Bayesian Statistics

2: Exchangeability and Conjugate Modeling

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu

http://www.ams.ucsc.edu/~draper

© 2005 David Draper (all rights reserved)
Contents

2.1 Probability as quantification of uncertainty about observables. Binary outcomes. *Case Study:* Hospital-specific prediction of patient-level mortality rates

2.2 Review of frequentist modeling and maximum-likelihood inference

2.3 Exchangeability as a Bayesian concept parallel to frequentist independence

2.4 Prior, posterior, and predictive distributions

2.5 Inference and prediction. Coherence and calibration

2.6 Conjugate analysis. Comparison with frequentist modeling

2.7 The exponential family; conjugate priors

2.8 Integer-valued outcomes; Poisson modeling. *Case Study:* Hospital length of stay for birth of premature babies

2.9 Continuous outcomes; Gaussian modeling. Multivariate unknowns; marginal posterior distributions. *Case Study:* Measurement of physical constants

2.10 References
2.1 Probability as quantification of uncertainty about observables. Binary outcomes

Case Study: Hospital-specific prediction of mortality rates. Let’s say you’re interested in measuring the quality of care (e.g., Kahn et al., 1990) offered by one particular hospital.

I’m thinking of the Dominican Hospital (DH) in Santa Cruz, CA; you may well have a different hospital in mind.

As part of this you decide to examine the medical records of all patients treated at the DH in one particular time window, say January 2002–December 2005, for one particular medical condition for which there is a strong process-outcome link, say acute myocardial infarction (AMI; heart attack).

(Process is what health care providers do on behalf of patients; outcomes are what happens as a result of that care.)

In the time window you’re interested in there will be about \( n = 400 \) AMI patients at the DH.
The Meaning of Probability

To keep things simple let's ignore process for the moment and focus here on one particular outcome: death status (mortality) as of 30 days from hospital admission, coded 1 for dead and 0 for alive.

(In addition to process this will also depend on the sickness at admission of the AMI patients, but for simplicity let's ignore that initially too.)

From the vantage point of December 2001, say, what may be said about the roughly 400 1s and 0s you will observe in 2002–05?

_The meaning of probability._ You are definitely uncertain about the 0–1 death outcomes $Y_1, \ldots, Y_n$ before you observe any of them.

Probability is supposed to be the part of mathematics concerned with quantifying uncertainty; can probability be used here?

In part 1 I argued that the answer was yes, and that three types of probability—classical, frequentist, and Bayesian—are available (in principle) to quantify uncertainty like that encountered here.
2.2 Review of Frequentist Modeling

I'll focus on the approaches with the most widespread usage—frequentist and Bayesian—in what follows.

How can the frequentist definition of probability be applied to the hospital mortality problem?

By definition the frequentist approach is based on the idea of hypothetical or actual repetitions of the process being studied, under conditions that are as close to independent identically distributed (IID) sampling as possible.

When faced with a data set like the 400 1s and 0s \((Y_1, \ldots, Y_n)\) here, the usual way to do this is to think of it as a random sample, or like a random sample, from some population that is of direct interest to you.

Then the randomness in your probability statements refers to the process of what you might get if you were to repeat the sampling over and over—the \(Y_i\) become random variables whose probability distribution is determined by this hypothetical repeated sampling.
Frequentist Modeling (continued)

On the previous page SD stands for standard deviation, the most common measure of the extent to which the observations \( y_i \) in a data set vary, or are spread out, around the center of the data.

The center is often measured by the mean \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \), and the SD of a sample of size \( n \) is then given by

\[
SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{1}
\]

The population size is denoted by \( N \); this is often much larger than the sample size \( n \).

With 0/1 (dichotomous) data, like the mortality outcomes in this case study, the population mean \( \mu \) simply records the proportion \( p \) of 1s in the population (check this), and similarly the sample mean \( \bar{y} \) keeps track automatically of the observed death rate \( \hat{p} \) in the sample.

As \( N \to \infty \) the population SD \( \sigma \) with 0/1 data takes on a simple form (check this):

\[
\sigma = \sqrt{p(1-p)}. \tag{2}
\]

It's common in frequentist modeling to make a notational distinction between the random variables \( Y_i \) (the placeholders for the process of making IID draws from the population over and over) and the values \( y_i \) that the \( Y_i \) might take on (although I'll abuse this notation with \( \hat{p} \) below).
Frequentist Modeling (continued)

Population

30-day mortality

\[ N = \begin{pmatrix} a & 0 & \cdots & 0 \\ 0 & a & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a \end{pmatrix} \]

Like IID

Sample

The observed data

30-day mortality

\[ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \]

\[ n = 400 \]

mean \( \bar{y} = \hat{p} = 0.18 \) (say)

\[ \text{long run mean of } \hat{p} \]

\[ E(\hat{p}) = \mu = p \]

\[ \text{long run variance of } \hat{p} \]

\[ \sqrt{\text{var}(\hat{p})} \]

\[ SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \]

\[ \text{est. } \sqrt{\text{var}(\hat{p})} \]

CLT

Density

\[ \text{when } n < N \quad (n \text{ is a lot smaller than } N) \]

\[ \text{IID = at random without replacement} \]

replacement
Frequentist Modeling (continued)

In the diagram on page 6 the relationship between the population and the sample data sets can be usefully considered in each of two directions:

- If the population is known you can think about how the sample is likely to come out under IID sampling—this is a probability question.

Here in this case \( p \) would be known and you’re trying to figure out the random behavior of the sample mean \( \bar{Y} = \hat{p} \).

- If instead only the sample is known your job is to infer the likely composition of the population that could have led to this IID sample—this is a question of statistical inference.

In this problem the sample mean \( \bar{y} = \hat{p} \) would be known and your job would be to estimate the population mean \( p \).

Suppose that \( N \gg n \), i.e., that even if SRS was used you are effectively dealing with IID sampling.

Intuitively both SRS and IID should be “good”—representative—sampling methods, so that \( \hat{p} \) should be a “good” estimate of \( p \), but what exactly does the word “good” mean in this sentence?

Evidently a good estimator \( \hat{p} \) would be likely to be close to the truth \( p \), especially with a lot of data (i.e., if \( n \) is large).

In the frequentist approach to inference quantifying this idea involves imagining how \( \hat{p} \) would have come out if the process by which the observed \( \hat{p} = 0.18 \) came to you were repeated under IID conditions.

This gives rise to the imaginary data set, the third part of the diagram on page 6: we consider all possible \( \hat{p} \) values based on an IID sample of size \( n \) from a population with \( 100p\% \) 1s and \( 100(1-p)\% \) 0s.
Frequentist Modeling (continued)

Let \( M \) be the number of hypothetical repetitions in the imaginary data set.

The long-run mean (as \( M \to \infty \)) of these imaginary \( \hat{p} \) values is called the expected value of the random variable \( \hat{p} \), written \( E(\hat{p}) \) or \( E_{\text{IID}}(\hat{p}) \) to emphasize the mechanism of drawing the sample from the population.

The long-run SD of these imaginary \( \hat{p} \) values is called the standard error of the random variable \( \hat{p} \), written \( SE(\hat{p}) \) or \( SE_{\text{IID}}(\hat{p}) \).

It's natural in studying how the hypothetical \( \hat{p} \) values vary around the center of the imaginary data set to make a histogram of these values: this is a plot with the possible values of \( \hat{p} \) along the horizontal scale and the frequency with which \( \hat{p} \) takes on those values on the vertical scale.

It's helpful to draw this plot on the density scale, which just means that the vertical scale is chosen so that the total area under the histogram is 1.

The long-run histogram of the imaginary \( \hat{p} \) values on the density scale is called the (probability) density of the random variable \( \hat{p} \).

The values of \( E(\hat{p}) \) and \( SE(\hat{p}) \), and the basic shape of the density of \( \hat{p} \), can be determined mathematically (under IID sampling) and verified by simulation.

It turns out that

\[
E_{\text{IID}}(\hat{p}) = p \quad \text{and} \quad SE_{\text{IID}}(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}},
\]

and the density of \( \hat{p} \) for large \( n \) is well approximated by the normal curve or Gaussian distribution (this result is the famous Central Limit Theorem (CLT)).
Frequentist Modeling (continued)

Suppose the sample of size \( n = 400 \) had 72 1s and 328 0s, so that \( \hat{p} = \frac{72}{400} = 0.18 \).

Thus you would estimate that the population mortality rate \( p \) is around 18\%, but how much uncertainty should be attached to this estimate?

The above standard error formula is not directly usable because it involves the unknown \( p \), but we can estimate the standard error by plugging in \( \hat{p} \):

\[
\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.18)(0.82)}{400}} \approx 0.019. \tag{4}
\]

In other words, I think \( p \) is around 18\%, give or take about 1.9\%.

A probabilistic uncertainty band can be obtained with the frequentist approach by appeal to the CLT, which says that (for large \( n \)) in repeated sampling \( \hat{p} \) would fluctuate around \( p \) like draws from a normal curve with mean \( p \) and SD (SE) 0.019, i.e.,

\[
0.95 \approx P_F \left[ p - 1.96 \text{SE}(\hat{p}) \leq \hat{p} \leq p + 1.96 \text{SE}(\hat{p}) \right] \\
= P_F \left[ \hat{p} - 1.96 \text{SE}(\hat{p}) \leq p \leq \hat{p} + 1.96 \text{SE}(\hat{p}) \right]. \tag{5}
\]

Thus (Neyman 1923) a 95\% (frequentist) confidence interval for \( p \) runs from \( \hat{p} - 1.96 \text{SE}(\hat{p}) \) to \( \hat{p} + 1.96 \text{SE}(\hat{p}) \), which in this case is from 0.180 – (1.96)(0.019) = 0.142 to 0.180 + (1.96)(0.019) = 0.218, i.e., I am “95\% confident that \( p \) is between about 14\% and 22\%”.

But what does this mean?
Frequentist Modeling (continued)

Everybody **wants** the confidence interval (CI) to mean

\[ P_F(0.142 \leq p \leq 0.218) = 0.95, \]  

but it **can't** mean that in the frequentist approach to probability: in that approach \( p \) is treated as a **fixed unknown constant**, which either **is** or **is not** between 0.142 and 0.218.

So what **does** it mean?

This is a kind of **calibration** of the CI process: about 95% of the nominal 95% CIs would include the true value, if you were to generate a lot of them via independent IID samples from the population.
Frequentist Modeling (continued)

The diagram on page 6 takes up a lot of space; it would be nice to have a more succinct summary of it.

A random variable $Y$ is said to follow the **Bernoulli distribution** with parameter $0 < p < 1$—this is summarized by saying $Y \sim B(p)$—if $Y$ takes on only the values 1 and 0 and

$$P(Y = y) = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{cases} = p^y (1 - p)^{1-y}. \quad (7)$$

A parameter is just a fixed unknown constant.

Another **popular name** for the parameter $p$ in this model is $\theta$.

Evidently what the population and sample parts of the diagram on page 6 are trying to say, in this notation, is that $(Y_1, \ldots, Y_n)$ are drawn **in an IID fashion** from the Bernoulli distribution with parameter $\theta$.

In the usual **shorthand**, which I'll use from now on, this is expressed as

$$Y_i \overset{\text{IID}}{\sim} B(\theta), \quad i = 1, \ldots, n \quad \text{for some } 0 < \theta < 1. \quad (8)$$

This is the **frequentist statistical model** for the AMI mortality data, except that we have forgotten so far to specify an important ingredient: **what is the population** of patients whose mean (underlying death rate) is $\theta$?

As a frequentist (recall page 5), to use probability to quantify your uncertainty about the 1s and 0s, you have to think of them as either literally a **random sample** or like a random sample from some population, either hypothetical or actual.
Frequentist Modeling (continued)

What are some possibilities for this population?

- All AMI patients who might have come to the DH in 2002–05 if the world had turned out differently; or

- Assuming sufficient time-homogeneity in all relevant factors, you could try to argue that the collection of all 400 AMI patients at the DH from 2002–05 is like a random sample of size 400 from the population of all AMI patients at the DH from (say) 1997–2006; or

- Cluster sampling is a way to choose, e.g., patients by taking a random sample of hospitals and then a random sample of patients nested within those hospitals. What we actually have here is a kind of cluster sample of all 400 AMI patients from the DH in 2002–2005 (and no patients from any other hospitals). Cluster samples tend to be less informative than SRS samples of the same size because of (positive) intracluster correlation (patients in a given hospital tend to be more similar in their outcomes than would an SRS of the same size from the population of all the patients in all the hospitals). Assuming the DH to be representative of some broader collection of hospitals in California and (unwisely) ignoring intracluster correlation, you could try to argue that these 400 1s and 0s were like a simple random sample of 400 AMI patients from this larger collection of hospitals.

None of these options is entirely compelling.

If you’re willing to pretend the data are like a sample from some population, interest would then focus on inference about the parameter $\theta$, the “underlying death rate” in this larger collection of patients to which you feel comfortable generalizing the 400 1s and 0s: if $\theta$ were unusually high, that would be prima facie evidence of a possible quality of care problem at the DH.
The Likelihood Function

Suppose (as above) that

\[ Y_i^{\text{IID}} \sim B(\theta), \quad i = 1, \ldots, n \quad \text{for some } 0 < \theta < 1. \quad (9) \]

Since the \( Y_i \) are independent, the joint sampling distribution of all of them, \( P(Y_1 = y_1, \ldots, Y_n = y_n) \), is the product of the separate, or marginal, sampling distributions \( P(Y_1 = y_1), \ldots, P(Y_n = y_n) \):

\[
P(Y_1 = y_1, \ldots, Y_n = y_n) = P(Y_1 = y_1) \cdots P(Y_n = y_n)
= \prod_{i=1}^{n} P(Y_i = y_i). \quad (10)
\]

But since the \( Y_i \) are also identically distributed, and each one is Bernoulli(\( \theta \)), i.e., \( P(Y_i = y_i) = \theta^{y_i} (1 - \theta)^{1-y_i} \), the joint sampling distribution can be written

\[
P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \theta^{y_i} (1 - \theta)^{1-y_i}. \quad (11)
\]

Let's use the symbol \( y \) to stand for the vector of observed data values \((y_1, \ldots, y_n)\).

Before any data have arrived, this joint sampling distribution is a function of \( y \) for fixed \( \theta \)—it tells you how the data would be likely to behave in the future if you were to take an IID sample from the Bernoulli(\( \theta \)) distribution.
The Likelihood Function (continued)

In 1921 Fisher had the following idea: after the data have arrived it makes more sense to interpret (11) as a function of $\theta$ for fixed $y$—he called this the **likelihood function** for $\theta$ in the Bernoulli($\theta$) model:

$$l(\theta|y) = l(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} \theta^{y_i} (1 - \theta)^{1-y_i}$$  \hspace{1cm} (12)

$$= P(Y_1 = y_1, \ldots, Y_n = y_n) \text{ but interpreted as a function of } \theta \text{ for fixed } y.$$  

Fisher tried to create a theory of inference about $\theta$ based only on this function—we will see below that this is an important ingredient, but **not the only important ingredient**, in inference from the Bayesian viewpoint.

The Bernoulli($\theta$) likelihood function can be **simplified** as follows:

$$l(\theta|y) = \theta^s (1 - \theta)^{n-s},$$  \hspace{1cm} (13)

where $s = \sum_{i=1}^{n} y_i$ is the **number of 1s** in the sample and $(n - s)$ is the **number of 0s**.

What does this function **look like**?

With $n = 400$ and $s = 72$ it's easy to get Maple to **plot it**:

```maple
rosalind 329> maple

\[
\begin{align*}
|\wedge| & \quad \text{Maple V Release 5 (University of California, Santa Cruz)} \\
_.\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\!\![15]$
```
The Likelihood Function (continued)

> plot( l( theta, 72, 400 ), theta = 0.12 .. 0.25 );

Does this function remind you of anything?
The Likelihood Function (continued)

Note that the likelihood function \( l(\theta|y) = \theta^s(1-\theta)^{n-s} \) in this problem depends on the data vector \( y \) only through \( s = \sum_{i=1}^{n} y_i \)—Fisher referred to any such data summary as a **sufficient statistic** (with respect to the given likelihood function).

It's often at least as useful to look at the **logarithm** of the likelihood function as the likelihood function itself:

\[
> ll := ( \theta, s, n ) \rightarrow \log( l( \theta, s, n ) );
\]

\[
> \text{plot}( ll( \theta, 72, 400 ), \theta = 0.12 .. 0.25 );
\]

![Graph of log likelihood function](image)

In this case, as is often true for large \( n \), the log likelihood function looks **locally quadratic around its maximum**.

Fisher had the further idea that the **maximum** of the likelihood function would be a good **estimate** of \( \theta \) (we'll look later at conditions under which this makes sense from the **Bayesian** viewpoint).
The Likelihood Function (continued)

Since the logarithm function is monotone increasing, it's equivalent in maximizing the likelihood to maximize the log likelihood, and for a function as well behaved as this you can do that by setting its first partial derivative with respect to \( \theta \) to 0 and solving:

\[
> \text{score} := \text{simplify} \left( \text{diff} \left( \text{ll} \left( \text{theta}, s, n \right), \text{theta} \right) \right);
\]

\[
s - n \, \text{theta} \\
\text{score} := - \frac{s - n \, \text{theta}}{\text{theta} \left( -1 + \text{theta} \right)}
\]

\[
> \text{solve} \left( \text{score} = 0, \text{theta} \right);
\]

\[
\frac{s}{n}
\]

\[
> \text{quit};
\]

bytes used=2125632, alloc=1376004, time=0.51

rosalind 330>

The function of the data that maximizes the likelihood (or log likelihood) function is called the maximum likelihood estimate (MLE) \( \hat{\theta}_{\text{MLE}} \).

Thus in this case \( \hat{\theta}_{\text{MLE}} \) is just the sample mean \( \bar{y} = \frac{s}{n} \), which we've previously seen is a sensible estimate of \( \theta \).

Note also that if you maximize \( l(\theta|y) \) and I maximize \( c \, l(\theta|y) \) for any constant \( c > 0 \), we'll get the same thing, i.e., the likelihood function is only defined up to a positive multiple; Fisher's actual definition was

\[
l(\theta|y) = c \, P(Y_1 = y_1, \ldots, Y_n = y_n)
\]

for any (normalizing constant) \( c > 0 \) (this will be put to Bayesian use below).

From now on \( c \) in expressions like the likelihood function above will be a generic (and often unspecified) positive constant.
Calibrating the MLE

Maximum likelihood provides a basic principle for estimation of a (population) parameter $\theta$ from the frequentist/likelihood point of view, but how should the accuracy of $\hat{\theta}_{\text{MLE}}$ be assessed?

Evidently in the frequentist approach we want to compute the variance or standard error of $\hat{\theta}_{\text{MLE}}$ in repeated sampling, or at least estimates of these quantities—let’s focus on the estimated variance $\hat{\mathcal{V}}(\hat{\theta}_{\text{MLE}})$.

Fisher (1922) proposed an approximation to $\hat{\mathcal{V}}(\hat{\theta}_{\text{MLE}})$ that works well for large $n$ and makes good intuitive sense.

In the AMI mortality case study, where

$\hat{\theta}_{\text{MLE}} = \hat{\theta} = \frac{s}{n}$ (the sample mean),

we already know that

$$V(\hat{\theta}_{\text{MLE}}) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad \hat{\mathcal{V}}(\hat{\theta}_{\text{MLE}}) = \frac{\hat{\theta}(1-\hat{\theta})}{n},$$

but Fisher wanted to derive results like this in a more basic and general way.
Calibrating the MLE (continued)

Imagine **quadrupling** the sample size in this case study from $n = 400$ to $n = 1600$ while keeping the observed death rate constant at 0.18—what would happen to the log likelihood function?

To answer this question, recall (p. 18) that as far as maximizing the likelihood function is concerned it's equally good to work with **any (positive) constant multiple** of it, which is equivalent to saying that we can **add any constant** we want to the log likelihood function without harming anything.

In the Maple plot below I've added a **different constant** to each of the log likelihood functions with $(s, n) = (72, 400)$ and $(288, 1600)$ so that they both go through the point $(\hat{\theta}_{\text{MLE}}, 0)$:

```maple
> plot( { ll( theta, 72, 400 ) - evalf( ll( 72 / 400, 72, 400 ) ),
         ll( theta, 288, 1600 ) - evalf( ll( 288 / 1600, 288, 1600 ) ) },
       theta = 0.12 .. 0.25, color = black );
```

![Maple plot]

---

20