# 3
# Likelihood

The purpose of models is to allow us to use past observations (*data*) to make predictions. In order to do this, however, we need a way of choosing a value of the parameter (or parameters) of the model. This process is called parameter *estimation* and this chapter discusses the most important general approach to it. In simple statistical analyses, these stages of model building and estimation may seem to be absent, the analysis just being an intuitively sensible way of summarizing the data. However, the analysis is only scientifically useful if we can generalize the findings, and such generalization must imply a model. Although the formal machinery of modelling and estimation may seem heavy handed for simple analyses, an understanding of it is essential to the development of methods for more difficult problems.

In modern statistics the concept which is central to the process of parameter estimation is *likelihood*. Likelihood is a measure of the *support* provided by a body of data for a particular value of the parameter of a probability model. It is calculated by working out how probable our observations would be if the parameter were to have the assumed value. The main idea is simply that parameter values which make the data more probable are better supported than values which make the data less probable. In this chapter we develop this idea within the framework of the binary model.

## 3.1 Likelihood in the binary model

Fig. 3.1 illustrates the outcomes observed in a small study in which 10 subjects are followed up for a fixed time period. There are two possible outcomes for each subject: *failure*, such as the development of the disease of interest, or *survival*. We adopt a binary probability model for the outcome for each subject in which failure has probability $\pi$ and survival has probability $1-\pi$. The complete tree would have many branches but only those corresponding to the observed study result is shown in full. To calculate the probability of occurrence of this result we simply multiply probabilities along the branches of the tree in the usual way:

$$\pi \times \pi \times (1-\pi) \times \cdots \times (1-\pi) = (\pi)^4(1-\pi)^6.$$



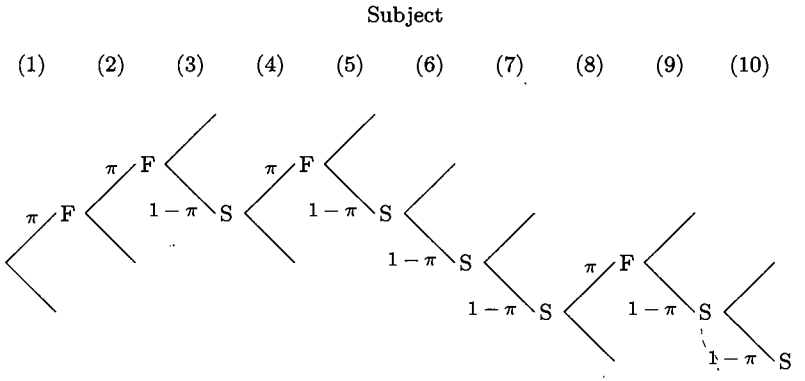



**Fig. 3.1.** Study outcomes for 10 subjects.

This expression can be used to calculate the probability of the observed study result for any specified value of $\pi$. For example, when $\pi = 0.1$ the probability is

$$(0.1)^4 \times (0.9)^6 = 5.31 \times 10^{-5}$$

and when $\pi = 0.5$ it is

$$(0.5)^4 \times (0.5)^6 = 9.77 \times 10^{-4}.$$

The results of these calculations show that the probability of the observed data is greater for $\pi = 0.5$ than for $\pi = 0.1$. In statistics this is often expressed by saying that $\pi = 0.5$ is more *likely* than $\pi = 0.1$, meaning that the former value is better supported by the data. In everyday use the words probable and likely mean the same thing, but in statistics the word likely is used in this more specialized sense.

**Exercise 3.1.** Is $\pi = 0.4$ more likely than $\pi = 0.5$?

The result of the expression

$$(\pi)^4(1-\pi)^6,$$

is a probability, but when we use it to assess the amount of support for different values of $\pi$ it is called a *likelihood*. More generally, if we observed $D$ failures in $N$ subjects, the likelihood for $\pi$ would be

$$(\pi)^D(1-\pi)^{N-D},$$

and we shall call this expression the *Bernoulli* likelihood, after the Swiss mathematician. Because there are so many possible outcomes to the study,

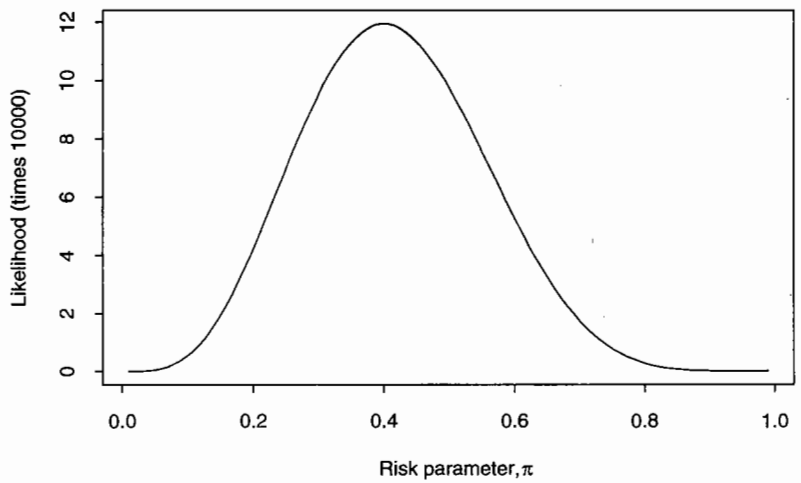



**Fig. 3.2.** The likelihood for $\pi$.

the likelihood (which is the probability of just one of these) is a small number. However, it is not the *absolute* value of the likelihood which should concern us, but its *relative* value for different choices of $\pi$.

Returning to our numerical example, Fig. 3.2 shows how the likelihood varies as a function of $\pi$. The value $\pi = 0.4$ gives a likelihood of $11.9 \times 10^{-4}$, which is the largest which can be achieved. This value of $\pi$ is called the *most likely value* or, more formally, the *maximum likelihood estimate* of $\pi$. It coincides with the observed proportion of failures in the study, 4/10.

## 3.2 The supported range for $\pi$

The most likely value for $\pi$ is 0.4, with likelihood $11.9 \times 10^{-4}$. The likelihood for any other value of $\pi$ will be less than this. How much less is measured by the *likelihood ratio*, which takes the value 1 when $\pi = 0.4$ and values less than 1 for any other values of $\pi$. This provides a more convenient measure of the degree of support than the likelihood itself. It can be used to classify values of $\pi$ as either supported or not according to some critical value of the likelihood ratio. Values of $\pi$ with likelihood ratios above the critical value are reported as 'supported', and values with likelihood ratios below this critical value as 'not supported'. The *supported range* for $\pi$ is the set of values of $\pi$ with likelihood ratios above the critical value. The choice of the critical value is a matter of convention.

For our observation of 4 failures and 6 survivors, the likelihood ratio as a function of $\pi$ is shown in Figure 3.3. We have used the number 0.258



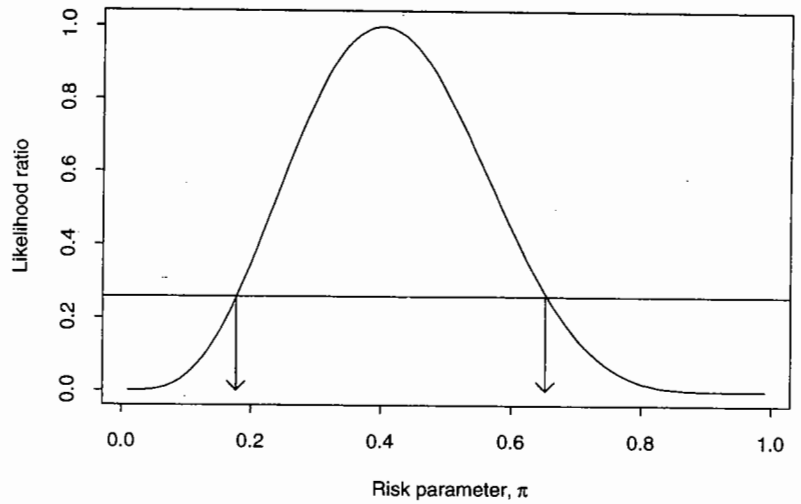



**Fig. 3.3.** The likelihood ratio for $\pi$.

for the critical value of the likelihood ratio and indicated the limits of the supported range with the two arrows. The range of supported values for $\pi$ is rather wide in this case: from 0.17 to 0.65.* For any choice of critical value the width of the supported range reflects the uncertainty in our knowledge about $\pi$. The main thing which determines this is the quantity of data used in calculating the likelihood. For example, if we were to observe 20 failures in 50 subjects, the most likely value of $\pi$ would still be 0.4, but the supported range would be narrower (see Figure 3.4).

Although the concept of a supported range based on likelihood ratios is intuitively simple, it requires some consensus about the choice of critical value. The achievement of this has not proved easy, since many scientists lack an intuitive feel for the amount of uncertainty corresponding to a stated numerical value for the likelihood ratio. As a result, statistical theorists have tried to find ways to measure the uncertainty about the value of a parameter in terms of *probability* which, it is argued, is more easily interpreted. The way of doing this which is most widely accepted in the scientific community is by imagining a large number of repetitions of the study. This approach is known as the *frequentist* theory of statistics and leads to a *confidence interval* for $\pi$ rather than a supported range. Another approach, often favoured by mathematicians, is based on a probability measure for the subjective 'degree of belief' that the parameter value lies in a stated *credible*

*These values were obtained from the graph, as illustrated. We shall be describing more convenient approximate methods for their computation in Chapter 9.

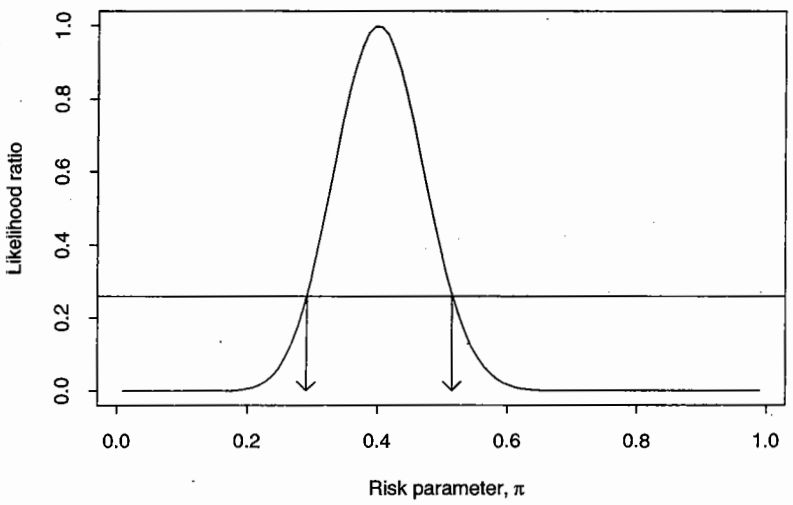



**Fig. 3.4.** The likelihood ratio based on 20 failures in 50 subjects.

*interval.* This is the *Bayesian* theory of statistics.

Luckily for applied scientists, these philosophical differences can be resolved, at least for the analysis of moderately large studies. In this case, we will show in Chapter 10 that the supported range based on a likelihood ratio criterion of 0.258 coincides approximately with a 90% confidence interval in the frequentist theory of statistics and a 90% credible interval in the Bayesian theory. We shall, therefore, set aside these difficulties for the present and continue to develop the idea of likelihood, which holds a central place in both theories of statistics and from which most of the statistical methods of modern epidemiology can be derived.

### 3.3 The log likelihood

The likelihood, when evaluated for a particular value of the parameter, can turn out to be a very small number, and it is generally more convenient to use the (natural) logarithm of the likelihood in place of the likelihood itself.† When combining log likelihoods from independent sets of data the separate log likelihoods are added to form the combined likelihood. This is because the likelihoods themselves, being the probabilities of independent sets of data, are combined by multiplication. The log likelihood for $\pi$, in this example, is

$$4\log(\pi) + 6\log(1-\pi).$$

†Readers not completely familiar with the logarithmic function, $\log(x)$ and its inverse, the exponential function, $\exp(x)$, are referred to Appendix A.



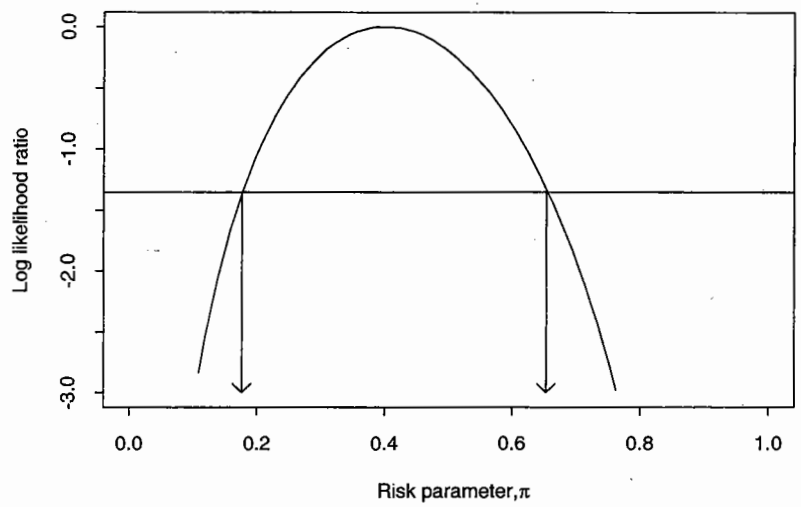



**Fig. 3.5.** The log likelihood ratio for $\pi$.

**Exercise 3.2.** Calculate the log likelihood when $\pi = 0.5$ and when $\pi = 0.1$.

The log likelihood takes its maximum at the same value of $\pi$ as the likelihood, namely $\pi = 0.4$, so its maximum is

$$4\log(0.4) + 6\log(0.6) = -6.730.$$

To obtain the log likelihood *ratio*, this maximum must be *subtracted* from the log likelihood. A graph of the log likelihood ratio is shown in Fig. 3.5. The supported range for $\pi$ can be found from this graph in the same way as from the likelihood ratio graph, by finding those values of $\pi$ for which the log likelihood ratio is greater than

$$\log(0.258) = -1.353.$$

**Exercise 3.3.** Calculate the log likelihood ratios for $\pi = 0.1$ and $\pi = 0.5$. Are these values of $\pi$ in the supported range?

In general, the log likelihood for $\pi$, when $D$ subjects fail and $N - D$ survive, is

$$D\log(\pi) + (N-D)\log(1-\pi).$$

We shall show in Chapter 9 that this expression takes its maximum value when $\pi = D/N$, the observed proportion of subjects who failed.

If the binary model is parametrized in terms of the odds parameter, $\Omega$, by substituting $\Omega/(1+\Omega)$ for $\pi$ and $1/(1+\Omega)$ for $(1-\pi)$, we obtain the log likelihood

$$D\log(\Omega) - N\log(1+\Omega).$$

This takes its maximum value when $\Omega = D/(N-D)$, the ratio of the number or failures to the number of survivors. The maximum value of the log likelihood is the same whether the log likelihood is expressed in terms of $\pi$ or $\Omega$.

### 3.4 Censoring in follow-up studies

In our discussion of follow-up studies of the occurrence of disease events, or failures, we have assumed that all subjects are potentially observed for the same fixed period. In most practical studies there will be some subjects whose follow-up is incomplete. This will occur

- when they die from other causes before the end of the follow-up interval;
- when they migrate and are no longer covered by the record system which registers failures;
- when they join the cohort too late to complete the follow-up period.

In all three cases the observation time for the subject is said to be censored. In fact, the first type of loss to follow-up, failure due to a *competing cause*, is rather different from the remaining two, but they are usually grouped together and dealt with in the same way. In Chapter 7 we shall discuss the justification for this practice. For the moment, we assume it to be reasonable.

Censoring puts our argument in some difficulty. The model allows for only two outcomes, failure and survival, while our data contains three, failure, survival, and censoring. For the present we shall avoid this difficulty with a simple pretence. As an illustration, suppose we have followed 1000 men for five years, during which 28 suffered myocardial infarction and 972 did not, but observation of 15 men was censored before completion of five years follow-up. If all 15 men were withdrawn from study on the *first* day of the follow up period, the size of the cohort would be 985 rather than 1000. Conversely, if they were all withdrawn on the *last* day, censoring could be ignored and the cohort size treated as a full 1000. When censoring is evenly spread over the study interval, we would expect an answer which lies somewhere in between these two extreme assumptions. This suggests treating the effective cohort size as 992.5 — mid-way between 985 and 1000. This convention is equivalent to the assumption that 7.5 subjects are censored on the first day of follow up and 7.5 on the last day.



**Table 3.1.** Genotypes of 7 probands and their parents

| Proband's genotype | Parents' genotypes | | Number |
|---|---|---|---|
| | Mother | Father | |
| (a,c) | (a,b) | (c,d) | 4 |
| (b,d) | (a,b) | (c,d) | 1 |
| (a,c) | (a,b) | (c,c) | 2 |

With only 15 subjects lost to follow up through censoring, this crude strategy for dealing with censoring is quite satisfactory, but if 150 were censored it could be seriously misleading. In Chapter 4 we shall see how this problem can be dealt with by extending the model.

### 3.5 Applications in genetics ★

The use of the log likelihood as a measure of support is of considerable importance in genetics. However, in that field it is conventional to use logarithms to the base 10 rather than natural logarithms. Since the two systems of logarithms differ only by a constant multiple (see Appendix A), this is only a trivial modification of the idea.

As an illustration of the use of log likelihood in genetics, we continue the example introduced in Exercises 2.4 and 2.5. Table 3.1 shows some hypothetical data which might have formed part of that collected in a study of an association between disease risk and presence of a certain HLA haplotype. If we were to observe a set of families over time, in order to relate the genotype to the eventual occurrence or non-occurrence of disease, then we could calculate a likelihood based on the probability of disease conditional upon genotype. However, such studies are logistically very difficult and are rarely done. Instead it is more usual to obtain, usually from clinicians, a collection of known cases of disease (*probands*) and their relatives, and to compare the genotypes of probands with the predictions from the model.

As in Exercise 2.5, we shall consider the model in which presence of a given haplotype, (a) say, leads to a risk of disease $\theta$ times as high as in its absence. Table 3.1 shows data concerning 7 probands and their parents. For each of the genetic configurations shown in the table, we derived the conditional probability of the genotype of a proband conditional on the genotypes of parents in Exercise 2.5 and we showed that these probabilities depend only on the risk ratio parameter $\theta$.

**Exercise 3.4.** Write down the expression for the log likelihood as a function of the unknown risk ratio, $\theta$, associated with presence of haplotype (a). What is the log likelihood ratio for the value $\theta = 1$ (corresponding to there being no increase in risk) as compared with $\theta = 6.0$ (which is the most likely value of $\theta$ in this case). Is the value $\theta = 1$ supported?

**Solutions to the exercises**

**3.1** The probability of the observed data when $\pi = 0.4$ is

$$0.4^4 \times 0.6^6 = 1.19 \times 10^{-3}.$$

which is more than the probability when $\pi = 0.5$. It follows that $\pi = 0.4$ is more likely than $\pi = 0.5$.

**3.2** The log likelihood when $\pi$=0.5 is

$$4\log(0.5) + 6\log(0.5) = -6.93.$$

The log likelihood when $\pi = 0.1$ is

$$4\log(0.1) + 6\log(0.9) = -9.84.$$

**3.3** The maximum log likelihood, occurring at $\pi = 0.4$, is

$$4\log(0.4) + 6\log(0.6) = -6.73$$

so that the log likelihood ratio for $\pi = 0.5$ is $-6.93 - (-6.73) = -0.20$. For $\pi = 0.1$ it is $-9.84 - (-6.73) = -3.11$. Thus 0.5 lies within the supported range and 0.1 does not.

**3.4** From the solution to Exercise 2.5, the conditional probabilities for each of the three genetic configurations are $\theta/(2\theta + 2)$, $1/(2\theta + 2)$, and $\theta/(\theta + 1)$. Thus, the log likelihood is

$$4\log\left(\frac{\theta}{2\theta+2}\right) + 1\log\left(\frac{1}{2\theta+2}\right) + 2\log\left(\frac{\theta}{\theta+1}\right).$$

At $\theta = 1.0$ this takes the value

$$4\log\left(\frac{1}{4}\right) + 1\log\left(\frac{1}{4}\right) + 2\log\left(\frac{1}{2}\right) = -8.318,$$

and at $\theta = 6.0$ (the most likely value) it is

$$4\log\left(\frac{6}{14}\right) + 1\log\left(\frac{1}{14}\right) + 2\log\left(\frac{6}{7}\right) = -6.337.$$

The log likelihood ratio for $\theta = 1$ is the difference between these, $-1.981$. Thus the parameter value $\theta = 1$ lies outside the limits of support we have suggested in this chapter.

# 3 Likelihood

In addition to C&H, see **Edwards** AWF. *Likelihood: Expanded Edition*. Cambridge University Press. 1972 and 1992. (print book) and **Padiwan** In All Likelihood: Statistical Modelling and Inference Using Likelihood (ebook) in the McGill library.

"*We need a way of choosing a value of the parameter(s) of the model*" (1st paragraph): It is clear from the later text that C&H do not mean to give the impression that one is only interest in a single value or point-estimate. For any method to be worthwhile, it needs to be able to provides some measure of uncertainty, i.e. an interval or range of parameter values.

*"In simple statistical analyses, these stages of model building and estimation may seem to be absent, the analysis just being an intuitively sensible way of summarizing the data."* Part of the reason is that (as an example) a sample mean may simply seem like a natural quantity to calculate, and it does not seem to require an explicit statistical model. Indeed, Miettinen, in his Mini-Dictionary (of Science, Medicine, Statistics and Epidemiological Research [1] ), has defined a *descriptive* statistic as *a statistic derived without any statistical model*. The mean can also be seen as the least squares estimate, in the sense that the sum of the squared deviations of the sample values from any other value than the sample mean would be larger than the sum of the squared deviations about the mean itself, i.e., the sample mean is a least squares estimate. But that purely arithmetic procedure still does not require any assumptions about the true value of the parameter value $\mu$, or about the shape of the distribution of the possible values on both sides of $\mu$. For the grade 6 exercise about the mean number of errors per page, it seemed to make sense to divide the total number of errors by the total number of pages; but what if the task was to estimate the mean weight of the pages? We discussed in class at least two different statistical models – that would lead to different estimates.

*"In modern statistics the concept which is central to the process of parameter estimation is likelihood."* *Older* and less sophisticated methods include the *method of moments*, and the *method of minimum chi-square* for count data. These estimators are not always efficient, and their sampling distributions are often mathematically intractable. For some types of data, the method of weighted least squares is a reasonable approach, and we will also see that iteratively-reweighed least squares is a way to obtain ML estimates without formally calculating likelihoods.

[1] https://link-springer-com.proxy3.library.mcgill.ca/book/10.1007%2F978-94-007-1171-6

Likelihood is *central* not just to obtain *frequentist-type* estimators per se, but also to allow *Bayesian analyses* to combine prior beliefs about parameter values to be updated with the data at hand, and arrive at what one's post-data beliefs should be.

Likelihood provides a very *flexible* approach to combining data, provided one has a probability model for them. As a simple example, consider the challenge of estimating the mean $\mu$ from several independent observations for a $N(\mu, \sigma)$ process, but where each observation is recorded to a different degree of numerical 'rounding' or 'binning.' For example, imagine that because of the differences with which the data were recorded, the $n = 4$ observations are $y_1 \in [4, 6)$, $y_2 \in [3, 4)$, $y_3 \in [5, \infty)$, $y_4 \in [-\infty, 3.6)$. Even if we were told the true value of $\sigma$, the least squares method cannot handle this uni-parameter estimation task.

"*The main idea is simply that parameter values which make the data more probable are better supported than values which make the data less probable.*" Before going on to *their first example*, with a parameter than in principle could take any values in the unit interval, consider a *simpler example* where there are just two values of $\pi$. We have sample of candies from one of two sources: American, where the expected distribution of colours is 30%:70% and the other Canadian where it is 50%:50%. In our sample of $n = 5$, the observed distribution is 2:3. Do the data provide more support for the one source than the other?

## 3.1 Likelihood in the binary model

Notice the *level of detail* at which the observed data are reported in Figure 3.1: not just the numbers of each (4 and 6) but the actual *sequence* in which they were observed. The Likelihood function uses the probability of the observed data. Even if we did not know the sequence, the probability of observing 4 and 6 would be ${}^{10}C_4 = 210$ times larger; however since we assume there is no order effect, i.e., that $\pi$ is constant over trials, the actual sequence does not contain any information about $\pi$, and we would not include this multiplier in the Likelihood. In any case, we think of the likelihood as a function of $\pi$ rather than of the observed numbers of each of the two types.: these data are considered fixed, and $\pi$ is varied.. contrast this with the tail area in a frequentist p-values, which includes other non-observed values more extreme than that observed. Likelihood and Bayesian methods do not do this.

"$\pi = 0.5$ *is more likely than* $\pi = 0.1$" Please realize that this statement by itself could be taken to mean that we should put more money on the 0.5 than the 0.1. It does not mean this. In the candy source example, knowing where

the candies were purchased, or what they tasked like, would be additional information that might in and of itself make one source more likely than the other. The point here is not to use terms that imply a prior or posterior probability distribution on $\pi$. The likelihood function is based just on the data, and in real life any extra prior information about $\pi$ would be combined with the information provided by the data. It would have been better if the authors had simply said "*the data provide more support for* "$\pi = 0.5$ *than* $\pi = 0.1$." Indeed, I don't think "$\pi = 0.5$ *is more likely than* $\pi = 0.1$" is standard terminology. The terminology "0.4 is the ML estimate of $\pi$" is simpler and less ambiguous.

## SOME HISTORY

*See* http://www.biostat.mcgill.ca/hanley/bios601/Likelihood/ *for several papers. Warning: JH does not pretend to be a professional historian, and it is only in the last decade or so that he has started to take an interest in these matters. So the reader is warned that the items and interpretations and accounts given here may not be 100% accurate.*

There is some dispute as to *who first used the principle of ML for the choice of parameter value*. The names of **Laplace** and **Gauss** are often mentioned; **Daniel Bernoulli** deserves mention too.[2] The *seldom mentioned* 1912 paper[3] by **Fisher, while still a student**, is a nice clean example, and shows how Likelihood (he did not use the word likelihood in the paper) is flexible and allows for the different bins sizes with which observations might be recorded, etc. It is worth reading that original paper, but don't spend too much time on section 5, where he deals with the ML estimation of the parameters $\mu$ and $\sigma$ of a Normal distribution: the ML estimate of $\sigma^2$ involves a divisor of $n$ rather than $n-1$, and embarrassment for Fisher, who was from early on, insisted on the correct degrees of freedom when assessing variation.

The *usual* reference is to papers by Fisher in the early 1920's, where he worked of many of the properties of ML estimators.

One interesting feature of the 1912 paper is that Fisher never defined the likelihood as a *product* of probabilities; instead **he defined the log-likelihood as a *sum* of log-probabilities**. This is very much in keeping with his summation of *information* over observations. Indeed, there is a lot in his writings about choosing the most *informative* configurations at which to observe the experimental or study units.

Stephen Stigler, an eminent and very readable historian of Statistics, has written extensively on **Laplace**. Indeed, Stigler considers Laplace to be one of the founders not just of mathematical statistics but of inference in general. His 1774 paper http://www.biostat.mcgill.ca/hanley/bios601/Likelihood/LaplaceInvProb1774-stigler1986.pdf (part 5 of which deals with fitting of a location parameter to 3 data points, when the dispersion is (a) known and (b) not. That paper (now translated by Stigler, and accompanied on the website by a Stigler commentary) was written when Laplace was just 25, and it includes what this website http://www.bayesian-inference.com/laplace calls **Laplace's First Law of Error** or the Laplace Distribution. It is an open-ended double-sided exponential, centred at zero, where the parameter $\mu$ represents the average absolute error, and thus the dispersion, i.e.,

$$pdf(x) = \frac{\mu}{2} \exp[-\mu|x|\ ].$$

see http://en.wikipedia.org/wiki/Laplace_distribution

The just cited website continues, telling us that "**In 1778, Laplace published his second law of errors**, noting that the frequency of an error was proportional to the exponential of the *square* of its magnitude. This was subsequently rediscovered by Gauss (possibly in 1795) and became known as the normal or Gaussian distribution." This is the familiar Normal Error curve, with its smoother and less peaked centre, and thinner tails.

The Laplace website continues that "Laplace published a Bayesian precursor of the Central Limit Theorem (CLT) in 1785, establishing the asymptotic normality of posterior distributions. In 1810, however, Laplace introduced the CLT as it applies to frequentist inference, and as it is mostly known today. Laplace created large-sample theory for both modes of probability (Hald, 2004, p. 30).

The frequentist version of the CLT states that the sample mean is approximately normally distributed in large samples when the variance is finite, regardless of the shape of the error distribution (Hald, 2004, p. 4). This allowed Laplace to work with almost any kind of data, rather than being restricted to binomial [or as we met recently, uniform] problems. Working with larger data sets, the CLT led Laplace to frequentist inference (Laplace, 1811)."

Part 5 of Laplace's 1774 paper used (like Daniel Bernoulli below) just 3 data points to illustrate his approach to obtaining a posterior distribution for the location parameter when the dispersion is/is not known. To do so, he proposes, and then integrates out, a prior for the dispersion, and arrives at a posterior distribution for the location parameter. Then, he uses the *median*

[2] see 'Daniel Bernoulli on maximum likelihood' by M. G. Kendall in Biometrika (1961), 48, 1 and 2, p. 1. here http://www.biostat.mcgill.ca/hanley/bios601/Likelihood/Kendall1961OnDanielBernoulliML.pdf See also 'A list of writings relating to the method of least squares' by Mansfield Merriman in Transactions of the Connecticut Academy of Arts and Sciences 1874-78, pp151-232.

[3] http://www.biostat.mcgill.ca/hanley/bios601/Likelihood/FisherLikelihoodPaper1912.pdf

*of that distribution* as his estimate of the location parameter (he has previously shown that the median minimizes the average absolute error).

This summary, from a well known late 19th century author of a widely used textbook on Least Squares (Merriman), on part V of that same 1774 paper (also 1774). It is part of a long listing of, and commentary on, writings up to then on Least Squares and related topics.

**1774 Laplace. 'Determiner le milieu que l'on doit prendre entre trois observations données d'un même phénomène.' *Mém. Acad. Paris, par divers savans* [étrangers], Vol. IV, pp. 634–644.**

**This is Part V of the 'Mémoire sur la probabilité des causes par les évènemens,' which occupies pages 621–656 of the volume. It contains the first attempt to deduce a rule for the combination of observations from the principles of Probability.**

**Laplace begins by saying that the law of probability of errors of observation may be represented by a curve whose equation is $y=\varphi(x)$, $x$ being any error and $y$ its probability; and this curve must have three properties: 1st, it must be symmetrical with reference to the axis of $y$, since positive and negative errors are equally probable; 2nd, the axis of $x$ must be an asymptote, since the probability of the error $\infty$ is 0; 3rd, the area of the curve must be unity, since it is certain that an error will be committed.**

**Laplace takes $\varphi(x)$ as $\varphi(x)=\frac{1}{2}me^{-m|x|}$ ($x$ being regarded as always positive), but his reasons for doing so are slight. With this law he finds the mean of three observations, regarding it as corresponding to an ordinate which divides a curve $u=\varphi(x_1)\varphi(x_2)\varphi(x_3)$ into two equal parts. His result is as follows: Let $M_1$, $M_2$ and $M_3$ be the three measurements, of which $M_1$ is the least; let $M_1+x$ be the mean; it is required to find $x$. Put $M_2-M_1=p$ and $M_3-M_2=q$: then $x$ is given by**

$$x = p + \frac{1}{m}\log.(1 + \tfrac{1}{2}e^{-mp} - \tfrac{1}{2}e^{-mq}),$$

**in which $m$ is a constant depending upon the precision of the observation. Laplace then shows that this value cannot agree with the rule of the arithmetical mean, and he computes a table for finding $x$ for certain given ratios of $q$ to $p$. For instance**

$$\begin{aligned} &\text{if } q = 0.0\,p \text{ ---------- } x = 0.860\,p,\\ &\quad q = 0.1\,p \text{ ---------- } x = 0.894\,p,\\ &\quad q = 0.2\,p \text{ ---------- } x = 0.916\,p,\\ &\quad q = 0.3\,p \text{ ---------- } x = 0.932\,p,\\ &\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

158 *Mansfield Merriman—List of Writings relating*

**Thus if three measurements of an angle give**

$$M_1 = a^\circ\, b'\, 0'', \quad M_2 = a^\circ\, b'\, 40'' \quad \text{and} \quad M_3 = a^\circ\, b'\, 50'',$$

**we have $p=40''$, $q=10''$ and $q=0.25p$; then from the table $x=37''$ and the adjusted result is $a^\circ\, b'\, 37''$, while by the usual rule of the average we would have $a^\circ\, b'\, 30''$. By Laplace's table the mean will lie nearer to the two observations which most nearly agree, than in the common method.**

**For remarks on this memoir see Todhunter, *History of Probability*, p. 469, and Glaisher, *Mem. Astron. Soc. Lond.*, 1872, Vol. XXXIX, pp. 121–123.**

Laplace's treatment of the 3 data-points problem focuses more on the posterior distribution than on the likelihood implicit in it; thus his approach does not maximize the likelihood per se (see Edwards p. 221-223 where he distinguishes between the method of Maximum *Likelihood* and the Method of Maximum *Probability*). And so, given we are in a chapter dealing only with likelihood. we will not replicate his worked examples. Instead we will use another lesser known worked example (by Daniel Bernoulli – it is difficult to keep track of the various Bernoulli's - there were so many of them) that has the likelihood as the ultimate target, even if it does not call it by that name.

Whereas you might be tempted to call the product $\phi(x_1)\phi(x_2)\phi(x_3)$ a Likelihood and to just go ahead and maximize it, that is not how Laplace saw it. Rather, because he assumed a flat prior for the unknown dispersion parameter $\mu$, it integrates out when he forms the posterior distribution for $x$, and so he has a *probability density function* for the posterior distribution of $x$ that (apart from any normalizing constants) is the same as the *likelihood function* for $x$.

The distinction between this example and the one by Daniel Bernoulli is very clear: Bernoulli maximizes his $\phi(x_1)\phi(x_2)\phi(x_3)$ directly, and he never invokes a posterior distribution for $x$. To him, it is simply 'the x value that has *the highest probability*'.

The following, one of the first attempts to use ML directly, albeit with a very rudimentary (semicircular) error distribution, is from **Daniel Bernoulli**, written somewhere $\leq$ 1778, when, Kendall tells us, he was 78. Indeed, it seems that in his 1774 piece, Laplace is aware that both Lagrange and Daniel Bernoulli had 'considered the same problem in manuscript memoirs that I have not seen', He added that "This announcement both added to the usefulness of the material and reminded me of my ideas on this topic. I have no doubt that these two illustrious geometers have treated the subject more successfully than I; however, I shall present my reflections here, persuaded as I am that through the consideration of different approaches, we may produce a less hypothetical and more certain method for determining the mean that one should take among many observations."

Even though we would all now agree that Laplace had the much more comprehensive approach (and ultimately even derived a different and more justifiable error distribution), it is worth starting with the simpler idea in Daniel Bernoulli's paper, which appeared in 1778. So here are some extracts from it, as well as an exercise on his Example 1. The entire piece is on the course webpage.

**10. First of all,** I would have every observer ponder thoroughly in his own mind and judge what is the greatest error which he is morally certain (though he should call down the wrath of heaven) he will never exceed however often he repeats the observation. He must be his own judge of his dexterity and not err on the side of severity or indulgence. Not that it matters very much whether the judgement he passes in this matter is fitting or somewhat flighty. Then let him make the radius of the controlling circle equal to the aforementioned greatest error; let this radius be $r$ and hence the width of the whole doubtful field $= 2r$.

11. After all these preliminaries it remains to determine the position of the controlling circle, since it is at the centre of this circle that the several observations should be deemed to be, as it were, concentrated. The aforesaid position is deduced from the fact that the whole complex of observations would occur more easily, and therefore more probably, for this location than for any other position of the circle. We shall have the true degree of probability for the whole complex of observations if we note the probability corresponding to the several observations that have been carried out and multiply all the probabilities by each other, just as we did in § 9. Then the product of the multiplication is to be differentiated and the differential put = 0. In this way we shall obtain an equation whose root will give the distance of the centre from any given point.

Put the radius of the controlling circle = $r$; the smallest observation = $A$; the second $A+a$; the third $A+b$; the fourth $A+c$, and so on; the distance of the centre of the controlling semicircle from the smallest observation = $x$, so that $A+x$ will denote the quantity which is most probably to be assumed on the basis of all the observations. By our hypothesis the probability for the first observation alone is to be expressed by $\sqrt{\{r^2-x^2\}}$; for the second observation by $\sqrt{\{r^2-(x-a)^2\}}$; for the third by $\sqrt{\{r^2-(x-b)^2\}}$; for the fourth by $\sqrt{\{r^2-(x-c)^2\}}$ and so on. Then I would have the several probabilities multiplied together according to the rules of the theory of probability, which gives

$$\sqrt{\{r^2-x^2\}}\times\sqrt{\{r^2-(x-a)^2\}}\times\sqrt{\{r^2-(x-b)^2\}}\times\sqrt{\{r^2-(x-c)^2\}}\times\ldots.$$

Finally, if the differential of this product is put = 0, the equation, by virtue of our hypotheses, gives the required value $x$ as having the highest probability. As, however, the aforesaid quantity is to be brought to its maximum value, it is obvious that its square will simultaneously be brought to the same state. So we can use, for ease of calculation, a formula which is composed entirely of rational terms, viz.

$$(r^2-x^2)\times\{r^2-(x-a)^2\}\times\{r^2-(x-b)^2\}\times\{r^2-(x-c)^2\}\times\ldots$$

and the differential is once more put = 0. For the rest, as many factors are to be taken as there were observations.

(...)

13. When we have three observations to deal with, viz. $A$; $A+a$ and $A+b$, we shall have three factors

$$\{r^2-x^2\}\times\{r^2-(x-a)^2\}\times\{r^2-(x-b)^2\},$$

for which we have to find the maximum value. If now these factors are actually multiplied together we shall obtain

$$\begin{gathered}r^6+2ar^4x-3r^4x^2-4ar^2x^3+3r^2x^4+2ax^5-x^6\\-a^2r^4-2ab^2r^2x+2b^2r^2x^2+2ab^2x^3-b^2x^4+2bx^5\\-b^2r^4+2br^4x-a^2b^2x^2-4br^2x^3-4abx^4\\+a^2b^2r^2-2a^2br^2x+4abr^2x^2+2a^2bx^3-a^2x^4\\+2a^2r^2x^2.\end{gathered}$$

If this expression is differentiated, and then after division by $dx$ is put = 0 to obtain the maximum value, the following general equation for any three observations whatsoever will result

$$\begin{gathered}2ar^4-6r^4x-12ar^2x^2+12r^2x^3+10ax^4-6x^5\\-2ab^2r^2+4b^2r^2x+6ab^2x^2-4b^2x^3+10bx^4\\+2br^4-2a^2b^2x-12br^2x^2-16abx^3\\-2a^2br^2+8abr^2x+6a^2bx^2-4a^2x^3\\+4a^2r^2x=0.\end{gathered}$$

The root of this equation, which is indeed of the fifth degree and consists of twenty terms, gives the distance of the centre of the controlling circle from the first observation, and the quantity $A+x$ gives the value which is most probably to be deduced from the three observations which have been made.

Example 1 {wording from Bernoulli}

Let us assume three observations $A$; $A+0.2$ and $A+1$, so that $a=0.2$ and $b=1$ and let the value to be assumed as most likely from these three observations be $A+x$. The common rule gives $x=0.4$. Let us see the new one which to my mind is more probable, and let us put $r=1$ (cf. paragraph 10, where he invokes a semi-circular probability density function). The following purely numerical equation results

$$1.92-0.32x-12.96x^2+4.64x^3+12x^4-6x^4=0,$$

the solution of which is approximately $x=0.44xx$, which exceeds the com-

monly accepted value by more than a tenth.

**Supplementary Exercise 3.1**

1. Plot the **log-**likelihood for Example 1 above, and visually obtain an approximate $x_{ML}$ value. Then use the `optim` or `optimize` function[4] in `R` (or its equivalent in your favourite package) to find the $x_{ML}$ value. Compare it with Bernoulli's solution of the quintic equation.

2. If, as does Bernoulli, we take $r = 1$, what variance does Bernoulli's 'distribution with a semi-circular pdf' have?

3. What value of the dispersion parameter in the Laplace Distribution[5] would lead to the same variance as is obtained with $r = 1$ in Bernoulli's distribution?

4. Using the Laplace Distribution with that same variance (i.e. matching the variances of the two error distributions), what is the ML value corresponding to the data in Bernoulli's Example 1?

For continuation of section 10 of Bernoulli, on the choice of $r$, see footnote[6]

{ (*For those interested*) **Continuing with Bernoulli's words...** }

This marked excess is due to the fact that the middle observation is much nearer to the first than to the third. From this it is easily deduced that the excess will be changed to a defect if the middle observation is nearer to the third than to the first, and that the nearer the middle observation is to the mean between the two extreme observations, the smaller will be this defect.

[4] `optim` recommended for $\geq 2$ dimensions.

[5] In the version described in Wikipedia, the dispersion parameter (what JH calls $\mu$) is replaced by a 'precision' parameter $b$)

[6] If you desire a rule on this matter common to all observers, I recommend you to suit your judgement to the actual observations that you have made: if you double the distance between the two extreme observations, you can use it, I think, safely enough as the diameter of the controlling circle, or, what comes to the same thing, if you make the radius equal to the difference between the two extreme observations. Indeed, it will be sufficient to increase this difference by half to form the diameter of the circle if several observations have been made; my own practice is to double it for three or four observations, and .to increase it by half for more. Lest this uncertainty offend any one, it is as well to note that if we were to make our controlling semicircle infinite we should then coincide with the generally accepted rule of the arithmetical mean; but if we were to diminish the circle as much as possible without contradiction, we should obtain the mean between the two extreme observations, which as a rule for several observations I have found to be less often wrong than I thought before I investigated the matter. Note by JH: today, with enough data, we would estimate both the dispersion parameter $r$ and the location parameter of primary interest.

To test this conjecture I retain the other values and change only the middle observation, as follows.

Example 2

Let $a$ now $= 0.56$, and as before $r = b = 1$. By the commonly accepted rule we shall have $x = 0.52$. Let us see what happens with ours. The equation of section 13 gives the following numerical equation

$$1.3728 + 3.1072x - 13.4784x^2 - 2.1244x^3 + 15.6x^4 - 6x^5 = 0.$$

which is approximately satisfied by $x = 0.51xx$. In accordance with our principles, the value of x is less than the arithmetical mean which is usually accepted, but the difference between the two is now quite small, viz. 0.00yy, exactly as I had anticipated would be the case. Hence it can also be seen that the greatest difference between the two estimates occurs when it so happens that two observations exactly coincide and only the third diverges. There are two cases, viz. when $a = 0$ and when $a = b$. I will expound the result in each case.

***Example* 3. Put $a = 0$, leaving the remaining denominations unaltered. Dividing by $2b - 2x$ we have the following numerical equation**

$$1 - 6x^2 - 2x^3 + 3x^4 = 0,$$

**which is approximately satisfied by $x = 0{\cdot}3977$, whereas the value of $x$ obtained from the common rule is $x = 0{\cdot}3333$. The former exceeds the latter by $0{\cdot}0644$. If, however, we put $a = b$ and divide by $2x$, the following equation results**

$$4 - 6x - 6x^2 - 10x^3 - 3x^4 = 0.$$

**This is approximately satisfied by $x = 0{\cdot}6022$, while the common value is $0{\cdot}6666$. So the difference between the two is once more $0{\cdot}0644$, but this time our new value is less than the common one, whereas previously it was greater. It is clear from this that our method takes better aim at a certain intermediate point than does the common method. Evidence of this sort does much to commend the method that I propose, and I will go a little more closely into this consideration, if so be that an *argumentum ad hominem* may be accepted in a matter which does not admit of mathematical demonstration.**

**16. If we combine the two cases in example 3, and suppose that six observations have been made, viz. $A, A, A + b$ and $A + b, A + b, A$, it is obvious that three observations support the value $A$ and the same number the value $A + b$. We see by § 12 that in this case both methods give the required mean value as $A + \frac{1}{2}b$, or for example 3, $A + 0{\cdot}5$; or, omitting the constant quantity $A$, simply $0{\cdot}5$. This value, derived from the six observations combined, will not be doubted by anyone. Now let us divide these six observations into two other triads, namely $A, A, A + 1$ and $A + 1, A + 1, A$. In this case, rejecting once more the quantity $A$, the commonly accepted rule gives for the first triad $0{\cdot}3$ and for the second $0{\cdot}\dot{6}$, both differing, the first by defect and the second by excess, by $0{\cdot}1\dot{6}$ from the mean $0{\cdot}5$. So for either triad of observations taken separately the common theory involves an error of $0{\cdot}1\dot{6}$, while ours involves an error of $0{\cdot}1022$, which is notably smaller. A great deal more evidence of this kind could be adduced to give further support to our fundamental argument; but I am afraid I should appear immoderate if I went on extending something which cannot be settled with certainty and absolute perfection. We have no higher aim than to be able to distinguish what is more probable from what is less.**

## 3.2 Supported range

The choice of critical value is much less standardized or conventional than say the one for a significance test, or confidence level, or a highest posterior density.

Fig 3.4 (based on 20/50) vs. Fig 3.3 (based on 4/10): the authors don't say it explicitly, but the sharpness of the likelihood function is measured formally by the second derivative at the point where it is a maximum.

## 3.3 The log likelihood

The (log-)likelihood is invariant to alternative monotonic transformations of the parameter, so one often chooses a parameter scale on which the function is more symmetric.

## 3.4 Censoring in follow-up studies

See applications below. These will be more relevant after we consider all of the fitting options, and the benefits/flexibility of a Likelihood approach.

## 3.5 Other fitting methods

We mentioned earlier that the method of least squares does not make an explicit assumption about the distribution of the deviations from the target or even that the observed data are a sample from a larger universe. Another older method, that does not make explicit assumptions about the variations about the postulated means, is the method of minimum chi-square. It was used for fitting simpler models for dose response data involving count data. This minimum chi-square criterion does not lead to simple methods of estimation, or to estimators with easily derived sampling distributions. Nevertheless, it is one of the thee methods (the others are ML – which requires a fully specified model for the variations, and LS, that does not) used in the java applet `http://www.biostat.mcgill.ca/hanley/MaxLik3D.swf`. The applet allows you to fit a linear model to the above-described 2-point data, and to monitor how the log-likelihood, the sum of squared deviations, and the chi-square goodness of fit statistics vary as a function of the entertained values of $\beta$.

The applet shows that the LS method which measures lack of fit on the same scale that the $y$'s are measured on (cf the two red lines). The min-$X^2$ method – applied to $y$'s that represent counts or frequencies, is similar, in that the "loss function" is $\sum(y-\hat{y})/\hat{y}^2$. The criterion for the ML fitting of a Poisson model is very different, in that it is measured on the probability or log-probability scale, a scale that is shown in blue, and projecting out from the $x-y$ plane.

**Under some Normal models with homoscedastic variation, the LS and ML methods give the same estimates for the parameter(s) that make up the mean**. If $y|x \sim Normal(\mu_x, \sigma^2)$, then Lik $= \prod(1/\sigma)\exp[-\{(y_i - \beta x_i)^2/2\sigma^2\}]$. This is maximized when the exponentiated quantity is minimized. The minimization is the same one involved in the LS estimation.

**Supplementary Exercise 3.2**.

Grouped Normal data, from section 12.2 of Fisher's paper.[7] Three hundred observed measurement errors ($\epsilon$'s) from a $N(0,\sigma)$ distribution are grouped (binned) in nine classes, positive and negative values being thrown together as shown in the following table:-

| Bin | 0-1 | 1-2 | 2-3 | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 | 8-9 | All |
|---|---|---|---|---|---|---|---|---|---|---|
| Frequency ($f$) | 114 | 84 | 53 | 24 | 14 | 6 | 3 | 1 | 1 | 300 |

Estimate $\sigma^2$ ...

1. as $(1/300)\sum f \times \epsilon_{mid}^2$. Note that we estimate it using a divisor of $n$ rather than $n-1$, since we do not have to estimate $\mu$ : the errors are deviations from *known* values, so $\mu = 0$ (structurally).
2. Using Sheppard's correction for the grouping, i.e, by subtracting $w^2/12$, where $w$ is the width of each bin, in this case 1. Incidentally, can you figure out why Sheppard subtracts this amount? Shouldn't grouping *add* rather than subtract noise?
3. Using the method of Minimum $\chi^2$.
4. By directly maximizing the (log) Likelihood, either by plotting or tabulating it and zooming in on the maximum, or by supplying the log-likelihood to a function such as `optim` or `optimize`, or by using a root-finding approach, such as that of Newton-Raphson, to obtain the root of the equation $dLogL(\sigma)/d\sigma = 0$, or $dLogL(\theta)/d\theta = 0$, where $\theta = \sigma^2$ or $\log(\sigma)$ or $\log(\sigma^2)$ or some other transform of $\sigma$.
5. Using the method of Maximum Likelihood, but using the 'EM' algorithm. For a compact description of the EM algorithm, see section 2.2 in the

[7] On the Mathematical Foundations of Theoretical Statistics, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 222 (1922), pp. 309-368. available here: `http://www.biostat.mcgill.ca/hanley/bios601/Likelihood/Fisher1922.pdf`

paper 'A statistician plays darts' by Tibshirani (Jr), Price and Taylor. *Hint*: be careful when you iteratively re-estimate the expected values: since we are focused on $\sigma^2$, at issue is the expected values of the squares of the values inside each bin, not the values themselves.

**Supplementary Exercise 3.3**.

Frequency data, the subject of Galton's 1894 correspondence with the Homing News and Pigeon Fanciers' Journal.[8]

*Significance* magazine ( `http://www.significancemagazine.org/`) has special Galton coverage in 2011, the 100th anniversary of his death – Galton was born in 1822, the same year, he noted himself, as the geneticist Gregor Mendel. In the article "Sir Francis Galton and the homing pigeon", Fanshawe writes...

> "The results for the 3,207 "old birds" are shown in the table. The table shows the proportion of birds in each category. Galton suggests summarising the figures by their mean and "variability", which he estimates as 976 and 124 yards per minute respectively. It is not clear which quantity Galton calls the "variability" – his figure appears too small to be a standard deviation.
>
> The second row of figures are Galton's, and arise from the proportions that would be expected by approximating the original data by a Normal distribution. The fit appears extremely good."

Using these frequencies and bin-boundaries[9] from the journal article, and the Normal distribution assumed by the journal and by Galton,

| Bin | -5 | 5-6 | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 | 12-13 | 13-14 | 14+ | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Freq | 22 | 43 | 164 | 284 | 598 | 645 | 683 | 396 | 132 | 120 | 120 | 3207 |

estimate $\mu$ and $\sigma$, and, where possible, using SE($\hat{\mu}$) and SE($\hat{\sigma}$), [10] form symmetric (frequentist) confidence intervals for $\mu$ and $\sigma$,

1. by concentrating the frequencies at the midpoints, and at suitably chosen values for the two open-ended categories
2. via the method of Minimum $\chi^2$, and
3. via the method of Maximum Likelihood. Then
4. determine whether Fanshawe is correct: i.e., is the "124 yards" measure of "variability" indeed too small to be a standard deviation (SD)?
5. Galton rarely used the SD.[11] Instead he – as Gosset often did – used the Probable Error (PE), i.e., 1/2 the IQR.[12]

   In a Gaussian distribution, how much smaller/larger is the PE than the SD?

   Does this factor explain how Galton arrived at the 124 yards per minute?

The sample size is so large here that the symmetric (z-based) CI for $\sigma$ is quite accurate. By what if the sample size were quite small? In this case you could use the tails of the (non-symmetric) distribution of the distribution of $s^2$ to derive an asymmetric first-principles frequentist confidence interval for $\sigma^2$, and by transformation, for $\sigma$.[13]

**Supplementary Exercise 3.4: Did Isaac Newton know the '$\sqrt{n}$ Law'?**

Both in his recent book *Seven Pillars of Statistical Wisdom*, and in his 1977 JASA article *Eight Centuries of Sampling Inspection: The Trial of the Pyx* [14] Stephen Stigler explains how the quality control scheme used by the Royal Mint was based on the – flawed – 'law' (or belief) that, for i.i.d. random variables $y \sim ?(\mu, \sigma)$,

$$\text{SD}\left[\sum_1^n y_i\right] = n \times \sigma, \quad \text{and SD}\left[\bar{y} = (1/n) \times \sum_1^n y_i\right] = \frac{\sigma}{n}.$$

It appears they were unaware that it is variances that add, not SD's, and that SD's of sums or means scale with $\sqrt{n}$,, not $n$.

[8] Material (3p of journal, Fanshawe's article, and `R` code) available here `http://www.biostat.mcgill.ca//hanley/bios601/Likelihood/.`

[9] 5-6 means 500-600 yards per minute, etc.

[10] Since $s^2 \sim (1/\nu) \times \sigma^2 \times ChiSq(d.f. = \nu)$, then $\text{Var}[s^2] = (1/\nu^2) \times \sigma^4 \times 2\nu$. By Delta method,

$$\text{Var}[s] \approx \text{Var}[s^2] \times \left\{\frac{ds}{ds^2}\right\}^2 = \underbrace{(1/\nu^2) \times \sigma^4 \times 2\nu} \times \underbrace{(1/4) \times \{1/\sigma^2\}^{-1}} = (1/\nu^2) \times \sigma^2,$$

so $\text{SE}[s] \approx (1/\nu^2)^{-1/2} \times \sigma$.

[11] Karl Pearson was the one who promoted the SD.

[12] Thus, it is equally probable (50:50) for an observation to be more/less than this amount from the middle (truth).

[13] Hint: (taking some semantic liberties) a first-principles 100(1-$\alpha$)% frequentist CI, $(L, U)$ for $\theta$ is the pair of *statistics* $(L, U)$, such that $\text{Prob}(\hat{\theta} \geq \hat{\theta}_{\text{observed}} \mid \theta = L) = \alpha/2$ and $\text{Prob}(\hat{\theta} \leq \hat{\theta}_{\text{observed}} \mid \theta = U) = \alpha/2$.

[14] `http://www.biostat.mcgill.ca/hanley/c323/pyx.pdf`

Understanding which version is correct is key in quality control, and in setting critical values to flag when the process has (been) altered – or in this application, to detect cheating. The Mint based their tolerances (they called it the 'remedy', because those who were caught had to pay back.) Stigler continues...

> "The most illustrious master of the Mint was Isaac Newton, who served in that position from 1699 until his death in 1727, having been Warden since 1696. It is natural to ask if this great scientist might have benefited from the remedy during his tenure with the Mint, either with or without the approval of the Crown. Such evidence as exists seems to indicate that he did not. Newton's manuscripts show him to have had a conscientious concern with the integrity of the Mint's product and to have put particular emphasis on the reduction of the variation in individual pieces. He was acutely aware of the possible benefit of the remedy to importers (through their returning of overweight coins for remelting) and sought to maintain a high standard for accuracy.
>
> Was Newton's grasp of statistical theory sufficient to allow him to take advantage of the remedy, had he wished? His published works have little to say about probability (Sheynin 1971), and his interpretations of data relevant to his physical theories were sometimes more influenced by his preconceptions than by statistical analysis (Westfall 1973). Nonetheless, in order to obtain some insight into Newton's understanding of the behavior of sample means, we shall examine a work he wrote while master of the Mint, where a**n interval estimate of a mean** is presented in a situation conceptually similar to the trial of the Pyx.
>
> The work in question was Newton's last, The *Chronology of Ancient Kingdoms Amended*, published post-humously in 1728. It had been written primarily as a defense of a short chronology of ancient history he had prepared for the Princess Caroline which had been printed in an unauthorized version in France in 1725. As part of his investigation of the dates of ancient events, Newton sought to show that earlier chronologists who had reckoned the average reign of ancient kings as being 35 to 40 years had been in error. By appealing to data from the more accurately recorded periods of history he meant to determine **an estimate of the mean length of a reign**, a quantity that could in turn be used to estimate lengths of eras when numbers of kings were more accurately recorded than were years. He wrote:
>
> For by the ordinary course of nature Kings Reign, one with another, about eighteen or twenty years a-piece: and if in some instances they Reign, one with another, five or six years longer, in others they Reign as much shorter: eighteen or twenty years is a medium. So the eighteen Kings of Judah who succeeded Solomon, Reigned 390 years, which is one with another 22 years a-piece ... (Newton 1728, p. 52).

Newton went on to list the data for eleven more periods of time; these data are presented in tabular form in Table 2.

**2. Newton's Data on Length of Kings' Reigns**

| Kingdom | Number of kings | Years | Mean reign (Newton) | Mean reign |
|---|---|---|---|---|
| Judah | 18 | 390 | 22 | 21.67 |
| Israel | 15 | 259 | 17¼ | 17.27 |
| Babylon | 18 | 209 | 11⅔ | 11.61 |
| Persia | 10 | 208 | 21 | 20.80 |
| Syria | 16 | 244 | 15¼ | 15.25 |
| Egypt | 11 | 277 | 25 | 25.18 |
| Macedonia | 8 | 138 | 17¼ | 17.25 |
| England (1066–1714) | 30 | 648 | 21½ | 21.60 |
| France (first 24) | 24 | 458 | 19 | 19.08 |
| France (second 24) | 24 | 451 | 18¾ | 18.79 |
| France (last 15) | 15 | 315 | 21 | 21.00 |
| France (all) | 63 | 1224 | 19½ | 19.43 |

*Source*: Adapted from Newton (1728, pp. 52–3).

After presenting the data, he repeated his contention that in "the course of nature," the reign of kings should be reckoned "at about eighteen or twenty years a-piece" (Newton 1728, p. 54). Where did Newton get his interval estimate of the mean? That he repeated "eighteen or twenty years" three times rather than even once quoting nineteen as a "medium" would indicate that he did indeed think of his estimate as an interval. Newton presented no calculations, but it is instructive to use the twelve average reigns of the third column of Table 2 to compute an overall mean plus or minus an estimated standard error of the mean, or $\bar{x} \pm \sqrt{12}$, in the manner of present day physicists. We find $19.10 \pm 1.01$. An alternative analysis would (a) note that the kings of France should not be included twice, eliminating the twelfth row of the table, (b) use the actual means of column 4 rather than Newton's approximations, (c) use a weighted (by number of kings) mean of the eleven

remaining numbers in column 4, and the corresponding maximum likelihood estimate of the standard deviation of this weighted mean. This would give 19.03 ± 1.01. Now, the point is not that Newton did either of these calculations. We can be certain he did not, although he may have calculated the mean of the third column. The point is that the ad hoc or intuitively derived interval he did present was in rough accord with something reasonable, i.e., roughly a 65 percent confidence interval. Had Newton specified an interval on the same basis as that apparently used to determine the remedy in the trial of the Pyx, he would have made a very different statement. As we have seen, the limits set for the trial of the Pyx correspond to values such that a sizable proportion (about 95 percent) of single coins fall between them; these limits were then, in effect, applied to average weights of large numbers of coins. An analysis of the length of reigns of the kings of England, data easily available to Newton, suggests that an interval such as 19 ± 11 includes about 65 percent of king's reigns. Yet Newton, who knew he would be applying his statement to aggregates of reigns, based his analysis upon aggregates and gave an interval appropriate to aggregates. Newton must have realized that 19 ± 1 would only contain a small fraction of individual reigns, and **it seems safe to say on the basis of this that he had at least an approximate intuitive understanding of the manner in which the variability of means decreased as the number of measurements averaged increased**. The application of this understanding to the trial of the Pyx would have been well within Newton's capabilities.

***Exercise for bios601:*** (1) What do you think of Stigler's evidence, and his conclusion? (2) Either analytically or numerically, try to re-produce the 19.03 ± 1.01, stating any additional assumptions made.

**Supplementary Exercise 3.5**.

Refer again to the article *Maternal Lead Levels after Alterations to Water Supply* in The Lancet 1981 `https://www-sciencedirect-com.proxy3.library.mcgill.ca/science/article/pii/S0140673681903846` .

1. The authors do not explain how exactly they handled the 9 tabulated intervals for blood lead concentrations. Argue why the boundaries for these 9 intervals may well have been `lower = c(0, seq(0.255,1.655,0.2) )` and `upper = c(seq(0.255,1.655,0.2), Inf)`
2. Fit two separate log-normal models to the reported 1977 and 1980 distributions, and plot the fitted distributions on the original scale.
3. How close do the fitted medians come to the reported ones?
4. How would you obtain a standard error for the estimated difference in medians?
5. Indicate how you could include the two separate models within one larger regression model that includes (0,1) term(s) for 'post-intervention'
6. Explain how you might fit this larger 'all-in-1' model (a) using 'canned' software (b) your own homegrown ML-fitting routine. [you don't have to actually develop the code; just the approach.]

## 3.6 Other Applications: exercises

### 3.6.1 2 datapoints and a model

One has 2 independent observations from the (no-intercept) model

$$E[y|x] = \mu_{y|x} = \beta \times x.$$

The $y$'s might represent the total numbers of typographical errors on $x$ randomly sampled pages of a large document, and the data might be $y = 2$ errors in total in a sample of $x = 1$ page, and $y = 8$ errors in total in a separate sample of $x = 2$ pages. The $\beta$ in the model represents the mean number of errors per page of the document. Or the $y$'s might represent the total weight of $x$ randomly sample pages of a document, and the data might be $y = 2$ units of weight in total for a sample of $x = 1$ page, and $y = 8$ units for a separate sample of $x = 2$ pages. The $\beta$ in the model represents the mean weight per page of the document.

We gave this 'estimation of $\beta$' problem $\{ (x, y) = (1, 2) \ \& \ (2, 8)\}$ to several statisticians and epidemiologists, and to several grade 6 students, and they gave us a variety of estimates, such as $\hat{\beta} = 3.6$/page, 3.33/page, and 3.45!

**Supplementary Exercise 3.6**

How can this be? The differences have to do with (i) what model they (implicitly or explicitly) used for the variation of each $y \mid x$ around the mean $\mu_{y|x}$ and (ii) the method of fitting.

1. From 1st principles derive the Method of Moments (MM), Least Squares (LS) and (if possible the) Maximum Likelihood (ML) estimators[15] of $\beta$ when
   (a) $y \mid x \sim ???(\mu_{y|x})$
   (b) $y \mid x \sim Poisson(\mu_{y|x})$
   (c) $y \mid x \sim N(\mu_{y|x}, \sigma)$ [assume $\sigma$ is known]
   (d) $y \mid x \sim N(\mu_{y|x}, \sigma^2 = x \times \sigma_0^2)$ [assume $\sigma_0^2$ is known]
2. Where possible, match the estimators with the various numerical estimates above.
3. One of the numerical estimates came from another fitting method, namely the (now seldom-used) method of Minimum Chi-square, which seeks the value of $\beta$ that minimizes $\sum \frac{(O-E)^2}{E} = \sum \frac{(y-\beta x)^2}{\beta x}$ in this example. Verify that the one remaining estimate of unknown origin is in fact obtained using this estimator.

See the (Flash) applet on `http://www.biostat.mcgill.ca/hanley/software/`

One of the messages of this exercise is that for one to use a likelihood approach, one must have a fully-specified probability model so that one can write the probability of each observed observation.

And, with different distributions of the $y$'s around the mean $\mu_{y|x} = E(y|x) = \beta \times x$, the probabilities (and thus the overall likelihood, and its maximum, would be different.

### 3.6.2 Application: Estimation of parameters of gamma distribution fitted to tumbler mortality data [interval-censored and right-censored data].

The important but seldom-visited article "Tumbler Mortality" by Brown and Flood in JASA in 1947 shows the "survival" of tumblers (Free Online Dictionary: a. A drinking glass, originally with a rounded bottom. b. A flat-bottomed glass having no handle, foot, or stem.) in a cafeteria. The article is available under Resources for Epidemiology and for Statistical Models. Note that whereas the authors used the word *truncation* for the observations on tumblers that were still in service at the end of the test, we would use the word '*right-censored*' today. Since inspections were only once a week, the lengths of service of the items that *did* fail are also censored, but *within* [in most instances] a 1-week interval. This type of censoring is called '*interval-censoring*'.

[15] [if the estimator doesn't dependent on the model, just say so.

**Supplementary Exercise 3.7**

Using the data in Table 1 for the article [contained in the various versions of the R code in the same link] , determine the MLEs of the two parameters of the gamma distribution, and compare them with those obtained by the original authors [they use a slightly approx. ML method]. Do so in two ways (they should give the same likelihood function, and thus the same MLEs):

1. using an *unconditional* approach, based on 549 contributions – one per tumbler, with each tumbler considered in isolation from the other 548 – so that each failure (unconditional) contributes one term and each (ULTIMATELY) censored observation (also unconditional) contributes another. [of course, there are 'multiplicities'; thus, instead of a sum of 549 log-likelihhods, you can use the multiplicities (and multiplication of a 1-item log-likelihood by the multiplicity) to reduce the computation].
2. using the binomial structure created by the authors: a row that has $n$ exposed tumblers *that* week (and that only considers whether the tumbler that began that week survived *that* week) makes $n$ Bernoulli-based log-likelihoods, (or 1 Binomial-based log-likelihood) for that *week*.

This exercise shows that there is more than 1 way to set up the likelihood.

### 3.6.3 Application: Estimation of parameters of a parametric distribution fitted to avalanche mortality data [all observations are censored – either left-censored or right-censored. Such data are often referred to as "current-status" data].

One example of status-quo data is data from a cross-sectional survey of menarche status in girls, or the prevalence of decayed-missing-or-filled (DMF) teeth (or say permanent dentition) in dental public health, or HIV prevalence in the general population or in specific sub-populations, such as partners of persons who contracted HIV though blood donations.

Another is the data from the Avalanche Survival Chances by Falk et al. in the journal *Nature* in 1994. The article and the data are available under Resources.

The authors fitted a non-parametric model. We will discuss in class which parametric models (or mixtures of different parametric models) might make sense. But, just to get some practice with this type of data, we will start with a very simply one, even if we know a priori it is too simplistic.

**Supplementary Exercise 3.8**

Using the raw data, and (for now) the simplistic parametric model we agreed on in class, determine the MLEs of the two parameters of this gamma distribution, and compare the fit with the fit of the smooth and non-parametric curves shown in the authors' article.

### 3.6.4 Application: Genetics of Blood Groups

[Premiminary] Questions 1-6 below were set by Olli Saarela for course EBIB607. See (on the BIOS601 website, under Resources for Likelihood) the relevant pages from section 3.7 of Edwards' book, where his main purpose is to illustrate the Likelihood ratio

$$\frac{Prob(data \mid \text{single tri-allelic locus})}{Prob(data \mid \text{two bi-allelic loci})},$$

and section 6.8 where he addresses the parameter fitting.

The example is from the book "Likelihood" by Edwards (1972) and originates from Bernstein (1924), who discovered the inheritance pattern of ABO blood groups. Formerly, it was thought that the ABO bloodgroup phenotype was determined by two biallelic loci with alleles $\{A, a\}$ and $\{B, b\}$, respectively. Under this model, the individual has a bloodgroup phenotype 'AB' when both alleles A and B are present, bloodgroup 'A' when allele A is present and allele B is not present, bloodgroup 'B' when allele B is present and allele A is not present, and finally, bloodgroup 'O' when neither A or B is present. The 6 sub-questions Saarela posed were the following [JH would immediately turn to trees to see what was going on]

1. List the possible genotypes under the two loci model (there are 9 in total).
2. In a population of 502 individuals, 42.2% had the observed bloodgroup phenotype 'A', 20.6% had the bloodgroup 'B', 7.8% had the bloodgroup 'AB', and the remaining 29.4% had the bloodgroup 'O'. Let us denote by $P(A = 1)$ the (marginal) probability that the allele A is present, and $P(B = 1)$ the (marginal) probability that the allele B is present. Recall from the lecture notes that the probabilities of mutually exlusive alternatives are additive, so that

   $$P(A = 1) = P(A = 1 \text{ and } B = 1) + P(A = 1 \text{ and } B = 0)$$

   and

   $$P(B = 1) = P(B = 1 \text{ and } A = 1) + P(B = 1 \text{ and } A = 0).$$

   Using the observed data, and estimating the probabilities by the corresponding empirical proportions, calculate the probabilities $P(A = 1)$, $P(A = 0)$, $P(B = 1)$ and $P(B = 0)$.
3. Under the two loci model, and assuming the loci to be independent (not in *linkage disequilibrium*), we can calculate the probability of each bloodgroup phenotype from the above four marginal probabilities; for instance the probability of the 'AB' phenotype is given by

   $$P(\text{'AB'}) = P(A = 1 \text{ and } B = 1) = P(A = 1)P(B = 1).$$

   Using the observed data, calculate the probabilities of the four bloodgroup phenotypes under this model.
4. An alternative model is that the ABO bloodgroup is in fact determined by a single triallelic locus with the alleles $\{A, B, O\}$. List the possible genotypes under this model (there are 6 in total), and the bloodgroup phenotype related to each of these.
5. Suppose that from the observed data, the allele frequencies (that is, the probabilities that a copy of the particular allele is inherited from a given parent) under the single locus model were estimated as $P(A = 1) = 0.2945$, $P(B = 1) = 0.1547$ and $P(O = 1) = 0.5508$. Using these estimates, calculate the four phenotype probabilities $P(\text{'AB'})$, $P(\text{'A'})$, $P(\text{'B'})$ and $P(\text{'O'})$ under the single locus model. (Hint: add the probabilities of different possible genotypes corresponding to a particular phenotype and assume that the maternal and paternal alleles are inherited independently).
6. Which model fits better to the observed phenotype proportions?

The 607 exercise illustrates the competing hypotheses (probability models) as to how the ABO bloodgroup phenotypes are determined.

Edwards used the Likelihood ratio to show the far greater support for the single tri-allelic locus hypothesis ($H_1$) over than the two bi-allelic loci hypothesis ($H_2$)

Part (5) of the 607 exercise gives, without the technical details, the (data-based) estimates of $P(A = 1)$, $P(B = 1)$, & $P(O = 1) = 1 - \{P(A = 1) + P(B = 1)\}$. The 607 students are then asked to calculate the 4 expected frequencies under this model, and to compare these fitted frequencies with those calculated under the other model.

When he uses $\hat{p}$, $\hat{q}$ & $\hat{r}$ to calculate the Likelihood ratio, in section 3.7 of his book, Edwards simply says that $p = P(A = 1)$, $q = P(B = 1)$, & $r =$

$P(O = 1)$ '*must be (were) obtained by iteration*'. In his section 6.8 (ML with a constraint among parameters) he gives several suggestions for estimation: one is to plot the log-Likelihood 'surface' as contours on a 'Streng diagram' (a triangle used for 3-nomial probabilities) and zoom in numerically; another is what he calls 'Fisher's method'; another is to use the Lagrangian multiplier method; and the last is the 'counting method', which he works through, starting on p139.

Nowadays, the simplest way is to simply write down the log-likelihood and directly maximize it. It is a 4-nomial, (effectively) 2-parameter log-likelihood, and thus easily maximized. Ignore the slightly 'non-integer' observed frequencies.[16]

| Phenotype | Genotype(s) | Theoretical Proportion | Observed Frequency* |
|---|---|---|---|
| '$A$' | $AA,\ AO$ | $p(p+2r)$ | $502 \times 0.422$ |
| '$B$' | $BB,\ BO$ | $q(q+2r)$ | $502 \times 0.206$ |
| '$AB$' | $AB$ | $2pq$ | $502 \times 0.078$ |
| '$O$' | $OO$ | $r^2$ | $502 \times 0.294$ |

**Supplementary Exercise 3.9**

1. Find $\hat{p}_{ML}$ and $\hat{q}_{ML}$ (and thus $\hat{r}_{ML}$ ) directly.

   *Hint*: To avoid inadmissible parameter regions ($< 0$ or $> 1$), you could re-express the 3 parameters as

   $$p = \frac{\exp[\alpha]}{\exp[\alpha]+\exp[\beta]+1};\ q = \frac{\exp[\beta]}{\exp[\alpha]+\exp[\beta]+1};\ r = \frac{1}{\exp[\alpha]+\exp[\beta]+1},$$

   and maximize over $\alpha$ and $\beta$.

2. Read through Edwards' description of the 'counting method', and comment on whether it qualifies as an application of what (in 1977) became known as the 'EM algorithm.'

[16] *Edwards tells us that 'the data are for Japanese in Korea, and are due to Kirihara. I cannot find and set of frequencies adding to 502 which would lead to the given proportions, so I have made the ensuing calculations in terms of the proportions themselves'. But for precision purposes, the 502 does matter, so JH suggests using the 'slightly non-integer' frequencies. Some software, expecting binomial or multinomial frequencies, would object, but you do not need to tell the `optim` function that the supplied frequencies in the log-likelihood are non-integer.

### 3.6.5 Application: Distribution of Observations in a Dilution Series.

(Again, text from (section 12.3 of) Fisher's 1922 paper). An important type of discontinuous distribution occurs in the application of the dilution method to the estimation of the number of micro-organisms in a sample of water or of soil.[17] [*note from JH: for simplicity, we replace some of Fisher's notation, by letting the expected or average concentration of micro-organisms be $\mu$ per cubic centimetre – Fisher had n per cubic centimetre.*] The method here presented was originally developed in connection with Mr. Cutler's extensive counts of soil protozoa carried out in the protozoological laboratory at Rothamsted, and although the method is of very wide application, this particular investigation affords an admirable example of the statistical principles involved.

In principle the method consists in making a series of dilutions of the soil sample, and determining the presence or absence of each type of protozoa in a cubic centimetre of the dilution, after incubation in a nutrient medium. The series in use proceeds by powers of 2, so that the frequency of protozoa in each dilution is one-half that in the last. The frequency at any stage of the process may then be represented by

$$\mu_x = \frac{\mu}{2^x},$$

when $x$ indicates the number of dilutions. Under conditions of random sampling, the chance of any plate made from the $x^{th}$ dilution receiving 0, 1, 2, 3 protozoa of a given species is given by the Poisson series

$$e^{-\mu_x}\left(1,\ \mu_x,\ \frac{\mu_x^2}{2!},\ \frac{\mu_x^3}{3!},\ \ldots\right),$$

and in consequence the (expected) proportion of sterile plates at dilution $x$ is

$$p_x = e^{-\mu_x},$$

and of fertile plates

$$q_x = 1 - e^{-\mu_x}.$$

In general we may consider a dilution series with dilution factor $a$ so that

$$\log p_a = -\frac{\mu}{a^x},$$

and assume that $n_a$ plates are poured from each dilution. The object of the method is to estimate $\mu$ from a record of the sterile and fertile plates. We can

[17] See related article, from Significance Magazine, on Petri Dishes – under Resources.

do so by treating the observed number of fertile plates at dilution $x$, say $n_x^+$, out of the $n_x$ poured from dilution $x$, as a realization of a binomial random variable, and thus writing the overall log likelihood as

$$logLik = \sum_x n_x^+ \times \log(1-p_x) + (n_x - n_x^+) \times \log(p_x),$$

when the summation is over the different dilutions.

**Supplementary Exercise 3.10**

Estimate $\mu$ from the following dilution series data [a=0.25, 0.5 denote 4 and 2 times the original concentration(a=1), and a=2, 3, ... denote 1/4, 1/8 ... times the original concentration(a=1)]:

| Dilution (a): | 0.25 | 0.5 | **1** | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|---|---|
| No. of plates ($n_a$) : | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| No. of fertile plates ($n_a^+$): | 5 | 5 | 5 | 5 | 4 | 3 | 2 | 2 | 0 | 0 |

### 3.6.6 Application: Pooled testing:- old and new uses

The following excerpts are from a 1976 article "Group testing with a new goal, estimation", in Biometrika, 62, 1, p. 181 by authors Sobel and Elashoff. They begin by referring to Dorfman, whose article, in the Annals of Mathematical Statistics, 1943, first used the ideas of group testing, with a binomial model, to reduce the number of medical tests necessary to find all members of a group of size N that have the syphilis antigen. They continued...

> Another aspect of the group-testing problem arises when one is interested not in the *classification of all the individuals* but in the *estimation of the frequency* of a disease, or of some property, when group-testing methods can be used. Given a random sample of size N, say, from a binomial population, the best estimate of the prevalence rate p, in the sense of minimizing the mean square error, will be obtained by testing each unit separately. However, if N is large and the tests are costly, then a different criterion, that includes testing costs, may indicate that group-testing designs should be used. We might expect benefits from group testing to increase as p decreases.
>
> [....] Example: Rodents are collected from the harbour of a large city, and, after being killed, dissected, etc., their liver is to be carefully examined under a microscope for the presence or absence of a specific type of bacterium. The goal of the study is to estimate the proportion p of rodents that carry this bacterium using an economical experimental design. In this application the cost of obtaining the animals is negligible compared to the cost of testing, i.e. the microscopic search. It was proposed that an economical design to estimate p should be possible by combining in a single sample a small portion of the liver from each of several test animals and then carrying out a microscopic search on a homogeneous mixture of these liver portions. The problem is to find the best number, say A, of liver portions to combine and how to estimate the prevalence rate p from such a design. In addition, if this bacterial type is present in some particular tests, then the pathologists want to know whether they should carry out another test on a subset of these same animals or go on to test a new group of A animals.
>
> [...] Thompson (1962) estimated the proportion of insect vectors capable of transmitting asteryellows virus in a natural population of the six-spotted leafhopper, an aphid. Instead of putting one insect with a previously unexposed aster test plant, he puts several insects with one test plant, for economic reasons, and waits to see if the plant develops the symptoms of this virus. If it does, then at least one of these insects carried the virus; otherwise it is assumed that none carried it. The statistical problem is to choose an optimal number A of insects to be put with one test plant.

**Contemporary uses**: (can also Google *Minipool testing*)

The following text is an excerpt from Canadian Blood Services : Customer Letter #2005-18, 2005-05-17, entitled "Planned Measures to Protect the Blood Supply from West Nile Virus (WNV) - 2005 Season."

> Dear Colleague:
>
> West Nile season is approaching once again and this letter is to inform you about enhanced measures Canadian Blood Services has put in place to further protect the safety of the blood supply during the 2005 season.
>
> For the summer of 2005, Canadian Blood Services will again use single-unit testing (SUT) to enhance the sensitivity of the West Nile Virus nucleic acid test. Minipool testing (6 samples/pool) is used throughout the year.
>
> - In the summer of 2005, a 'trigger' will be used to initiate SUT. SUT will be initiated in a health region when a presumptive positive blood donor is detected using minipool testing, OR the

prevalence of recent confirmed human cases in the preceding two weeks exceeds 1/1,000 population in rural areas, or 1/2,500 in urban areas.

- SUT will cease in a health region when there have been no positive donors for two weeks or the occurrence of WNV cases in the population falls below the aforementioned population triggers.

**Supplementary Exercise 3.11**

Suppose that in order to estimate the prevalence ($\pi$) of a characteristic in a population, one tests $N$ randomly sampled objects by pooling them into $n_b$ batches of size $k$ (so that $N = n_b \times k$) and determining, for each batch, i.e. collectively, if at least one of its members is positive. Suppose that $n_{b+}$ batches are found to be positive. Develop estimators of $\pi$ using the method of moments, and using minimum $\chi^2$ and Maximum Likelihood criteria.

### 3.6.7 Application: Measuring one's accuracy at darts

In 2011, Tibshirani (junior!) et al.[18] published a very instructive essay. In addition to its innovative use of a personalized heatmap to show the optimal strategy for throwing darts, it provides an engaging example for teaching several statistical concepts and techniques, such as fast Fourier transforms, the EM algorithm, Monte Carlo integration, importance sampling, and the Metropolis Hastings algorithm. It is a delightful blend of the applied and the theoretical, the algebraic and the graphical.

It also continues the tradition of statisticians' fascination with the imagery of marksmen (Turner, 2010). In her chapter on metaphor and reality of target practice, Klein (1997) writes of 'men reasoning on the likes of target practice' and describes how this imagery has pervaded the thinking and work of natural philosophers and statisticians. Klein shows a frequency curve, by Yule, for 1,000 shots from an artillery gun in American target practice. Pearson used it in his 1894 lectures on evolution; he decomposed the frequency curve into two chance distributions centered slightly to the right and left of the target, gave reasons why this might occur, and used it to illustrate the interplay between random variation and natural selection. He also used it in his 1900 paper in one of the illustrations of his test of goodness of fit. Incidentally, Klein also reminds us of the origin of the term '*stochastic.*' In Liddell and Scott (1920) we find the following entries:

| | |
|---|---|
| $\sigma\tau o\chi o\varsigma$ | an aim, shot. a guess, conjecture. |
| $\sigma\tau o\chi a\sigma\mu\alpha$ | a missile aimed at a mark; an arrow, javelin. |
| $\sigma\tau o\chi a\sigma\tau\iota\kappa o\varsigma$ | able to hit: able to guess, shrewd, sagacious. |

Since the optimal aiming spot in darts – and thus the heatmap provided by the online applet – depends strongly on one's accuracy, much of the Tibshirani et al. article is devoted to the challenge of estimating the (co)variance parameter(s) that describes this accuracy. All of the estimators rely on the data generated by throwing $n$ darts, aiming each time at the centre of the board, i.e., the double-bulls-eye, and recording the result for each throw.

The authors noted that they would lose considerable information by not measuring the actual *locations* where the darts land but considered this to be too time-consuming and error-prone. Instead, they chose the individual *scores* produced by the throws (the 44 possible scores are 0:22, 24:28, 30, 32:34, 36, 38:40, 42, 45, 48, 50, 51, 54, 57, 60). Based on $n = 100$ throws by authors 1 and 2, assuming the simplest variance model (equal, uncorrelated vertical and horizontal Gaussian errors), their standard deviations were estimated to be $\hat{\sigma} = 64.6$ and 26.9 respectively (the applet gives $\hat{\sigma}$ to 2 decimal places)

Our follow-up letter provides a measure of the statistical precision of these accuracy estimates (for example, we calculate that the 95% limits to accompany the reported point estimate 64.6 derived from 100 scores are approximately 56 and 75). More importantly, we show that more precise estimates of $\sigma$ can often be achieved with the same number of throws (or the same precision with fewer throws) if one uses a simpler yet more informative version of the result from each throw.

Here, as in the letter, we focus on the simplest variance model, where horizontal and vertical errors, $e_x$ and $e_y$, are Gaussian, centered on (0,0), independent of each other and of the same amplitude, i.e., $\sigma_{e_x} = \sigma_{e_y} = \sigma; \rho_{e_x,e_y} = 0$.

We first consider the most mathematically tractable, but least practical, method of estimating $\sigma$, namely to measure the exact $(x, y)$ locations where the $n$ darts land. We then consider the almost as mathematically tractable, but much more practical – and almost as statistically efficient – method of estimating $\sigma$, namely to merely record in which 'ring' each dart lands. We leave to later the the authors' more complex – but sometimes less efficient –

[18] Tibshirani, R.J., Price, A, and Taylor, J. A statistician plays darts. J. R. Statist. Soc. A (2011) 174, Part 1, 213-226. [See also the follow-up letter from S. Sadhukhan, Z Liu, and J Hanley, along with the references • Klein, J.L. (1997). Statistical Visions in Time: A History of Time Series Analysis 1662- 1938. pp. 3-11. Cambridge. Cambridge University Press. • Liddell, H.G. and Scott R. (1920). A Lexicon, abridged from Liddell and Scott's Greek-English Lexicon. p. 653. London. Oxford at the Clarendon Press. • Tibshirani, R.J., Price, A, and Taylor, J. A statistician plays darts. J. R. Statist. Soc. A (2011) 174, Part 1, 213-226. • Turner, E.L. and Hanley, J.A. (2010) Cultural imagery and statistical models of the force of mortality: Addison, Gompertz and Pearson. J. R. Statist. Soc. A, 173, Part 3, 483-499.

method based on actual 0-60 scoring system used in darts games.

Denote by $e_{c,i}$ the error in the $c$-th co-ordinate (1='$x$', 2='$y$') of the $i$-th dart.

**Supplementary Exercise 3.12**

1. Show that $(1/2n)\sum_c\{\sum_i e_{c,i}^2\}$ is an unbiased estimator of $\sigma^2$ and that it is the method-of-moments, the LS, and the ML estimator.

   What sampling statistical distribution does this estimator follow?

   Use the two separate $\alpha/2$ tails of this (slightly non-symmetric) distribution to derive an asymmetric first-principles frequentist confidence interval for $\sigma^2$.[19]

   Suppose that for each dart thrown, one calculates the squared distance from the center, ie $d_i^2 = e_{1,i}^2 + e_{2,i}^2$. Show that $(1/n)\sum_i d_i^2$ is an unbiased estimator of $2\sigma^2$. What sampling statistical distribution does each $d_i^2$ follow? What is a common name for the distribution of the square root of this random variable?

2. Suppose we simply divide the dartboard into 7 '*rings*'[20] and record which one the dart lands in: 1. the double-bulls-eye; 2. the single-bulls-eye; the ones formed by the: 3. single-bulls-eye and inner triple; 4. inner and outer triple; 5. outer triple and inner double; and 6. inner and outer double, wires respectively; and 7. beyond the outer double wire (i.e., the throw misses the board). In other words, we divide the dartboard into just 7 regions. Suppose that the distribution of the results of $n = 100$ throws is as follows:

| ring: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | all |
|---|---|---|---|---|---|---|---|---|
| frequency: | 0 | 6 | 77 | 5 | 12 | 0 | 0 | 100 |

   Calculate (and plot) the logLik($\sigma^2$) function and find the MLE of $\sigma^2$.

[19] Hint: (taking some semantic liberties) a first-principles 100(1-$\alpha$)% frequentist CI, $(L, U)$ for $\theta$ is the pair of *statistics* $(L, U)$, such that Prob($\hat{\theta} \geq \hat{\theta}_{\text{observed}} \mid \theta = L) = \alpha/2$ and Prob($\hat{\theta} \leq \hat{\theta}_{\text{observed}} \mid \theta = U) = \alpha/2$.

[20] In fact, the innermost region is a circle, the next 5 are rings, and the outermost one is all of the remaining area.

## 3.7 An application of the EM algorithm: rounding (and thus 'heaping' of the frequencies) of values

Just like having to code one's own Newton-Raphson algorithm is being replaced by the availability of functions such as `optim`, and as the use of MCMC methods is growing, it is also the case that the EM algorithm – which used to be seem as a derivative-free approach to parameter estimation in a large class of problems – may be becoming less common.

But it is still the simplest way for a large number of situations, not all of which are immediately of the 'missing' or 'incomplete' data type. The original 1977 paper by Dempster, Laird and Rubin has examples in contexts where its applicability not have been immediately obvious.

As motivation for why it is an important item in the biostatistician's repertoire ('toolbox'), consider the following example of 'heaped' frequencies resulting from the rounding of values.

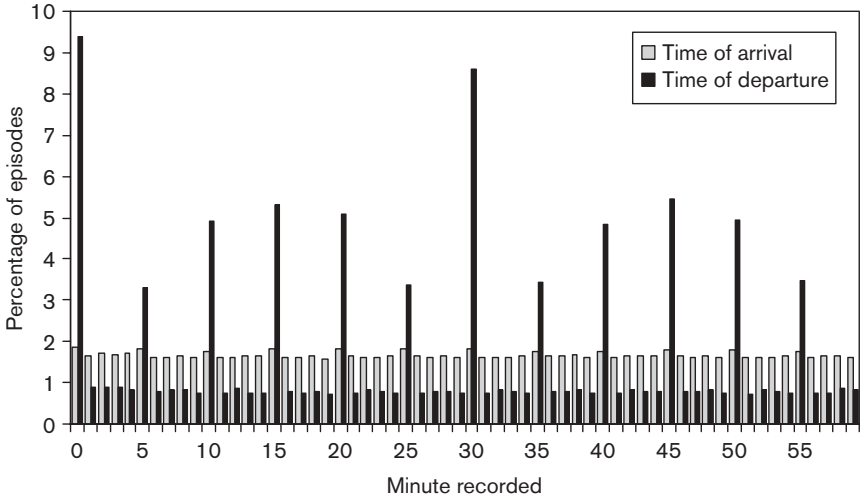
> ***Digit preference bias in the recording of emergency department times***
>
> **Objective** Digit preference bias has previously been described in a number of different clinical settings. The paper aimed to assess whether digit preference bias affects the recording of the time patients arrive and leave emergency departments.
>
> **Method** An observational study of 137 emergency departments in England and Wales was conducted. Each department was asked to submit details of the time of arrival and time of departure from the emergency department for each patient attending during April 2004. In addition, interviews with the lead clinician were undertaken to determine the method used to record the time of departure. The degree of digit preference bias was assessed using a modification of Whipple's index.
>
> **Results** One hundred and twenty-three (86.9%) departments submitted data detailing 648 203 emergency department episodes. 114875 (18.0%) episodes had a recorded minute of departure of '0' or '30', with a further 2 81 890 (44.1%) having other values with a terminal digit of '0' or '5'. The mean modified Whipple's index for time of departure was 316.9 (range 70.9484.4). Linear regression demonstrates a small but significant inverse relationship between the modified Whipple's index and the mean total time in department (b = -0.05, 95% CIs -0.09 to -0.0004, P = 0.048).
>
> **Conclusion** Some departments show considerable digit preference

bias in the recording of time of departure from the emergency department. Such bias may cause difficulty in assessing changes in the performance of departments.

Thomas E. Locker and Suzanne M. Mason. European Journal of Emergency Medicine 13:99101

### Supplementary Exercise 3.13

Refer to the graph below, and to the (.csv) file containing the 60 minute-specific arrival and departure frequencies, which JH has extracted from it.




Distribution of the recorded minute of arrival and departure.

1. The authors reported that (from their Graph 1 – above) "it can be seen that there is little digit preference bias in the recording of time of arrival with a mean MWI of 108.6 (range 66.1279.4)".

   The digit preference seems to be limited to the rounding of a certain proportion (we will call it $\pi_5$) of all arrival times to the nearest 5 and recording the remaining proportion $(1-\pi_5)$ using the exact arrival minute.

   For now, take this to be a reasonable working model, and estimate the model parameters, i.e., the probabilities $\pi_1(=1-\pi_5)$ and $\pi_5$, and the 'mean number of arrivals per bin,' $\mu$.[21] Do so (i) using grade 6 arithmetic, and (ii) by setting it up in, and fitting, a linear model.

2. The authors reported that "Some departments show considerable digit preference bias in the recording of time of departure from the emergency department.'

   Suggest a reasonable working model, and outline how you mighty estimate the model parameters by setting up, and fitting, a linear model.

### Supplementary Exercise 3.14

Refer to the article 'Age at menarche in Warsaw girls in 1965', by Halina Milicer and Franciszek Szczotka in Human Biology, Vol. 38, No. 3 (September, 1966), pp. 199-203, and to the frequencies and interval midpoints in the `menarche` dataset in the `MASS` `R` package.

1. Try to replicate the fitted constants and $X^2$ test statistic in the article, and to understand why they are different from the constants and goodness of fit statistics produced by the fit in the `R` example in the `menarche` documentation.

2. Fit the probit model by ML, but from scratch – using `optim` with your own log-likelihood function.

3. Instead of fitting the Gaussian (Normal) model, fit the Gamma model and the log-Normal model.

4. Is there a way to fit the $N(\mu, \sigma)$ parameters using the *boundaries* (rather than the centres) of the intervals? i.e., recognizing that the ages within each bin are spread out, and not all at the mid-point?

### Supplementary Exercise 3.15

This exercise is motivated by the article 'Trends In The Black-White Life Expectancy Gap Among US States, 1990-2009' by Harper S, MacLehose RF, and Kaufman JS, in the journal Health Affairs 33, No. 8 (2014): 1375-1382. The authors fit age-, sex-, year- and race-specific mortality rates from 'data obtained from the (U.S.) National Vital Statistics System.[22]

[21] You might think of $\mu$ as superfluous and irrelevant, but in the more complex case we are leading up to (numbers of car crashes in each 1-minute bin from 8pm to 10pm, before and after cell phone calls switch over to being 'free' after 9pm) $\mu$ will not be constant, but will decline over the 2 hour period.

[22] 'The system, which is maintained by the National Center for Health Statistics (NCHS), collects information on all deaths occurring in the United States each year. For the fifty states and the District of Columbia.' They calculated state-specific all-cause mortality rates for infants, children ages 1-4, and the remaining five-year age categories (up to ages 85 and older) from 1990 to 2009 for black and white males and females. They extracted the data using software from the National Cancer Institute?s Surveillance, Epidemiology, and End Results Program,18 for which the NCHS provides the underlying mortality data

'The NCHS reports zero counts for deaths when, for a given group (for example, black males ages 15-19 in Wyoming in 2001), no deaths have been reported. **However, for reasons of confidentiality, the NCHS suppresses the number of deaths if the count is 1-9**. For these categories (12.9 percent of all 69,920 state-age- sex-race observations),' the authors 'imputed the unobserved count using a truncated[23] Poisson regression to accommodate the fact that the true number of deaths was censored. [24]

For these two reasons (the sometimes-interval-censored numerators, and the need to smooth adjacent rates) they 'used Markov chain Monte Carlo methods to obtain state-specific annual mortality rates for each race by sex and age.'

For this exercise, we will deal only with the presence of interval-censored counts. We wish to estimate two mortality rates, one for females, one for males. The data come from a 1-year age interval (while people were in their 58th year of life) in the year 2012 in a European country. But they come in 20 segments, since the death certificates were processed in batches. The count of deaths (D) in each segment is accompanied by the numbers of person years (PY) in the sub-population. If the numerator was 1, 2, 3, or 4 it was 'suppressed' ('coarsened' would be a better term) and is shown below as -124.

1. Use the Poisson model to estimate the log of the death rate among males, i.e., the parameter of interest is $\theta_M = \log(\lambda_M)$.
2. Repeat for the parameter $\theta_F = \log(\lambda_F)$ for females.
3. Use these 2 estimates to estimate the rate ratio $\theta_M/\theta_F$.

| Female | | Male | |
|---|---|---|---|
| D | PY | D | PY |
| 8 | 1490 | 35 | 2865 |
| 7 | 2030 | -124 | 101 |
| 5 | 679 | -124 | 272 |
| 5 | 2087 | 20 | 2185 |
| -124 | 591 | -124 | 264 |
| 19 | 4001 | 37 | 4559 |
| -124 | 653 | 5 | 947 |
| 9 | 1884 | 41 | 5123 |
| 16 | 3293 | 14 | 2069 |
| 0 | 74 | 13 | 1439 |
| 8 | 2222 | 12 | 747 |
| -124 | 324 | -124 | 399 |
| -124 | 258 | 15 | 1553 |
| -124 | 606 | 26 | 2826 |
| -124 | 1254 | 18 | 1875 |
| 18 | 2373 | 20 | 2157 |
| 12 | 1830 | 11 | 1074 |
| -124 | 583 | 8 | 1338 |
| 10 | 2591 | 10 | 675 |
| 14 | 2906 | -124 | 960 |

D: Number of Deaths; PY: Person-Years.

## 3.8 Bayesian approach to parameter estimation

Given that the Bayesian approach is a very important and conceptually different way of making inference about the parameters of a model, and even though they mentioned Bayes rule in Chapter 2, it is surprising that Clayton and Hills do not make a statement about the Bayesian approach until Chapter 10; and even then, they do not give it much space. Maybe it's because they wanted the reader to become quite comfortable with Likelihood (which provides the Bridge between the prior and posterior distributions) before doing so.

and population es-timates.'

They 'used data for nineteen age categories, forty-six state entities, two races, two sexes, and twenty years. Thus, there were 69,920 crude mortality rates that could be calculated from the data. However, given the small black populations in many states, these crude estimates suffer from a lack of precision.

They obtained denominator counts from the Census Bureau.' They modeled the number of deaths in each state-age-sex-race category in each year using Poisson regression, with the size of each group as an offset term in the regression.

[23] Technically, it should be called interval-censored, and the term 'truncated' avoided. A distribution is truncated if some of the distribution is missing. Here it would happen if a segment where the count was zero was not even reported. A good example from biology is the distribution of the sizes of fish: if the net used to catch them has too large a mesh, the small fish will be missed, and the observed distribution will be distorted version of the truth; likewise if the distribution of how many times people visit their doctor over a year was estimated by only looking at the files of those who visited at least once. The distribution of heights of 5 year old children visiting Disney World would be distorted if it was estimated from only those short enough to pass under a bar used to limit certain rides to smaller children, or at the other end if they had to be a certain minimum height to go on a ride.

The difference between censoring and truncation is sometimes described by saying that in a *censoring* situation, one *records* the fact that an observation is censored, whereas in a *truncation*, these observations are *not even recorded – they are missed.*

[24] Furthermore, 'because mortality rates were less stable in categories that had a small number of deaths (for example, in places with relatively small populations or where death rates were very low),' they 'opted to smooth all state-age-sex-race death rates across years using a conditional auto-regressive prior specification. This specification allowed the death rates for each state-age-sex-race group to be smoothed toward the rate in adjacent years. For example, the death rate for black men ages 35?39 in New Hampshire in 2003 was smoothed using data for those in 2002 and 2004 to provide a more stable estimate.