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 π 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:

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

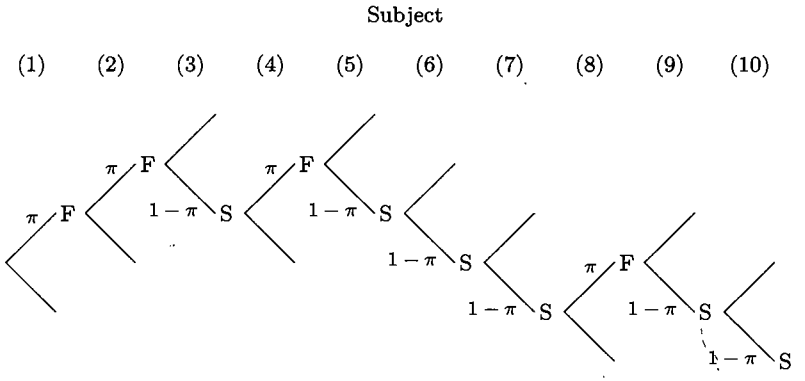

Fig. 3.1. Study outcomes for 10 subjects.

This expression can be used to calculate the probability of the observed study result for any specified value of π. 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 π it is called a *likelihood*. More generally, if we observed D failures in N subjects, the likelihood for π 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,

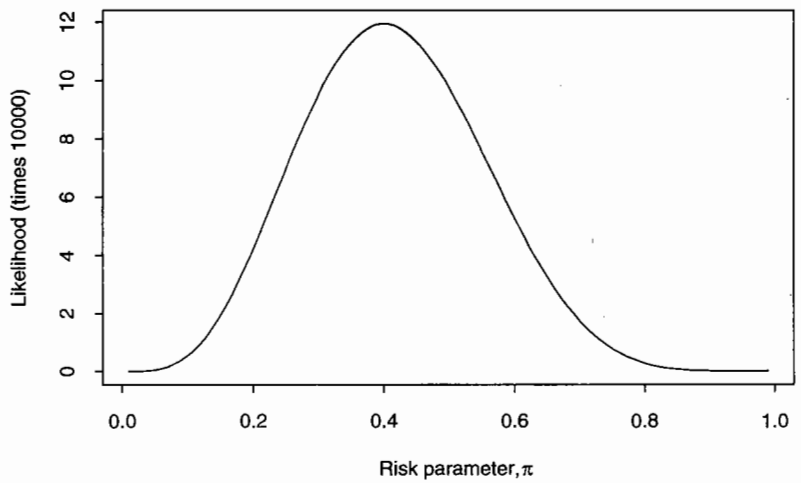

Fig. 3.2. The likelihood for π.

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 π.

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

3.2 The supported range for π

The most likely value for π is 0.4, with likelihood 11.9×10^{-4}. The likelihood for any other value of π 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 π. This provides a more convenient measure of the degree of support than the likelihood itself. It can be used to classify values of π as either supported or not according to some critical value of the likelihood ratio. Values of π 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 π is the set of values of π 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 π is shown in Figure 3.3. We have used the number 0.258

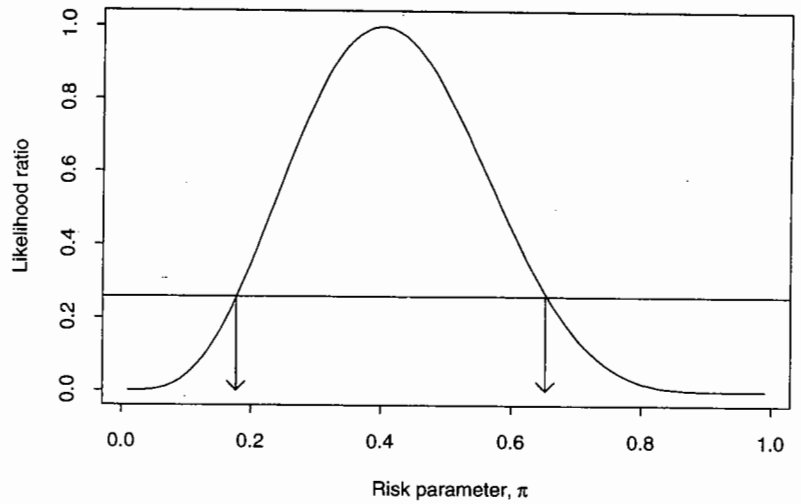

Fig. 3.3. The likelihood ratio for π.

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 π 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 π. 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 π 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 π 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.

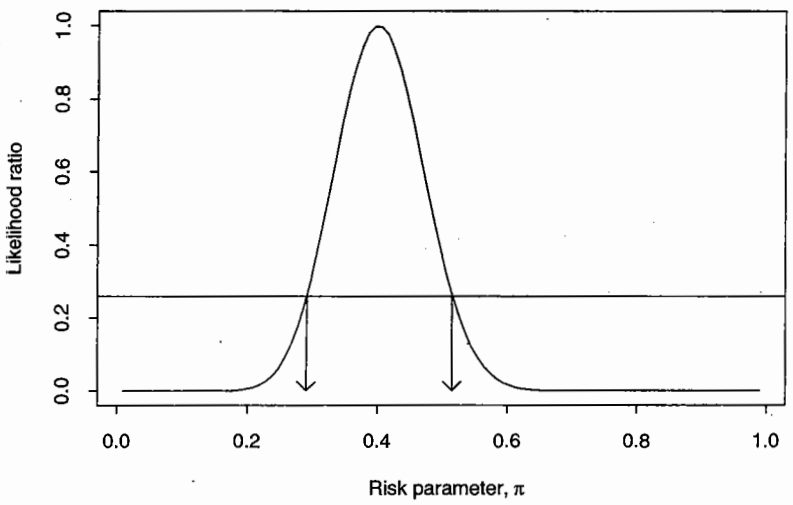

Fig. 3.4. The likelihood ratio based on 20 failures in 50 subjects.

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 π, 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.

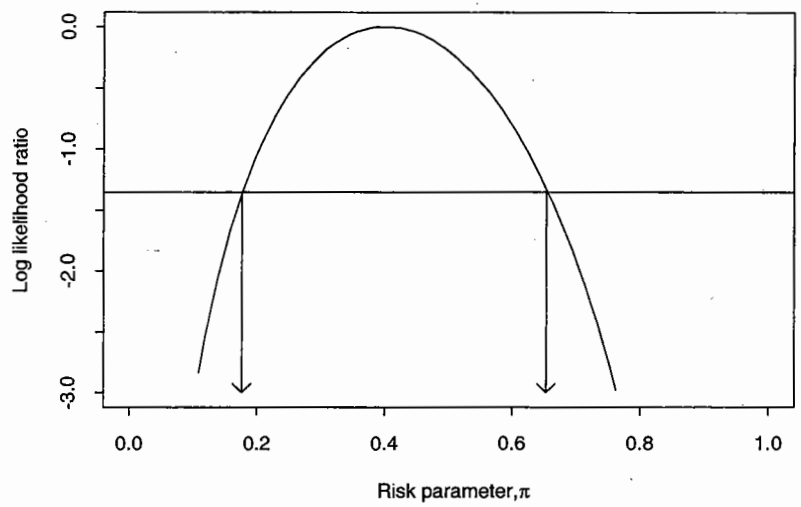

Fig. 3.5. The log likelihood ratio for π.

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 π 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 π can be found from this graph in the same way as from the likelihood ratio graph, by finding those values of π 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 π in the supported range?

In general, the log likelihood for π, 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, Ω, by substituting $\Omega/(1+\Omega)$ for π 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 π or Ω.

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 θ 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 θ.

Exercise 3.4. Write down the expression for the log likelihood as a function of the unknown risk ratio, θ, 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 θ 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 π=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

"*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 they 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. 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 μ, or about the shape of the distribution of the possible values on both sides of μ. 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.

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 μ 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 σ, 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 vlaues of π. 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 π is constant over trials, the actual sequence does not contain any information about π, and we would not include this multiplier in the Likelihood. In any case, we think of the likelihood as a function of π rather than of the observed numbers of each of the two types.: these data are considered fixed, and π 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 π. The likelihood function is based just on the data, and in real life any extra prior information about π 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 π" is

simpler and less ambiguous.

History: there is some dispute as to who first used the principle of ML for the choice of parameter value. The name of Gauss is often mentioned. The *seldom mentioned* 1912 paper 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 μ and σ of a Normal distribution: the ML estimate of σ^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. His 1912 paper can be found in the digital archives in Adelaide, Australia (he spent his last years there) but JH has put a copy in the Resources folder.

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.

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/felxibility 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 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 β.

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 3.1. Grouped Normal data (from Fisher's paper[1]).Three hundred observed measurement errors (ϵ'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:-

[1] 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

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 σ^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 μ : 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 χ^2.
4. Using the method of Maximum Likelihood.

Supplementary Exercise 3.2. Frequency data, the subject of Galton's 1894 correspondence with the Homing News and Pigeon Fanciers' Journal.[2]

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[3] 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

[2]Material (3p of journal, Fanshawe's article, and R code) avialable under Resources.
[3]5-6 means 500-600 yards per minute, etc.

estimate μ and σ, and, where possible, using SE($\hat{\mu}$) and SE($\hat{\sigma}$), [4] form symmetric (frequentist) confidence intervals for μ and σ,

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 χ^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.[5] Instead he – as Gosset often did – used the Probable Error (PE), i.e., 1/2 the IQR.[6]

   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 σ 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 σ^2, and by transformation, for σ.[7]

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

[5]Karl Pearson was the one who promoted the SD.
[6]Thus, it is equally probable (50:50) for an observation to be more/less than this amount from the middle (truth).
[7]Hint: (taking some semantic liberties) a first-principles 100(1-α)% frequentist CI, (L, U) for θ is the pair of *statistics* (L, U), such that $\text{Prob}(\hat{\theta} \geq \hat{\theta}_{\text{observed}} \mid \theta = L) = \alpha/2$ and $\text{Prob}(\hat{\theta} \leq \hat{\theta}_{\text{observed}} \mid \theta = U) = \alpha/2$.

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 β 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 β in the model represents the mean weight per page of the document.

We gave this 'estimation of β' 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 3.3

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 both the LS and (if possible the) ML estimators of β 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 σ is known]
   (d) $y \mid x \sim N(\mu_{y|x}, \sigma^2 = x \times \sigma_0^2)$ [assume σ_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 β 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*'.

Supplementary Exercise 3.4

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 3.5

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: Distribution of Observations in a Dilution Series.

(Again, text from 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.[8] [*note from JH: for simplicity, we replace some of Fisher's notation, by letting the expected or average concentration of micro-organisms be μ 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 μ from a record of the sterile and fertile plates. We can 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

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

when the summation is over the different dilutions.

Supplementary Exercise 3.6 Estimate μ 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

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

3.6.5 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 form 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 3.7 Suppose that in order to estimate the prevalence (π)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 π using the method of moments, and using minimum χ^2 and Maximum Likelihood criteria.

3.6.6 Application: Measuring one's accuracy at darts

In 2011, Tibshirani (junior!) et al.[9] 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.

[9]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.

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 σ 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 σ, 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 σ, namely to merely record in which 'ring' each dart lands. We leave to later the the authors' more complex – but sometimes less efficient – 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 3.8

1. Show that $(1/2n)\sum_c\{\sum_i e_{c,i}^2\}$ is an unbiased estimator of σ^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 σ^2.[10]

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

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*' [11] 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(σ^2) function and find the MLE of σ^2.

3.7 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.

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