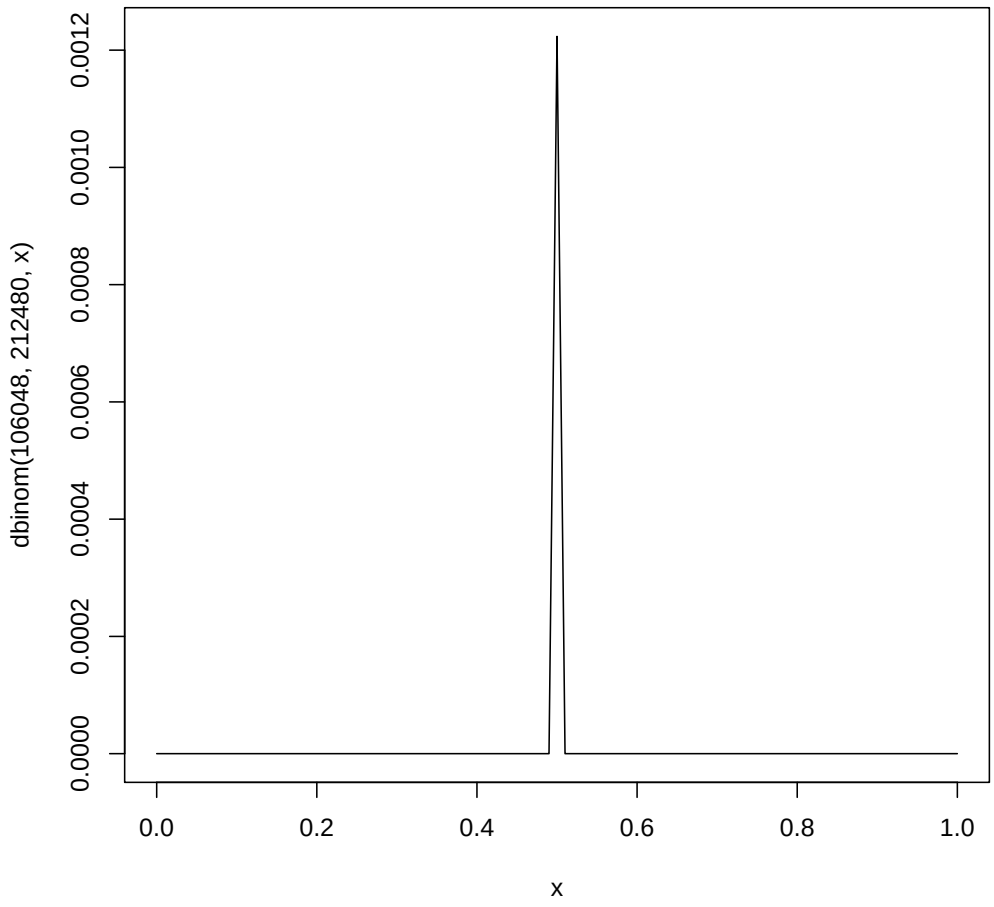


Figure 18.4. The coin tossing likelihood.

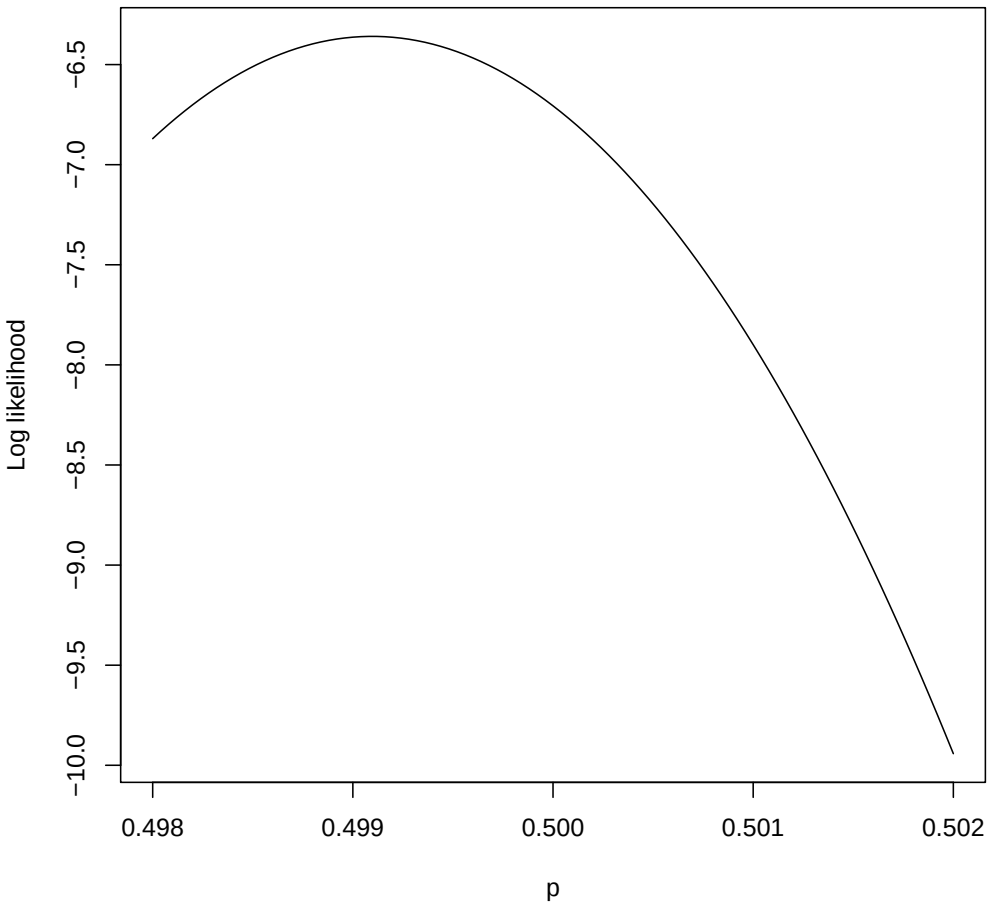


Figure 18.5. The coin tossing log likelihood, zoomed in.

It appears from the pictures that the maximum occurs a near 0.5, but where? We already know that the maximum likelihood estimate is $106048/212480 \approx 49.91\%$, but for many estimation problems, we need to find the maximum using numerical methods. Fortunately R, Mathematica, Matlab, and similar programs all have built-in commands to find maxima numerically. Let's try one out.

In R, the `optimize` command will minimize a function. To get it to maximize, use the `maximize = TRUE` option. But first you have to define the function that you want to maximize:

```
L = function (x) dbinom(106048,212480, x)
optimize(L, interval = 0:1, maximum = TRUE)
```

which produces the output

```
$maximum
[1] 0.9999339

$objective
[1] 0
```

This tells you that R computed the maximizer to be $p = 0.9999339$ and that the resulting value of the likelihood is 0. This as you know, is not at all correct.

What about Mathematica? If we run the equivalent

Update this annually.

```
l[p_] := PDF[BinomialDistribution[212480, p], 106048]
NMaximize[{l[p], 0 <= p <= 1}, p]
```

in Mathematica 12.0, it produces

```
{0., {p -> 0.641915}}
```

That is, Mathematica tells us the MLE estimate is $p = 0.64$. This is better, but still ridiculously incorrect.

Welcome to the world of numerical computation.

You can't always take the word of a computer as the truth.

Lets' try to figure out the problem. In the case of R, it may help to know what 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.

I personally did not find that very useful, but it had the word "search" in the description. That suggests the algorithm is searching around for a maximum. But look at Figure 18.4. 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] 0.499102

$objective
[1] 0.001730914
```

which looks pretty good.

Let's check out Mathematica.

```
NMaximize[{l[p], 0.4 <= p <= 0.6}, p]
```

produces

```
{0., {p -> 0.528383}}
```

which is still abysmal.

Another tactic worth trying is what I mentioned earlier—take the log-likelihood function:

```
L = function (x) dbinom(106048, 212480, x, log = TRUE)
optimize(L, interval = 0:1, maximum = TRUE)
```

which produces the output

```
$maximum
[1] 0.4991049

$objective
[1] -6.359123
```

This is a tiny bit different, but it hardly matters. Note here that the value of the objective function is negative. That's because the objective is the logarithm of the likelihood, and the likelihood here is less than one.

Naively taking the logarithm in Mathematica causes it to blow up:

```
NMaximize[{Log[l[p]], 0.4 <= p <= 0.6}, p]
```

produces the following error:

```
NMaximize::nnum: The function value Indeterminate is not a number at {p} = {0.409146}.
```

This is followed by a very long expression representing the number Mathematica is trying to evaluate. But if we exercise just a little judgment and instead use

```
ll[p_] := 106048 Log[p] + (212480 - 106048) Log[1 - p]
NMaximize[{ll[p], 0 <= p <= 1}, p]
```

we get

```
{-147280., {p -> 0.499096}}
```

which is correct to six decimal places.

If all 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, [a]

Доверяй, но проверяй
[Trust, but verify].

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

18.7 The likelihood function for independent experiments

Often a random experiment is actually a sequence of n independent random experiments wit the same likelihood, or a set of n independent observations of identically distributed random variables $X_1, \dots, X_n$. If R denotes the range of each X_i, then the set $\boldsymbol{S}$ of experimental outcomes is R^n, or better yet $\bigcup_{n=1}^{\infty} R^n$.

Let $X_1, \dots, X_n$ be independent and identically distributed with common pmf or pdf

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

Given observations $X_1 = x_1, \dots, X_n = x_n$, the (joint) likelihood function is

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

Taking logarithms gives

$$\ln L(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \ln f(x_i;\theta).$$

18.7.1 Example (Independent and identically distributed normals) Let $X_1, \dots, X_n$ be independent and identically distributed $N(\mu, \sigma^2)$ random variables. We don't know μ and σ^2, but given the sample $x_1, \dots, x_n$, the likelihood function is

Larsen–Marx [12]: pp. 290–291

$$L(\mu, \sigma^2; x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

We may ignore constants and write

$$L(\mu, \sigma^2; x_1, \dots, x_n) \propto \sigma^{-n} \prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

or, by taking logs we would get (up to a constant)

$$\ln L(\mu, \sigma^2; x_1, \dots, x_n) \propto -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

To find the maximizer of the log-likelihood we set both partials $\partial/\partial\mu$ and $\partial/\partial\sigma^2$ to zero. Now

$$\frac{\partial}{\partial\mu}\ln L(\hat{\mu},\widehat{\sigma^2}) = \frac{1}{\widehat{\sigma^2}}\sum_{i=1}^{n}(x_i-\hat{\mu}) \tag{1}$$

and (treating σ^2 as a single symbol),

$$\frac{\partial}{\partial\sigma^2}\ln L(\hat{\mu},\widehat{\sigma^2}) = -\frac{n}{2}\frac{1}{\widehat{\sigma^2}} + \frac{1}{2}\left(\frac{1}{\widehat{\sigma^2}}\right)^2\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \tag{2}$$

Setting (1) to zero implies

$$\hat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{3}$$

That is, the MLE of μ is the sample average.

Note that

$$\boldsymbol{E}\,\hat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n} X_i}{n} = \mu.$$

That is, if the dgp is governed by parameters μ and σ^2, then the expectation of the $\hat{\mu}_{\text{MLE}}$ is μ. In the case the MLE estimator of μ for normal random variable is an **unbiased estimator**.

Multiplying (2) by $2(\widehat{\sigma^2})^2$ and setting it to zero gives:

$$-n\widehat{\sigma^2} + \sum_{i=1}^{n}(x_i-\hat{\mu})^2 = 0,$$

or letting

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} (= \hat{\mu}),$$

we get

$$-n\widehat{\sigma^2} + \sum_{i=1}^{n}(x_i-\bar{x})^2 = 0,$$

so

$$\widehat{\sigma^2}_{\text{MLE}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \tag{4}$$

Need to introduce these concepts earlier or not refer to them,

While $\hat{\mu}_{\text{MLE}}$ is unbiased, $\hat{\sigma}^2_{\text{MLE}}$ is *biased*. To see this, let's compute its expectation. We start with the expectation:

$$\boldsymbol{E}\,\hat{\sigma}^2_{\text{MLE}} = \boldsymbol{E}\,\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}.$$

So let's start with $\boldsymbol{E}(X_i-\bar{X})^2$. First, let

$$Z = \sum_{j\neq i} X_j.$$

Then Z has mean $(n-1)\mu$ and variance $(n-1)\sigma^2$ as the sum of $n-1$ independent $N(\mu,\sigma^2)$ random variables. Moreover

$$\boldsymbol{E}\,Z^2 = (n-1)^2\mu^2 + (n-1)\sigma^2.$$

since for any rv $\boldsymbol{Var}Y = \boldsymbol{E}(Y^2) - (\boldsymbol{E}\,Y)^2$. Also note that X_i and Z are independent, so

$$\boldsymbol{E}\,X_iZ = (\boldsymbol{E}\,X_i)(\boldsymbol{E}\,Z) = (n-1)\mu^2.$$

Finally observe that

$$\bar{X} = \frac{X_i + Z}{n}.$$

Thus

$$\begin{aligned}
\boldsymbol{E}(X_i - \bar{X})^2 &= \boldsymbol{E}\left(X_i - \frac{X_i+Z}{n}\right)^2 \\
&= \boldsymbol{E}\left(\frac{n-1}{n}X_i - \frac{1}{n}Z\right)^2 \\
&= \frac{1}{n^2}\,\boldsymbol{E}\left((n-1)^2X_i^2 - 2(n-1)X_iZ + Z^2\right) \\
&= \frac{1}{n^2}\left((n-1)^2(\mu^2+\sigma^2) - 2(n-1)^2\mu^2 + (n-1)^2\mu^2 + (n-1)\sigma^2\right) \\
&= \frac{1}{n^2}\left(\left[(n-1)^2 - 2(n-1)^2 + (n-1)^2\right]\mu^2 + \left[(n-1)^2 + (n-1)\right]\sigma^2\right) \\
&= \frac{1}{n^2}n(n-1)\sigma^2 \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

It follows from (4) that

$$\boldsymbol{E}\,\hat{\sigma}^2_{\text{MLE}} = \frac{n-1}{n}\sigma^2.$$

That is, on average, σ^2_{MLE} underestimates the variance. The reason is this. If we knew μ, we want to use as our estimate

$$\frac{\sum_{i=1}^n (x_i-\mu)^2}{n},$$

but we don't know μ, so we use $\bar{x}$ instead. But remember $\bar{x}$ is the value of μ that minimizes the above sum of squares, so it should be smaller on average than using the true μ.

Thus $\hat{\sigma}^2_{\text{MLE}}$ is biased, but the bias tends to zero as $n \to \infty$.

An unbiased estimate of σ^2 is given by

$$S^2 = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-1} = \frac{n}{n-1}\widehat{\sigma^2}_{\text{MLE}}.$$

Now go back and realize that the computation of the expectations depends only on the fact that the X_i are independent and identically distributed with mean μ and variance σ^2, not that they are normal. □

18.8 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 [15] 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 [15, 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.

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 115 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. We will just ignore tie games, since they are effectively not complete games. That gives us 111 (after the 2019 Series) 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, $k = 0, \dots, 3$. It follows from your earlier homework that

$$L(p; N_0, N_1, N_2, N_3) = \underbrace{\frac{N!}{N_0!N_1!N_2!N_3!}\left[\prod_{k=0}^{3}\binom{3+k}{k}^{N_k}\right]}_{\text{independent of } p}\prod_{k=0}^{3}\left[p^4(1-p)^k + p^k(1-p)^4\right]^{N_k}.$$

I don't know how to solve for the maximizer analytically, so numerical methods must be used.

18.9 Confidence intervals for Normal means if σ is known

So far we have looked at **point estimates**, and barely made a dent in the subject. (Erich L. Lehmann's classic *Theory of Point Estimation* [13] runs to about 500 pages.) But there is more than just point estimation.

Interval estimates are closely related to hypothesis testing (coming up soon) and convey more information than point estimates.

Go back to the Normal estimation case. The maximum likelihood estimator $\hat{\mu}_{\text{MLE}}$ of the mean μ is just the sample mean $\bar{x} = \sum_i x_i/n$, but how "good" is that estimate? If $X_1, \dots, X_n$ are independent and identically distributed $N(\mu, \sigma^2)$, then

$$\hat{\mu}_{\text{MLE}} = \frac{X_1 + \cdots + X_n}{n} \sim N(\mu, \sigma^2/n),$$

so by standardizing $\hat{\mu}$ we have

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Let Z be a standard normal random variable and define z_α by

$$P(Z > z_\alpha) = \alpha,$$

It is a fact that

$$z_{0.025} = 1.96.$$

Therefore

$$P\left(-1.96 \leqslant \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \leqslant 1.96 \right) = 0.95$$

But this event is also equal to the event

$$\left(\hat{\mu} - \frac{1.96\sigma}{\sqrt{n}} \leqslant \mu \leqslant \hat{\mu} + \frac{1.96\sigma}{\sqrt{n}} \right).$$

So another way to interpret this is

$$P\big(\mu \in [\hat{\mu} - (1.96\sigma/\sqrt{n}),\ \ \hat{\mu} + (1.96\sigma/\sqrt{n})]\big) = 95\%$$

even though μ is not random. The random interval

$$I = [\hat{\mu} - (1.96\sigma/\sqrt{n}),\ \ \hat{\mu} + (1.96\sigma/\sqrt{n})]$$

is called a 95% **confidence interval** for μ. More generally we have the following

To get a $1 - \alpha$ **confidence interval** for μ when σ is known, set

$$I = \left[\hat{\mu} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}},\ \hat{\mu} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \right]. \tag{5}$$

Then

$$P(\mu \in I) = 1 - \alpha.$$

18.9.1 Interpreting confidence intervals

Remember that μ is not random, rather the interval $I(X) = [\hat{\mu} - 1.96\sigma/\sqrt{n},\ \hat{\mu} + 1.96\sigma/\sqrt{n}]$ is random, since it is based on the random variable $\hat{\mu}$. But once I calculate I, μ either belongs to I or it doesn't, so what am I to make of the 95% probability? I think the way to think about it is this:

No matter what the values of μ and σ are, following the procedure "draw a sample X from the distribution $N(\mu, \sigma^2)$, and use (5) to calculate the interval $I(X)$," the interval $I(X)$ will then have a 95% probability of containing μ.

This is not the same as saying, I used (5) to calculate the interval I, so no matter what the values of μ and σ are, the interval I has a 95% probability of containing μ.

It is the *procedure*, not the interval per se, that gives us the confidence.

Figure 18.6 shows the result of using this procedure 100 times to construct a symmetric 95% confidence interval for μ, based on (pseudo-)random samples of size 5 drawn from a standard normal distribution. Note that in this instance, five of the 100 intervals missed the true mean 0.

Figure 18.6. Here are one hundred 95% confidence intervals for the mean from a Monte Carlo simulation of a sample of size 5 independent standard normals. The intervals that do not include the true mean 0 are shown in red.

Table 18.1. WTF?

I am putting this here, because I don't where it belongs. The activity of estimation has given rise to a fair amount of jargon. The definitions I give here agree with *The Cambridge Dictionary of Statistics* [10].

When T is an estimator of the parameter θ (and so a random variable),

- the **sampling error** is the difference between T and θ.
- the **sampling distribution** is the distribution of T.
- the **standard error** of T is the standard deviation of T, or the standard deviation of the sampling distribution.
- **Error bars** are graphical devices used to plot estimates and to give some idea of their variability. There is not a universal practice on how long error bars, should be—it is field-, and perhaps journal-specific. But usually they are the length of the standard error (one standard deviation of the estimator) or a 95% confidence interval for the estimator.

18.10 ⋆ Another fallacy in the interpretation of confidence intervals

According to Cumming, Williams, and Fidler [7], the following statement is a common misunderstanding of confidence intervals, at least among psychological researchers.

> **Fallacy:** "A 95% confidence interval I is constructed for the mean μ of a normal distribution. Thus there is a 95% probability that the estimate $\hat{\mu}$ from an independent replication will fall into the interval I."

I don't know how common this fallacy is in the general research population, but here is an example. Gilbert, King, Pettigrew, and Wilson [11] writing in the prestigious journal *Science* argue:[3]

> If all 100 of the original studies examined by OSC had reported true effects, then sampling error alone should cause 5% of the replication studies to "fail" by producing results that fall outside the 95% confidence interval of the original study [...]

But here is why it is a fallacy. Let's take a really simple case so we can see what is goin on. Imagine that we are drawing a sample X of size one from a standard Normal$(0, 1)$ distribution. The MLE estimate of the mean is then just the sample value x. The 95% confidence interval is then $[x - 1.09, x + 1.96]$. If we take an independent second sample, Y, the question is, what is

$$P(Y \in [x - 1.96, x + 1.96])?$$

The answer clearly depends on x. If by some fortunate stroke of luck $x = 0$, then 1.96 was chosen so that for a standard normal random variable Y, we have $P(Y \in [-1.96, 1.96]) = 0.95$. But if x is not zero, then the probability will be smaller than 0.95.

So what we want to know is

$$P(\,|Y - X| \leqslant 1.96), \quad \text{where } X \text{ and } Y \text{ are independent standard normals.}$$

[3] Just so you don't think I only pick on psychologists, while Gilbert and Wilson are social psychologists, King and Pettigrew are political scientists.

This is given by the double integral of the joint density $f(x, y)$ over the strip of height (and depth) 1.96 around the diagonal in (x, y) space:

$$\int_{-\infty}^{\infty}\int_{x-1.96}^{x+1.96} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dy\, dx = 0.834.$$

(We cannot evaluate this integral in closed form, but it is tractable numerically. The numerical value above was computed by Mathematica 12.) Cumming and Maillardet [6] refer to this as the **capture probability** of the 95% confidence interval.

Now observe that when σ is known, if the experiment is to take a sample of size n, construct a 95% confidence interval based on the sample, then take another independent sample of size n, then the probability of capture is unchanged: Let X be the mean of a sample of size n. The 95% confidence interval is $X \pm 1.96\sigma/\sqrt{n}$. If Y is another independent sample mean (of size n), then

$$P\big(|Y - X| \leqslant 1.96\sigma/\sqrt{n}\big) = P\bigg(\left|\frac{\sqrt{n}}{\sigma}Y - \frac{\sqrt{n}}{\sigma}X\right| \leqslant 1.96\bigg),$$

but $(\sqrt{n}/\sigma)Y$ and $(\sqrt{n}/\sigma)X$ are independent standard normals, and we have just computed this probability as 83.4%.

If σ is unknown, then we have to estimate it, and our estimate will have a χ^2 distribution (and will be independent of our estimate of the mean, (see Corollary 11.5.2 below). This means (see Proposition 18.12.3 below) the length of the confidence interval follows a χ^2 distribution, so we need to add a third dimension to the integral above. This will either make a good exercise, or I will add the result to these notes at a later time.

18.11 Considerations in constructing confidence intervals

There are two more points worth noting.

- Suppose we know μ, and we want to choose an interval I so that the standard normal random variable $Z = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}$ lies in I with probability $1 - \alpha$. Any interval $[a, b]$ satisfying $\int_a^b \frac{1}{2\pi} e^{-z^2/2}\, dz = 1 - \alpha$ has this property.

> Because of the shape (symmetric and unimodal) of the normal distribution, the symmetric interval $[-\frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$ is the *shortest* such interval.

- Because of the properties of the standard normal distribution, the length of the interval $\left[\hat{\mu} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \hat{\mu} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right]$ does not depend on μ.

- For distributions that are not symmetric, you may want to construct asymmetric confidence intervals. I can think of at least two principles you could use.

1. Choose the *shortest* interval $[a, b]$ containing your point MLE $\hat{\theta}$ that has $P_{\hat{\theta}}\big([a, b]\big) = 1 - \alpha$. This would be the interval where the likelihood (= density) is highest. Since $\hat{\theta}$ maximizes the likelihood, we know it will be in the interval.

Oops. How do we know that an interval is the shortest set? Maybe we would be better off taking two short intervals instead one long one. For unimodal (single-peaked) densities, this won't happen.

2. The other principle you might consider is to choose an interval $[a, b]$ so that $P(\theta < a) = P(\theta > b) = \alpha/2$, bearing in mind the above interpretation of the probability.

In the normal case, these two principles are not in conflict and procedure for constructing the interval described above is consistent with both.

18.12 Confidence intervals for Normal means if σ is *not* known

The confidence interval given by (5) depends on the standard deviation σ. You might ask, when might I know σ, but not know μ? Maybe in a case like this: I can imagine the variance in a measurement of weight using a balance beam scale depends on the friction in the balance bearing. I can also imagine that the mean measurement of a sample's mass depends on the sample's actual mass. I might have a lot of experience with this particular of scale, so that I know the variance σ^2, but the mean of the measurement depends on which sample I am weighing. To get a good estimate of the weight, I might make several measurements,[4] and I could then use this procedure to generate a confidence interval. (I just made this up, and it sounds plausible, but do any of you chemists or engineers have any real information on such scales?)

So what do we do if we don't know σ? We can use an MLE estimate $\hat{\sigma}$ of σ to calculate a confidence interval. The catch is that $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ is *not* a Standard Normal random variable.

18.12.1 Theorem *[12, Theorem 7.3.5, p. 393] For a sample $X_1, \dots, X_n$ of independent and identically distributed Normal $N(\mu, \sigma^2)$ random variables, the statistic*

$$T_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student t-distribution with $n-1$ degrees of freedom.

We will discuss the t-distribution in more detail later. For now, we discuss the mechanics of constructing confidence intervals based on the t-distribution..

18.12.2 Definition (t-distribution cutoffs) *Larsen–Marx [12, p. 395] define $\boldsymbol{t_{\alpha,n}}$ by*

$$P(T_n \geqslant t_{\alpha,n}) = \alpha,$$

where T_n has the Student t-distribution with n degrees of freedom.

Then

$$P\bigg(-t_{\alpha/2,n-1} \leqslant \frac{\bar{X} - \mu}{S/\sqrt{n}} \leqslant t_{\alpha/2,n-1} \bigg) = 1 - \alpha$$

or we can turn the inequality "inside out" to get the equivalent statement

$$P\big(\bar{X} - t_{\alpha/2,n-1} S/\sqrt{n} \leqslant \mu \leqslant \bar{X} + t_{\alpha/2,n-1} S/\sqrt{n}\big) = 1 - \alpha.$$

In other words,

18.12.3 Proposition *Given the sample values $x_1, \dots, x_n$ from n independent and identically distributed draws from a normal distribution, a $1-\alpha$ confidence interval for μ is the interval*

$$\left(\bar{x} - \frac{s\, t_{\alpha/2,n-1}}{\sqrt{n}},\ \ \bar{x} + \frac{s\, t_{\alpha/2,n-1}}{\sqrt{n}} \right).$$

Figure 18.7 shows the result of using this procedure 100 times to construct a symmetric 95% confidence interval for μ, based on (pseudo-)random samples of size 5 drawn from a standard normal distribution. Note that in this instance, 7 of the 100 intervals missed the true mean 0.

[4] My grandfather was a carpenter, so I am quite familiar with the old saw, "Measure twice, cut once." (Sorry, I couldn't help myself.)

Compare this figure to Figure 18.6, where the variance was known. In that case, all the confidence intervals had the same width. When the variance is estimated from the sample, this is no longer the case. We shall show later that the sample mean and the sample variance are stochastically independent, so a short confidence interval is not necessarily a "better" confidence interval.

18.13 Digression: The quantiles z_α

We already covered this in Section 10.7.

Statisticians have adopted the following special notation. Let Z be a Standard Normal random variable, with cumulative distribution function denoted Φ.

18.13.1 Definition *For $0 < \alpha < 1$, define z_α by*

$$P(Z > z_\alpha) = \alpha,$$

Larsen–Marx [12]: p. 307

see Figure 18.8, or equivalently

$$P(Z \leqslant z_\alpha) = 1 - \alpha.$$

Then

$$z_\alpha = \Phi^{-1}(1 - \alpha)$$

This is something you can look up with R or Mathematica's built-in quantile functions. (Remember the **quantile function** is Φ^{-1}.) By symmetry,

$$P(Z < -z_\alpha) = \alpha \qquad \text{and} \qquad P(\,|Z| > z_\alpha) = 2\alpha$$

so

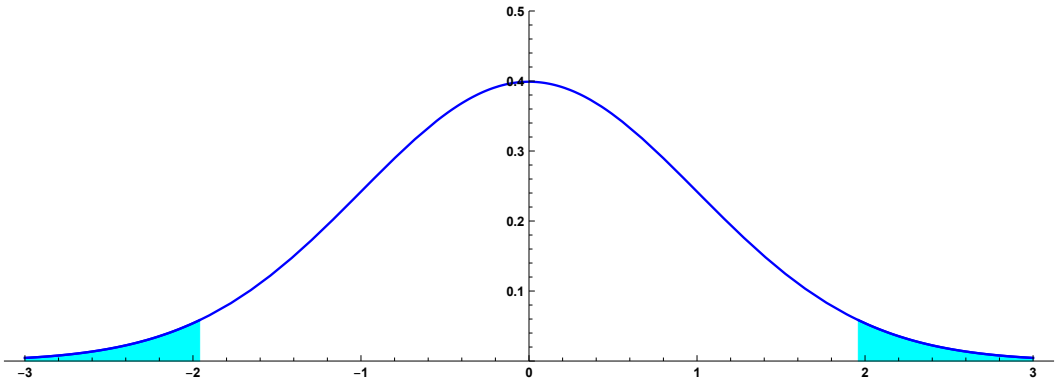
$$P(\,-z_\alpha \leqslant Z \leqslant z_\alpha) = 1 - 2\alpha.$$

The last inequality is often expressed as

$$P\big(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\big) = 1 - \alpha.$$

Here are some commonly used values of α and the corresponding z_α to two decimal places.

α	z_α	$1 - 2\alpha$
0.1	1.28	0.80
0.05	1.64	0.90
0.025	1.96	0.95
0.01	2.33	0.98
0.005	2.58	0.99



This shaded area is the probability of the event ($|Z| > 1.96$), which is equal to 0.05. Values outside the interval $(-1.96, 1.96)$ are often regarded as unlikely to have occurred "by chance."

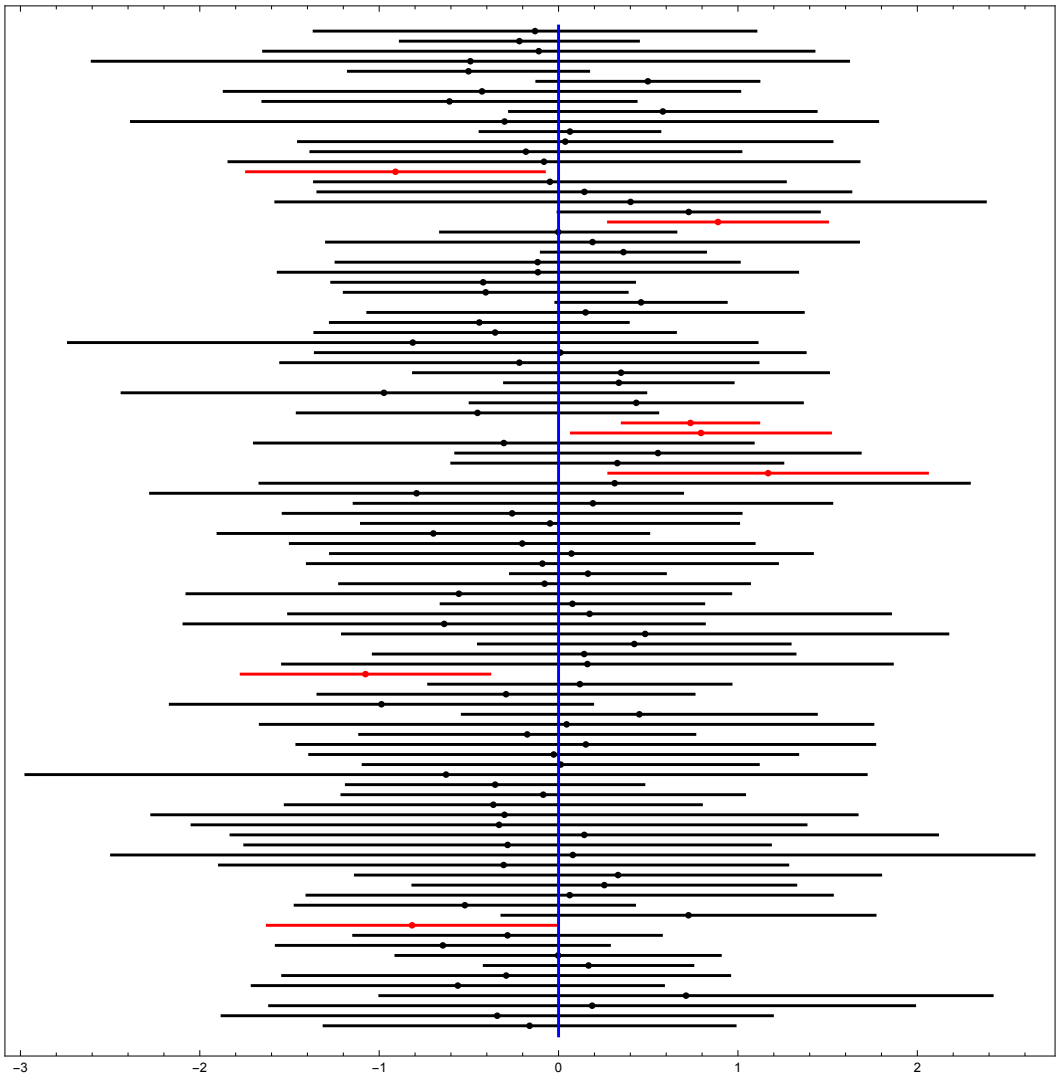


Figure 18.7. Here are one hundred Monte Carlo simulations of a 95% confidence interval for the mean of a standard normal, based on a sample of size five for each simulation. The confidence interval depends on the *estimated* standard deviation, so not all intervals are the same length. The intervals that do not include 0 are shown in red. There are seven of them.

Notice that the third red interval from the top is relatively short, so the estimate appears to be relatively precise. Nevertheless, it is wrong in the sense that the true parameter is not in the confidence interval. This is a case of what is technically known as "bad luck."

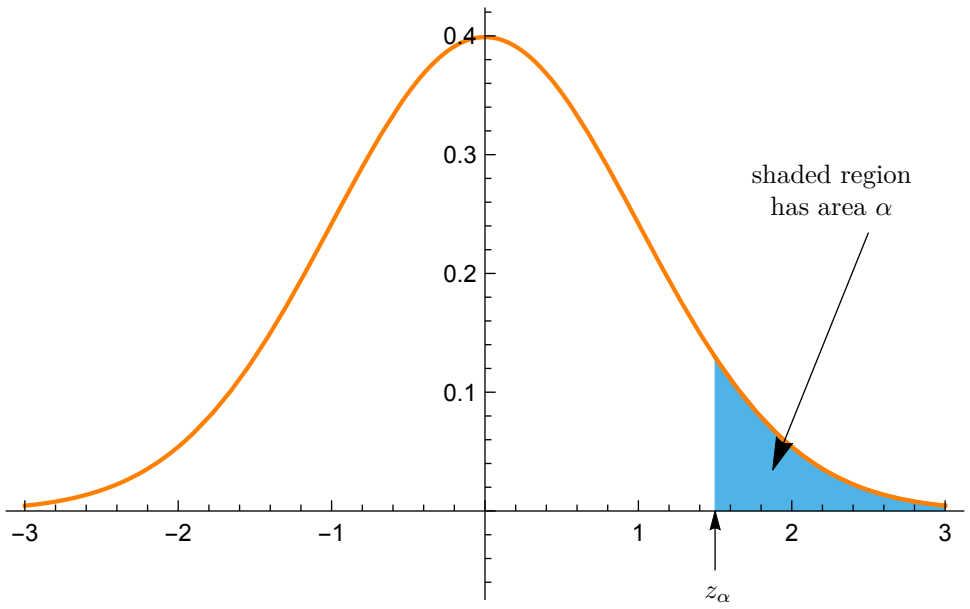


Figure 18.8. Given α, z_α is chosen so that the shaded region has area $= \alpha$.

18.13.1 *t*-quantiles versus *z*-quantiles

The values z_α, which are used to construct a $(1-\alpha\%)$ confidence intervals based on knowing the standard deviation σ, can be very misleading for small sample sizes, when σ is estimated by the unbiased version of the MLE estimate. The following Table 18.2 gives z_α and $t_{\alpha,n}$ for various values of α and n. This shows how the critical value of a test changes with the number of degrees of freedom.

	degrees of freedom n										
	1	2	4	8	16	32	64	128	256	512	
α	$t_{\alpha,n}$										z_α
0.10	3.08	1.89	1.53	1.4	1.34	1.31	1.29	1.29	1.28	1.28	1.28
0.05	6.31	2.92	2.13	1.86	1.75	1.69	1.67	1.66	1.65	1.65	1.64
0.025	12.71	4.3	2.78	2.31	2.12	2.04	2.	1.98	1.97	1.96	1.96
0.01	31.82	6.96	3.75	2.9	2.58	2.45	2.39	2.36	2.34	2.33	2.33
0.005	63.66	9.92	4.6	3.36	2.92	2.74	2.65	2.61	2.6	2.59	2.58

Table 18.2. $t_{\alpha,n}$ compared to z_α for various degrees of freedom n and significance levels α.

Section 22.20 describes the commands in R and Mathematica that can be used to compute these quantiles.

Bibliography

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

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and regression trees*. Boca Raton, Florida: CRC Press.

[3] 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

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

[5] 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

[6] G. Cumming and R. Maillardet. 2006. Confidence intervals and replication: where will the next mean fall? *Psychological Methods* 11(3):217–227. DOI: 10.1037/1082-989X.11.3.217

[7] G. Cumming, J. Williams, and F. Fidler. 2004. Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics* 3(4):299–311. DOI: 10.1207/s15328031us0304_5

[8] 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

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

[10] B. S. Everitt and A. Skrondal. 2010. *The Cambridge dictionary of statistics*, 4th. ed. Cambridge: Cambridge University Press.

[11] D. T. Gilbert, G. King, S. Pettigrew, and T. D. Wilson. 2016. Comment on "Estimating the reproducibility of psychological science". *Science* 352(6277):1037–1038. DOI: 10.1126/science.aad7243

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

[13] E. L. Lehmann. 1983. *Theory of point estimation*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley and Sons.

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

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

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

[17] ——— . 1990. *Envisioning information*. Cheshire Connecticut: Graphics Press.

[18] ——— . 2006. *Beautiful evidence*. Cheshire Connecticut: Graphics Press.

[19] J. W. Tukey. 1977. *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley.