Statistical Inference: An Integrated Bayesian/Likelihood Approach

Murray Aitkin

July 3, 2009
Chapter 1

Theories of Statistical Inference

1.1 Example

We begin with a simple example. We draw a simple random sample of size 40 from a large but finite population of families and for each family record the family income for the previous tax year. From this sample we wish to draw an inference about the population mean family income for that tax year. How is this to be done? The sample of incomes, rounded down to the nearest thousand dollars, is given below.

<table>
<thead>
<tr>
<th>Family income, in units of 1000 dollars</th>
</tr>
</thead>
<tbody>
<tr>
<td>26 35 38 39 42 46 47 47 52 53 55 55 56</td>
</tr>
<tr>
<td>58 60 60 60 60 65 65 67 67 70 71 72</td>
</tr>
<tr>
<td>75 77 80 81 85 93 96 104 104 107 119 120</td>
</tr>
</tbody>
</table>

Theories of inference can be divided into two classes: those which use the likelihood function (defined below) as an important, or the sole, basis for the theory, and those which do not give the likelihood any special status.

Within the first class, there is a division between theories which regard the likelihood as the sole function of the data which provides evidence about the model parameters, and those which interpret the likelihood through other factors.

Within the second class, there is a division between theories which take some account of a statistical model for the data, and those which are based exclusively on the properties of estimates of the parameters of interest in repeated sampling of the population. Comprehensive discussions of the main theories can be found in Welsh (1996) and Lindsey (1996) to which we refer frequently. The discussion
here is limited to the generality and simplicity of the theories. We illustrate these theories with reference to the population income problem above.

1.2 Statistical models

Theories which use the likelihood require a statistical model for the population from which the sample is taken, or more generally for the process which generates the data. Inspection of the sample income values shows that (in terms of the measurement unit of $1000) they are integers, as are the other unsampled values in the population. So the population of size $N$ can be expressed in terms of the population counts $N_J$ at the possible distinct integer values of income $Y_J$, or equivalently by the population proportions $p_J = N_J/N$ at these values.

A (simplifying) statistical model is an approximate representation of the proportions $p_J$ by a smooth probability distribution depending on a small number of model parameters. The form of the probability function is chosen (in this case of a large number of distinct values of $Y$) by matching the cumulative distribution function (cdf) of the probability distribution to the empirical cumulative distribution of the observed values. A detailed discussion of this process is given in Aitkin, Francis and Hinde (2005) and Aitkin, Darnell, Francis and Hinde (2008). We do not give details here, but the matching process leads to the choice of an approximating continuous cdf model $F(y | \lambda)$, and corresponding density function $f(y | \lambda) = F'(y | \lambda)$; the probability $p_J$ of $Y_J$ is approximated by $F(Y_J + \delta/2 | \lambda) - F(Y_J - \delta/2 | \lambda)$ where $\delta$ is the measurement precision ($= 1$ in the units of measurement). When the variable $Y$ is inherently discrete on a small number of values, as with count data, the values $p_J$ are approximated directly by a discrete probability distribution model.

The income sample above is clearly skewed with a longer right-hand tail of large values, so an approximating model with right skew would be appropriate. The gamma, lognormal and Weibull distributions are possible choices.

1.3 The Likelihood Function

Given a simple random sample $y = y_1, \ldots, y_n$ of size $n$ drawn from the population (assumed for the moment to be large compared to the sample), and an approximating statistical model $F(y | \lambda)$, the likelihood function $L(\lambda | y)$ (of the model parameters $\lambda$) is the probability of the observed data as a function of
1.3. THE LIKELIHOOD FUNCTION

if the measurement precision is high relative to the variability in the data.

In general the parameter vector \( \lambda \) can be partitioned into a subvector \( \theta \), of parameters of interest, and a subvector \( \phi \) of nuisance parameters. We want to draw conclusions about the parameters of interest \( \theta \), but the model depends as well on the nuisance parameters \( \phi \).

For example, we use the gamma distribution with parameters \( \mu \) (the mean) and \( r \) (the shape parameter) as the specified model, where \( \mu \) is the parameter of interest and \( r \) is the nuisance parameter:

\[
f(y | \mu, r) = \frac{r^r \Gamma(r)}{\Gamma(r) \mu^r} \exp(-ry/\mu) y^{r-1}.
\]

This parametrization is convenient because it gives orthogonality in \( \mu \) and \( r \) in the information matrix. For the more common parametrization giving the gamma density

\[
f^*(y | \mu, r) = \frac{1}{\Gamma(r) \mu^r} \exp(-y/\mu) y^{r-1},
\]

the mean is \( r\mu \), which immediately complicates inference about the mean. We discuss this important issue further in Chapter 2.

Then for high measurement precision \( \delta \), the gamma likelihood function can be written as

\[
L(\mu, r | y) = \prod_{i=1}^{n} \left[ \frac{r^r \exp(-ry_i/\mu) y_i^{r-1}}{\Gamma(r) \mu^r} \cdot \delta^n \right]
\]

where \( T = \sum_i y_i, P = \prod_i y_i \). An important theoretical point is that the likelihood function here depends on the data through only the two data functions \( T \) and \( P \) (and the sample size \( n \)). These sufficient statistics are all that is needed to compute the likelihood function – we do not need the data values themselves.

We will generally drop the \( y \) from the notation for the likelihood function, but it is always implicit in its definition that the data have been observed. We now describe briefly the theories and how they deal with inference about the population mean income.

---

1 The likelihood is frequently defined to be any constant multiple of the probability of the observed data, but in our approach likelihoods are probabilities.
1.4 Theories

1.4.1 Pure likelihood theory

This theory (see for example Edwards 1972 or Lindsey 1996) uses the likelihood exclusively to draw inferences about the model parameters. It is fully conditional on the observed data, so denies the relevance of repeated sampling, but also explicitly rules out prior distributions for model parameters. Since the likelihood is the probability of the observed data, it can be used as a measure of relative plausibility of different parameter values: the relative evidence for different values $\lambda_1$ and $\lambda_2$ of the model parameter $\lambda$ is assessed through their likelihood ratio $L(\lambda_1)/L(\lambda_2)$. A likelihood ratio of 10 means that the data support the value $\lambda_1$ ten times as strongly as the value $\lambda_2$.

So we can compare the data support for pairs of values $(\mu, r)$ like $(60, 3)$ and $(65, 3)$, or $(60, 3)$ and $(65, 5)$. Figure 1.1 shows a form of “contour plot” of the two-parameter likelihood in $\mu$ and $r$, by computing the likelihood over a 100-point equally-spaced grid in $\mu$ and $r$, chosen to cover the region of appreciable likelihood. The size of the plotting symbol in the plot is proportional to the likelihood.

![Gamma likelihood, family income data](image)

Figure 1.1: Gamma likelihood, family income data

(The “banding” in the edges of the graph is due to the finite steps in the size of the plotting symbol.)

In pure likelihood theory this plot represents the data information about both parameters. For two general values of $\mu$ with the same $r$, $(\mu_1, r)$ and
(\mu_2, r) the likelihood ratio is

\[
\frac{L(\mu_1, r)}{L(\mu_2, r)} = \frac{r^{nr} \exp(-rT/\mu_1) P_{nr} \cdot \delta^n/\Gamma^n(r)\mu_1^{nr}}{r^{nr} \exp(-rT/\mu_2) P_{nr} \cdot \delta^n/\Gamma^n(r)\mu_2^{nr}} = \left(\frac{\mu_2}{\mu_1}\right)^{nr} \exp(-rT[1/\mu_1 - 1/\mu_2]).
\]

This depends critically on the nuisance parameter r, as the log of the likelihood ratio increases or decreases in magnitude linearly with r. However the theory does not specify how nuisance parameters are to be treated, and so has very restricted application, to simple cases where either there are no nuisance parameters, or the likelihood function is separable in the parameters of interest, that is, it can be expressed as

\[L(\theta, \phi) = L_1(\theta) \cdot L_2(\phi),\]

where \(L_1\) and \(L_2\) are functions respectively of \(\theta\) only and \(\phi\) only. If this is the case, it follows immediately that

\[
\frac{L(\theta_1, \phi)}{L(\theta_2, \phi)} = \frac{L_1(\theta_1)}{L_1(\theta_2)},
\]

and the nuisance parameter can be ignored: every section through the likelihood parallel to the \(\theta\) axis gives the same likelihood ratio for \(\theta_1\) to \(\theta_2\), whatever the value of \(\phi\).

In other cases (the vast majority) some treatment of the nuisance parameters is necessary, and different ad hoc treatments may lead to different conclusions. Lindsey (1996) gave a very detailed discussion of this theory. A common approach is to eliminate the nuisance parameter by profiling: replacing the unknown \(\phi\) by \(\hat{\phi}(\theta)\), its MLE given \(\theta\). In the case of separable parameters, this replaces \(\phi\) by its MLE \(\hat{\phi}\), giving the correct likelihood ratio, but in general \(\hat{\phi}(\theta)\) defines a curved path through the parameter space, and the resulting profile likelihood is over-precise.

### 1.4.2 Bayes theory

Bayes theory was dominant (indeed, the only theory) from the 1800s to the 1920s. It is fully conditional on the observed data. This is the data set we have, and conclusions about the population from which it was drawn are based on the likelihood function \(L(\lambda)\) (representing the data information) and the prior (probability) distribution \(\pi(\lambda)\) of the model parameters, representing the information we have about these parameters external to, and in advance of, the sample data. Inference is expressed through the posterior distribution \(\pi(\lambda | y)\) of the model parameters, updated from the prior by the likelihood through Bayes’s (or Bayes’) theorem:

\[
\pi(\lambda | y) = \frac{L(\lambda)\pi(\lambda)}{\int L(\lambda)\pi(\lambda) d\lambda}.
\]
If $\lambda$ can take just one of the two values $\lambda_1$ and $\lambda_2$, with prior probabilities $\pi_1$ and $\pi_2$, the ratio of posterior probabilities (the posterior odds for $\lambda_1$ to $\lambda_2$) is

$$
\frac{\pi(\lambda_1 \mid y)}{\pi(\lambda_2 \mid y)} = \frac{L(\lambda_1)\pi_1}{L(\lambda_2)\pi_2} = \frac{L(\lambda_1)}{L(\lambda_2)} \cdot \frac{\pi_1}{\pi_2},
$$

so that the posterior odds is equal to the likelihood ratio multiplied by the prior odds.

So the likelihood ratio plays the same role in Bayes theory that it does in pure likelihood theory – to provide the data evidence for one parameter value over another – but this is complemented in Bayes theory by the prior information about these values – their prior probabilities.

The theory requires that we express prior information as a probability distribution. In many cases we may not have well-developed information or views which are easily expressed as a probability distribution, and much use is made, by many Bayesians, of weak or non-informative priors, which are “uninformative” relative to the information in the data: the data were presumably collected to obtain information about parameters for which we had little prior information, and so the prior should reflect this lack of information. A non-informative prior for the case above of two parameter values would be one with equal prior probabilities, leading to the posterior odds being equal to the likelihood ratio.

A subdivision of Bayes theorists regards non-informative priors as at best undesirable (especially when they are improper), and at worst denying the whole point and advantage of the Bayesian approach, which is to accommodate both sample data and external information in the same unified probabilistic framework. It argues that all prior distributions should reflect the actual information available to the analyst; this may mean that different analysts using different prior distributions come to different conclusions. Analysts who have difficulty formulating priors need to be trained in prior elicitation (Garthwaite et al 2005).

Arguments about the prior, and the meaning, existence and uniqueness of “noninformative” priors, will be discussed more fully in Chapter 2. As stated above, we want in general to draw conclusions about the parameters of interest $\theta$, but the model depends as well on the nuisance parameters $\phi$. This is achieved in Bayes theory by a standard probability procedure: we integrate the joint posterior distribution $\pi(\theta, \phi \mid y)$ over $\phi$ to give the marginal posterior distribution $\pi(\theta \mid y)$:

$$
\pi(\theta \mid y) = \int \pi(\theta, \phi \mid y) d\phi.
$$

For the income example, we need to specify the prior distribution for $(\mu, r)$. This choice is discussed further in Chapter 2. A simple approach is to specify independent flat priors for $\mu$ and $r$. The posterior density $\pi(\mu, r \mid y)$ is then proportional to the likelihood $L(\mu, r)$.

Figure 1.1, which showed the likelihood, can also be interpreted, for the independent flat priors, as the joint posterior density defined over the $100 \times 100$
grid of $\mu_k$ and $r_\ell$ for $k, \ell = 1, \ldots, 100$. Strictly speaking it is a posterior mass function $\pi(\mu_k, r_\ell \mid y)$ rather than a density function. The marginal posterior mass function of $\mu$ is then simply obtained by summing over $r_\ell$:

$$
\pi(\mu_k \mid y) = \sum_{\ell=1}^{100} \pi(\mu_k, r_\ell \mid y).
$$

Figure 1.2 shows this mass function, and Figure 1.3 shows the cumulative mass function; these are drawn as continuous curves for visual appeal.

The mean of the posterior mass function $\sum_k \pi(\mu_k \mid y) \mu_k$ is 67.44. Approximate credible intervals and percentiles for the posterior distribution of $\mu$ can be obtained by linear interpolation from the cumulative mass function: at $\mu = 60.5$ the cdf is 0.0221 and at 60.75 it is 0.0265, so the approximate 2.5 percentile is 60.66; at $\mu = 74.5$ the cdf is 0.9725 and at 74.75 it is 0.9761, so the approximate 97.5 percentile is 74.67; the approximate 95% credible interval for $\mu$ is then [60.66, 74.67].

This simple approach suffers from two difficulties. First, it is limited to small numbers of parameters: with three parameters 100-point grids would require $10^6$ points, and the three-dimensional likelihood is difficult to visualise. Second, the marginal distributions of the individual parameters are computed over only 100 points, which does not give sufficient precision for accurate percentiles. We adopt a different approach, by taking advantage of the form of the posterior resulting from the exponential family to which the gamma distribution belongs.

We now specify independent flat priors for log $\mu$ and $r$; we note below the effect of this change on the posterior distribution for $\mu$. These priors are improper (as is the flat prior on $\mu$) – they do not integrate to 1 over the infinite ranges for $r$ and $\mu$ – but this does not cause any difficulty if the posterior is proper.

The role of $r$ is clear if we re-express the likelihood in $\mu$ and $r$: writing $\theta = 1/\mu$,

$$
L(\mu, r \mid y) = \frac{e^{nr}}{\Gamma(nr)} \exp(-rT/\mu) P^{r-1} \cdot \delta^n
$$

$$
L(\theta, r \mid y) = \frac{(rT)^{nr}}{\Gamma(nr)} e^{-rT^2 \theta^{nr}} \cdot \frac{P^{r-1} \Gamma(nr)}{T^{nr} \Gamma(nr)} \cdot \delta^n.
$$

With the prior $\pi(\mu, r) = d\mu dr/\mu$ which is equivalent to $\pi(\theta, r) = d\theta dr/\theta$, the joint posterior distribution of $\theta$ and $r$ can be factored into the conditional posterior distribution of $\theta$ given $r$, and the marginal posterior distribution of $r$.

The conditional density of $\theta$ given $r$ is

$$
\pi(\theta \mid r, y) = \frac{(rT)^{nr}}{\Gamma(nr)} e^{-rT^2 \theta^{nr-1}},
$$

a gamma distribution with parameters $rT$ and $nr$. (It is easily seen that the effect of changing the prior on $\mu$ to uniform is to change the prior on $\theta$ to $d\theta/\theta^2$;
Figure 1.2: Posterior mass function, mean income

Figure 1.3: Posterior cumulative distribution, mean income (1)
1.4. THEORIES

this changes the conditional gamma distribution of \( \mu \) given \( r \) to gamma(\( rT, nr - 1 \)). For \( n = 40 \) and \( r > 2 \) this makes a negligible difference to the posterior.)

The marginal density of \( r \) is not of any standard form:

\[
\pi(r \mid y) = c \cdot \frac{P^{r-1} \Gamma(nr)}{T^{nr} \Gamma(n)(r)},
\]

where \( c \) is the integrating constant of the density. Analytic integration over \( r \) is therefore not possible. However simulation from the marginal density of \( \mu \) is straightforward, and is discussed at length in Chapter 2. We first compute the posterior mass function of \( r \) over a fine grid, draw \( M \) random values \( r^{[m]} \) of \( r \) from this mass function, and for each \( r^{[m]} \) draw a random \( \theta^{[m]} \) from the conditional gamma density of \( \theta \) given \( r^{[m]} \); then random values \( \mu^{[m]} \) of \( \mu \) are obtained from \( \mu^{[m]} = 1/\theta^{[m]} \).

The advantage of this approach is that a much finer grid for \( r \) can be used, and the draws of \( \mu \) are not on a discrete grid; the disadvantage is that the random draws \( \mu^{[m]} \), when ordered, define the cdf of \( \mu \) only up to the random variation inherent in the \( M \) simulation values. This can be reduced, but not eliminated, by increasing \( M \). We use this approach extensively in this book.

Figure 1.4 shows the cumulative distribution function of \( \mu \) for a sample of 10,000 random values of \( \mu \) generated in this way.

![Figure 1.4: Posterior distribution, mean family income (2)](image)

It is visually identical to that based on the grid computation.

From the ordered values we can immediately construct approximations to the posterior mean of \( \mu \), by simply averaging the 10,000 random values, and
to the 95% central credible interval for \( \mu \), from the 250-th and 9750-th ordered values. This gives the posterior mean as 67.22, and the approximate interval as \([60.54, 74.62]\). The credible interval agrees very closely with the (interpolated) interval \([60.66, 74.67]\) from the marginal posterior with the flat prior on \( \mu \). The posterior means agree less well.

The interval limit values will differ slightly with different random seeds and different random number generators on different computers: their precision depends on the simulation sample size and can be expressed in the usual probabilistic way. We give further details of this approach in Chapter 2.

### 1.4.3 Likelihood-based repeated sampling theory

This theory was dominant from the 1930s to the 1990s. The theory uses the likelihood function to provide both maximum likelihood estimates of parameters and likelihood-based confidence intervals, but these are interpreted through their behaviour in (hypothetical) repeated sampling from the same population (whence the Bayesian term frequentist to describe the theory). Prior distributions are not used for single sample inference.

The interpretation of the likelihood function for the income example is now reversed compared with the Bayesian interpretation: the parameters \( \mu \) and \( r \) are still fixed constants, but do not have prior distributions. It is the sufficient statistics \( T \) and \( P \) which have the probability distributions, over hypothetical repeated samples drawn from the same population. The same likelihood function

\[
L(\mu, r \mid y) = \frac{r^{nr}}{\Gamma(n)r \mu^{nr}} \exp\left(-\frac{rT}{\mu}\right) P^{r-1} \cdot \delta^n
\]

can be re-written differently:

\[
L(\mu, r \mid y) = \frac{r^{nr}}{\Gamma(nr) \mu^{nr}} \exp\left(-\frac{rT}{\mu}\right) \frac{T^{nr-1} \Gamma(nr)}{T^{nr-1} \Gamma(n)}
\]

which is very similar to the Bayesian re-expression, but it now refers to the sampling distributions of \( T \) and \( P \). In repeated sampling, \( T \) has a gamma\((r/\mu, nr)\) distribution, and \( P \) given \( T \) has a distribution determined by the second component.

Without information about \( r \), we cannot use \( T \) to draw inferences about \( \mu \). The standard approach to this difficulty, in the two-parameter exponential family (see for example Cox and Hinkley 1974 or Welsh 1996), is to condition the repeated sampling inference on the value of the sufficient statistic for the nuisance parameter. Since the statistic on which we are conditioning may have a distribution which depends also on the parameter of interest, it is not clear that we retain full information about the parameter of interest by this approach. For example, if \( r \) were the parameter of interest, we would condition on the value of \( T \), to eliminate \( \mu \) from the likelihood. Since the marginal distribution of \( T \) is gamma\((r/\mu, nr)\), the conditional likelihood of \( r \) given \( T \) is simply the second
1.4. THEORIES

The term above:

$$CL(r \mid T) = \frac{P^{r-1}\Gamma(nr)}{T^{nr-1}\Gamma^n(r)}$$

which does not depend on $\mu$, and the log conditional likelihood is

$$\log CL(r \mid T) = (r - 1) \log P + \log \Gamma(nr) - [(nr - 1) \log T + n \log \Gamma(r)].$$

The conditional distribution of $\log P = \sum_i \log y_i$ given $T$ can be obtained, with some effort, from the log-Dirichlet distribution of $\log(y_i/T)$. However since the distribution of $T$ depends on $r$, we may have lost information about $r$ in this process.

But we are interested in $\mu$, not $r$, so we need the conditional distribution of $T$ given $P$. Since the distribution of $P$ has no simple form, we cannot easily determine the necessary conditional distribution, and $r$ has to be eliminated from the likelihood by other means.

One possible approach is through the profile likelihood, in which the unknown $r$ is replaced by its MLE $\hat{r}(\mu)$ given $\mu$, and the resulting profile likelihood in $\mu$ is treated as a single-parameter likelihood. Differentiation of the log-likelihood shows that the MLE of $r$ given $\mu$ is the solution of

$$\log r - \psi(r) = \mu + \bar{y}/\mu + (\log P)/n - 1,$$

where $\psi(\cdot)$ is the digamma function. Aitkin, Francis and Hinde (2005) and Aitkin, Darnell, Francis and Hinde (2008) gave detailed discussions of profile likelihoods in a much broader context. The properties of such profile likelihoods are in general not the same as those of single-parameter likelihoods: the replacement of unknown nuisance parameters by functions of the parameter of interest increases the apparent information about the parameter of interest, and makes inferential statements based on the profile likelihood over-precise.

Even when the conditional likelihood can be determined easily, an obvious question is the relevance of the form of conditional sampling. We have to imagine a sub-family of samples in which the sufficient statistic $P$ is exactly the same as in our observed sample, and imagine then the variation in these samples of the other sufficient statistic $T$. Even in sampling experiments from a known gamma population, it would be very difficult to evaluate the properties of the conditional likelihood obtained in this way, since $P$ has a continuous distribution in repeated sampling, and so need never provide the same value in repeated sampling.

The approach to conditioning on the sufficient statistics for nuisance parameters is not universally accepted amongst all frequentists, even when the conditioning statistic is fully ancillary – does not depend at all on the parameter of interest. When the distribution of the conditioning statistic depends explicitly on the parameter of interest, the case for conditioning is even less persuasive. For example in a 2x2 contingency table from two binomial populations, arguments over the appropriateness of the conditional likelihood, obtained by conditioning on the other margin of the table, continue long after Fisher’s insistence on the correctness and optimality of the conditioning approach.
We do not need to take a position on this issue, which is often ignored in practice:

“The choice of the appropriate set of hypothetical repetitions is in principle fundamental, although in practice much less often a focus of immediate concern.” Cox (2006 p. 198).

However there is a close relation between Bayesian and conditional frequentist procedures:

“Pierce (1973) showed that ... good conditional procedures can only be achieved from a Bayesian analysis based on a proper prior, and that good frequentist properties can only hold for procedures which are in a sense limits of proper Bayesian procedures.” (Welsh 1996 p. 164)

For the 2 × 2 table, Altham (1969) showed that Fisher’s conditional likelihood arises as a posterior distribution for a specific prior.

The important point for our income example is that the frequentist theory does not provide a straightforward theoretical answer to inference about \( \mu \) in the gamma distribution. For distributions outside the exponential family (like the Weibull or finite mixtures) the conditioning approach is not available, and other methods of eliminating nuisance parameters have to be found. Grice and Bain (1980) gave approximate small-sample approaches for the gamma mean.

In large samples reliance can be placed on the asymptotic normality of functions of the observed data, but this is not a general theory.

1.4.4 “Model-guided” survey sampling theory

The term model-guided or model-assisted (as used in Särndal, Swensson and Wretman 1992) is relatively new in survey sampling theory, which was extensively developed in the 1950s. It refers to the usefulness of model-based estimators of model parameters, but without reliance on the correctness of the model. Without a formal model, inference is based on the repeated sampling distribution of the sample selection indicators, not of the population values themselves. The survey design determines the inference, hence the term design-based (as opposed to model-based) inference.

Our income mean example gives a simple example of the approach. We change notation slightly. \( Y \) is the variable of interest, in a finite population of size \( N \). The population values of \( Y \) are \( Y_1, Y_2, ..., Y_N \). The population mean \( \mu \) is

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} Y_i / N
\]

and the population variance is

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 / N.
\]
1.4. THEORIES

(In the survey literature the variance denominator is usually $N - 1$.)

We draw a simple random sample (SRS) without replacement of fixed predetermined size $n$, and obtain observed values $y_1, ..., y_n$, with sample mean $\bar{y}$ and sample variance

$$s^2 = \sum_i (y_i - \bar{y})^2 / (n - 1).$$

Define indicator variables $Z_1, Z_2, ..., Z_I, ..., Z_N$: let

$$Z_I = 1 \text{ if population member } I \text{ is selected}$$
$$= 0 \text{ if population member } I \text{ is not selected}.$$

Then

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$= \frac{1}{n} \sum_{I=1}^{N} Z_I Y_I / \sum_{I=1}^{N} Z_I$$

$$= \frac{1}{n} \sum_{I=1}^{N} Z_I Y_I / n.$$

Inference about $\mu$ is based on the repeated sampling properties of the random variable $\bar{y}$ as an estimator of $\mu$. The fundamental inferential principles are: the $Y_I$ are fixed constants, and the $Z_I$ are Bernoulli random variables, with

$$\Pr[Z_I = 1] = \frac{\text{no. of samples containing unit I}}{\text{no. of samples of size n}} = \frac{(N-1)}{n} = \frac{n}{N} = \pi,$$

the sampling fraction. The properties of $\bar{y}$ are easily established.

$$E[Z_I] = E[Z_I^2] = \pi, \ Var[Z_I] = \pi(1 - \pi) = \frac{n}{N}(1 - \frac{n}{N}).$$

Hence

$$E[\bar{y}] = \frac{1}{n} \sum_{I=1}^{N} E[Z_I] Y_I$$

$$= \frac{1}{N} \sum_{I=1}^{N} Y_I = \mu.$$

So as a random variable, $\bar{y}$ is unbiased for $\mu$. For the variance of $\bar{y}$ we need the joint distribution of pairs of the $Z_I$. These are not independent:

$$\Pr[Z_I = 1, Z_J = 1] = \Pr[Z_I = 1] \Pr[Z_J = 1 \mid Z_I = 1]$$

$$= \frac{n}{N} \cdot \frac{n-1}{N-1}.$$
and so

\[
\text{Cov}[Z_I, Z_J] = \frac{n(n-1)}{N(N-1)} \left( \frac{n}{N} \right)^2
\]

\[
= -\frac{1}{N-1} n \left(1 - \frac{n}{N} \right)
\]

\[
= -\pi(1 - \pi)/(N - 1)
\]

\[
\text{Var}[\bar{y}] = \frac{\sum_i Y_i^2 \text{Var}[Z_i]/n^2 + \sum_i \sum_{\neq J} Y_i Y_J \text{Cov}[Z_I, Z_J]/n^2}{(n-1)}
\]

\[
= \frac{1 - n/N}{nN(N-1)} [(N-1) \sum_i Y_i^2 - \sum_i \sum_{\neq J} Y_i Y_J]
\]

\[
= \frac{1 - n/N}{n(N-1)} \sum_i (Y_i - \mu)^2
\]

\[
= (1 - n/N)\sigma^2/n
\]

\[
= (1 - \pi)\sigma^2/n
\]

if the population variance is defined by the $N - 1$ denominator.

The first term $(1 - n/N) = (1 - \pi)$ is a finite population correction: for a small sample fraction, $\text{Var}[\bar{y}] \approx \sigma^2/n$, but as $\pi \to 1$, $n \to N$, and $\text{Var}[\bar{y}] \to 0$, since the sample exhausts the population.

For the sample variance,

\[
\text{E}[s^2] = \text{E} \left[ \frac{\sum_i y_i^2}{n} - \bar{y}^2 \right] / (n-1)
\]

\[
= \text{E} \left[ \frac{\sum_i Z_i Y_i^2}{n} - \bar{y}^2 \right] / (n-1)
\]

\[
= \left\{ n \sum_i Y_i^2 / N - n(\text{Var}[\bar{y}] + \text{E}[\bar{y}^2]) \right\} / (n-1)
\]

\[
= \left\{ n \sum_i Y_i^2 / N - n([1 - 1/N] \sigma^2/n + \mu^2) \right\} / (n-1)
\]

\[
= (1 - 1/N)\sigma^2.
\]

So $s^2$ is an almost unbiased estimator of $\sigma^2$, regardless of any distribution model for $Y$, and under the Bernoulli model, $\bar{y}$ is the minimum variance linear unbiased estimator of $\mu$.

For confidence interval statements about $\mu$, the theory uses the Central Limit Theorem in its general form. The sample mean has expectation $\mu$ and variance $(1 - 1/N)\sigma^2$ in repeated sampling, and since it is a (weighted) linear combination of (correlated) random variables $Z_I$ as $n \to \infty$ (and $N \to \infty$), the sampling distribution of

\[
z = \frac{\sqrt{n}(\bar{y} - \mu)}{\sigma} \to N(0, 1)
\]
as does that of

\[ t = \frac{\sqrt{n}(\bar{y} - \mu)}{s}, \]

giving the usual large-sample confidence interval

\[ \bar{y} - z_{1-\alpha/2} s / \sqrt{n} < \mu < \bar{y} + z_{1-\alpha/2} s / \sqrt{n}. \]

The accuracy of the confidence interval coverage depends on the sample size \( n \) – it may be quite inaccurate for small \( n \) – and may depend on other properties of the \( Y \) population. Without other information about this population, we cannot say more.

For the example, we have \( \bar{y} = 67.1, s^2 = 500.87 \), and the (approximate) 95% confidence interval for the population mean is [60.1, 74.0]. Remarkably, we seem to be able to make the same inferential statement about \( \mu \) without any model for the population values \( Y \), or invoking the Central Limit Theorem for the sampling distribution of \( \bar{y} \) as a function of \((y_1, \ldots, y_n)\!\). However, when we move to regression models, this approach becomes more difficult. The same approach to simple linear regression of \( Y \) on \( X \) expresses the usual regression estimator

\[ b = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \]

in the indicator variable form:

\[ b = \frac{\sum_I (Y_I - \bar{y})(X_I - \bar{x})}{\sum_I (X_I - \bar{x})^2} \]

with

\[ \bar{y} = \frac{\sum_{I=1}^N Z_I Y_I}{\sum_{I=1}^N Z_I}, \]
\[ \bar{x} = \frac{\sum_{I=1}^N Z_I X_I}{\sum_{I=1}^N Z_I}. \]

Now the repeated sampling distribution of \( b \) is much more complex, because it is a ratio of two linear functions of the \( Z_I \), which does not have a simple asymptotic distribution, though by the Mann-Wald theorem the denominator term can be replaced by its expectation in the limiting distribution.

A more important question is \textit{why} we are using the population version of the linear regression, in the absence of any model for the population. This is the role of the \textit{guiding} model: if the conditional distribution of \( Y \) given \( X \) has a linear regression on \( X \) with constant variance, the least squares estimate would be the minimum variance unbiased estimate, and if the conditional distribution of \( Y \) given \( X \) were in addition normal, the regression estimator would be maximum likelihood and optimal.
So if the model were to hold, the sample survey estimator would be identical to the optimal estimator, but if the model does not hold, the sample survey estimator should still provide a good estimate of the corresponding population quantity

\[ B = \sum_I Z(Y_I - \mu_Y)(X_I - \mu_X)/\sum_I (X_I - \mu_X)^2. \]

From the viewpoint of model-based likelihood theory, this approach is unsatisfactory. The argument is clear if we construct the likelihood as the probability of all observed data. The data are both the sample selection indicators \( Z_I \) and the observed response variables \( y_i \) for the selected population members, so we need a population model for the \( Y_I \) as well.

The fundamental probability relation we use is

\[
\Pr[Y_I, Z_I] = \Pr[Z_I | Y_I] \Pr[Y_I] = \Pr[Y_I | Z_I] \Pr[Z_I],
\]

where \( \Pr[Z_I | Y_I] \) is the sample selection model for \( Z \) – it specifies how the selection probability of population member \( I \) depends on the value of the response \( Y_I \) for that member – and \( \Pr[Y_I] \) is the population model for \( Y \). For simple random sampling,

\[
\Pr[Z_I | Y_I] = \Pr[Z_I] = \pi_Z I (1 - \pi) I - Z_I,
\]

the Bernoulli model above. Correspondingly,

\[
\Pr[Y_I | Z_I] = \Pr[Y_I]
\]

– the model for the selected population values is the same as that for the unselected values. So

\[
\Pr[Y_I, Z_I] = \Pr[Y_I] \Pr[Z_I],
\]

and the likelihood is

\[
L = \Pr[y_1, \ldots, y_n] \cdot \Pr[Z_1, \ldots, Z_N] = \Pr[y_1, \ldots, y_n] \cdot \frac{1}{\binom{N}{n}} = \Pr[y_1, \ldots, y_n] \cdot \frac{n}{N} \cdot \frac{n - 1}{N - 1} \cdots \frac{1}{N - n + 1}.
\]

The last term in the selection probabilities is completely known from the design – it is just a constant, and plays no part in likelihood-based statistical inference.

Inferential statements about the parameters are based on ratios of likelihoods, and these constant terms cancel in the ratios. Thus, regardless of the kind of model we might have for the \( Y_I \), any inference through likelihood ratios does not depend on the sample design, if this is non-informative, in the sense described above – that \( \Pr[Z_I | Y_I] = \Pr[Z_I] \) – membership of the \( I \)-th population member in the sample does not depend on the value of the response \( Y_I \).
In survey sampling theory, this difficulty is countered by the difficulty of the
dependence of model-based inference on the correctness of the model – if this is
incorrect, the conclusions from the analysis could be wrong. Since every model is
by definition wrong (as it is a simplification), the risk of wrong conclusions from
the model-based approach is inherent in the approach. Since survey samplers
are frequently working to a time line for analysis, they cannot spend much time
on investigating and validating suitable models for the response \( Y \). Also, if the
sample design is informative, likelihood-based inference becomes much more
difficult because the form of dependence in \( \Pr[Z_i \mid Y_i] \) needs to be specified and
included in the likelihood.

In Chapter 2 we address the probability model specification difficulty, using a minimal probability model and prior which make no restrictive smooth
assumptions – the model is always correct!

### 1.5 Non-model based repeated sampling

Other data analysis methods have been invented which do not use specific models, or if they do, do not use likelihood-based methods to analyse the data. Some inventors of methods are dissatisfied with the model-based approach – they feel models are restrictive and stand in the way of creative, innovative approaches
to data analysis. The resulting methods are inherently ad hoc from a theoretical viewpoint because they are not based on general theoretical principles. So their properties have to be assessed by repeated sampling experiments from known data structures, in terms of test sizes or coverage properties of intervals in repeated sampling.

Some methods are closely related to formal statistical models, and when viewed or restructured in terms of models can be improved or evaluated by standard statistical theory. Neural networks are an example: they have received much statistical attention, and the multi-layer perceptron or feed-forward neural network has been re-expressed as a formal latent variable model by several statisticians, including Aitkin and Foxall (2002), who pointed out difficulties with the likelihood in the conventional formulation, and gave improved algorithms for fitting the re-expressed model by maximum likelihood.

The danger of repeated sampling inference which ignores the model is clear from the simple case of the uniform distribution with known scale but unknown location. We have a random sample \((y_1, \ldots, y_n)\) from the uniform distribution with density

\[
    f(y \mid \theta) = \begin{cases} 
        1 & \text{for } y \in (\theta - 0.5, \theta + 0.5) \\
        0 & \text{for } y \notin (\theta - 0.5, \theta + 0.5). 
    \end{cases}
\]

A logical analysis can be based on the repeated-sampling distribution of \( \bar{y} \), which rapidly approaches \( N(\theta, 1/(12n)) \). So an approximate 95% confidence interval for \( \theta \) is \( \bar{y} \pm 1.96/\sqrt{12n} \). As an example, here is a sample of 10 from this distribution in which the true \( \theta \) is 0.9:
with mean $\bar{y} = 0.976$, and approximate 95% confidence interval $0.976 \pm 0.179 = [0.797, 1.155]$. The confidence interval will indeed have the correct coverage asymptotically. However the likelihood function is

$$L(\theta) = \begin{cases} 1 & \text{for } \theta \in (y_{(n)} - 0.5, y_{(1)} + 0.5) \\ 0 & \text{for other } \theta, \end{cases}$$

where $y_{(1)}$ and $y_{(n)}$ are the smallest and largest order statistics of the sample. So the interval $[y_{(n)} - 0.5, y_{(1)} + 0.5] = [0.761, 1.049]$ is a 100% confidence interval for $\theta$ – we are certain that $\theta$ lies in this interval, and equally certain that it does not lie outside it! The approximate 95% repeated-sampling interval is longer, and contains impossible values in the interval $(1.049, 1.155]$.

Further, there is no preference for one value of $\theta$ over another within the 100% confidence interval – all values of $\theta$ are equally well supported.

The inference based on the sampling distribution of $\bar{y}$, though formally correct asymptotically in its statement of coverage probability, is irrelevant to the actual information in the data, because it does not recognise the properties of the statistical model for the data.

A more careful analysis (Welsh 1996 pp. 157-9) leads to a similar conclusion for the sampling distribution of any estimator of $\theta$, like the mid-range, defined as $(y_{(1)} + y_{(n)})/2$: unless this is conditioned on the ancillary statistic (the range $y_{(n)} - y_{(1)}$), similar impossible inferences result.