

# One-Day Short Course on Bayesian Modeling, Inference and Prediction

## 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`

*Sponsored by the Boston Chapter  
of the American Statistical Association*

10 December 2004, 8am–5.30pm  
Hotel@MIT, 20 Sidney Street, Cambridge MA

© 2004 David Draper (all rights reserved)

## **2: Exchangeability and Conjugate Modeling**

### **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 would probably 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 2000–December 2003**, 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 let's ignore that initially too.)

From the vantage point of December 1999, say, **what may be said** about the roughly 400 1s and 0s you will observe in 2000–03?

*The meaning of probability.* You are definitely **uncertain** about the 0–1 death outcomes  $Y_1, \dots, Y_n$  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.

The **classical** approach turns out to be **impractical** to implement in all but the simplest problems; we'll focus here on the **frequentist** and **Bayesian** stories.

## 2.2 Review of Frequentist Modeling

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  $(Y_1, \dots, Y_n)$  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.

In the absence of any **predictor information** the off-the-shelf **frequentist model** for this situation is of course

$$Y_i \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta), \quad i = 1, \dots, n \quad (1)$$

for some  $0 < \theta < 1$ , but what is the **population** to which it's appropriate to **generalize outward** from the 400 1s and 0s that will be observed?

## Frequentist Modeling (continued)

Here are some **possibilities**:

- All AMI patients who **might have** come to the DH in 2000–03 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 2000–03 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 2000–2003. 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.

## The Likelihood Function

Suppose (as above) that the frequentist model is

$$Y_i \stackrel{\text{IID}}{\sim} B(\theta), \quad i = 1, \dots, n \quad \text{for some } 0 < \theta < 1. \quad (2)$$

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

$$\begin{aligned} P(Y_1 = y_1, \dots, Y_n = y_n) &= P(Y_1 = y_1) \cdots P(Y_n = y_n) \\ &= \prod_{i=1}^n P(Y_i = y_i). \end{aligned} \quad (3)$$

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

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i}. \quad (4)$$

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

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 (as you know) Fisher had the following idea: **after** the data have arrived it makes more sense to interpret (4) as a function of  $\theta$  for fixed  $y$ —this is the **likelihood function** for  $\theta$  in the Bernoulli( $\theta$ ) model:

$$\begin{aligned} l(\theta|y) &= l(\theta|y_1, \dots, y_n) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} & (5) \\ &= P(Y_1 = y_1, \dots, Y_n = y_n) \text{ but interpreted} \\ &\quad \text{as a function of } \theta \text{ for fixed } y. \end{aligned}$$

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:

$$l(\theta|y) = \theta^s (1 - \theta)^{n-s}, \quad (6)$$

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**:

```
rosalind 329> maple
```

```
  | \ ^ / |      Maple V Release 5 (University of California, Santa Cruz)
  . _ | \ |   | / | _ . Copyright (c) 1981-1997 by Waterloo Maple Inc. All rights
  \  MAPLE  / reserved. Maple and Maple V are registered trademarks of
  <-----> Waterloo Maple Inc.
    |      Type ? for help.
```

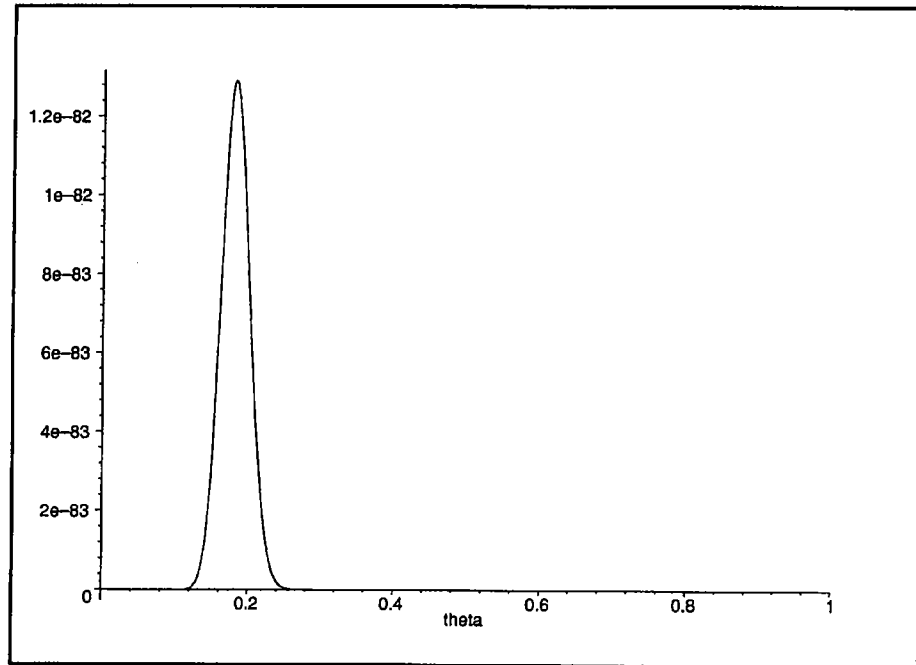
```
> l := ( theta, s, n ) -> theta^s * ( 1 - theta )^( n - s );
```

```
                s                (n - s)
l := (theta, s, n) -> theta  (1 - theta)
```

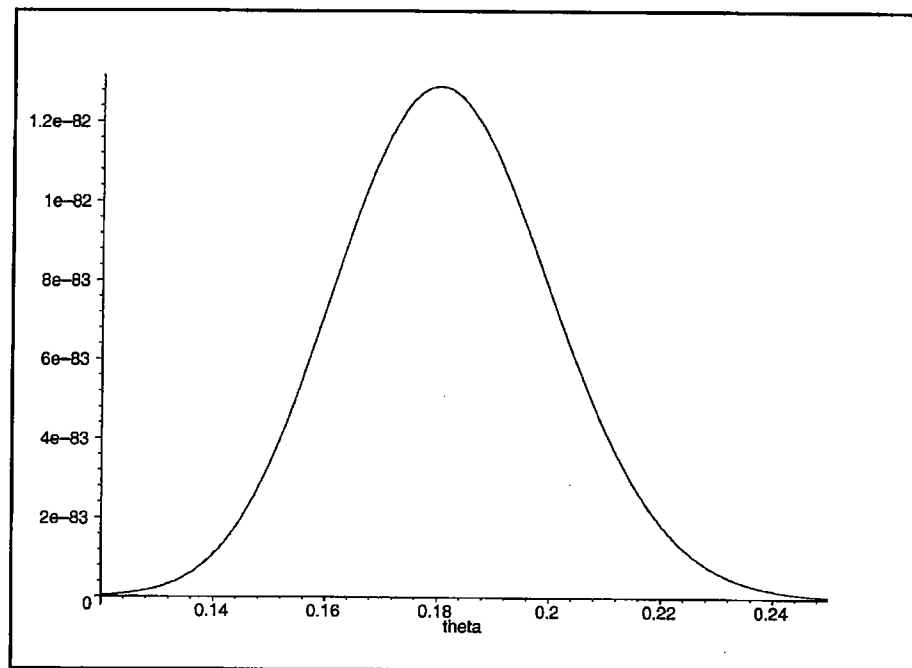
```
> plotsetup( x11 );
```

```
> plot( l( theta, 72, 400 ), theta = 0 .. 1 );
```

# The Likelihood Function (continued)



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



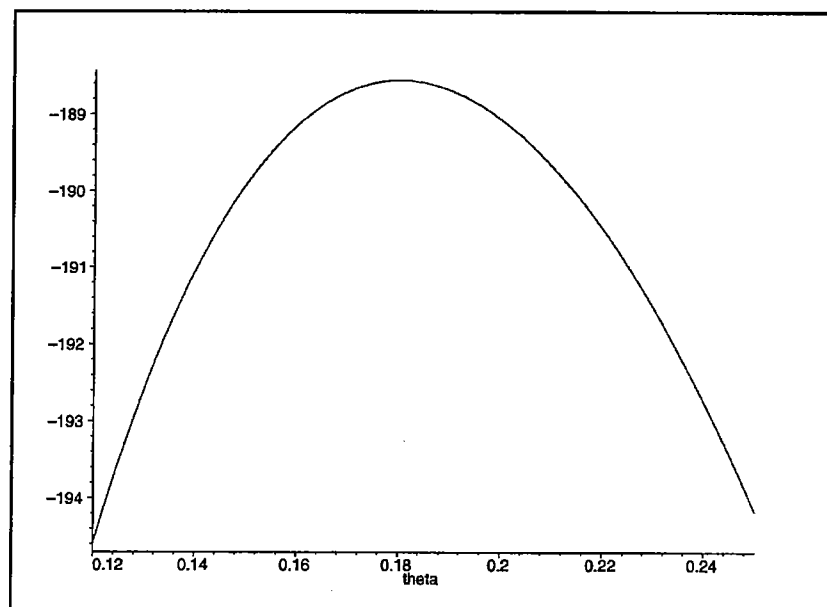
As you can see, this looks a lot like a **Gaussian distribution** (not yet density-normalized) for  $\theta$ , which is the **Bayesian** way to **interpret** the likelihood function (see below).

## 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  $s = \sum_{i=1}^n y_i$** —(as you know) 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 ) -> log( l( theta, s, n ) );  
> plot( ll( theta, 72, 400 ), theta = 0.12 .. 0.25 );
```



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

Fisher had (as you know) 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:

```
> score := simplify( diff( ll( theta, s, n ), theta ) );
```

$$\text{score} := - \frac{s - n \text{ theta}}{\text{theta} (-1 + \text{theta})}$$

```
> solve( score = 0, theta );
```

$$s/n$$

```
> 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 (as you know) the **maximum likelihood estimate (MLE)**  $\hat{\theta}_{\text{MLE}}$ .

Thus in this case  $\hat{\theta}_{\text{MLE}}$  is just the **sample mean**  $\frac{s}{n}$ , 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

$$l(\theta|y) = c P(Y_1 = y_1, \dots, Y_n = y_n)$$

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 estimated versions of these quantities—let's focus on the estimated variance  $\hat{V}(\hat{\theta}_{\text{MLE}})$ .

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

In the AMI mortality case study, where

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

we already know that

$$V(\hat{\theta}_{\text{MLE}}) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad \hat{V}(\hat{\theta}_{\text{MLE}}) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \quad (7)$$

but Fisher wanted to derive results like this in a more **basic** and **general** way.

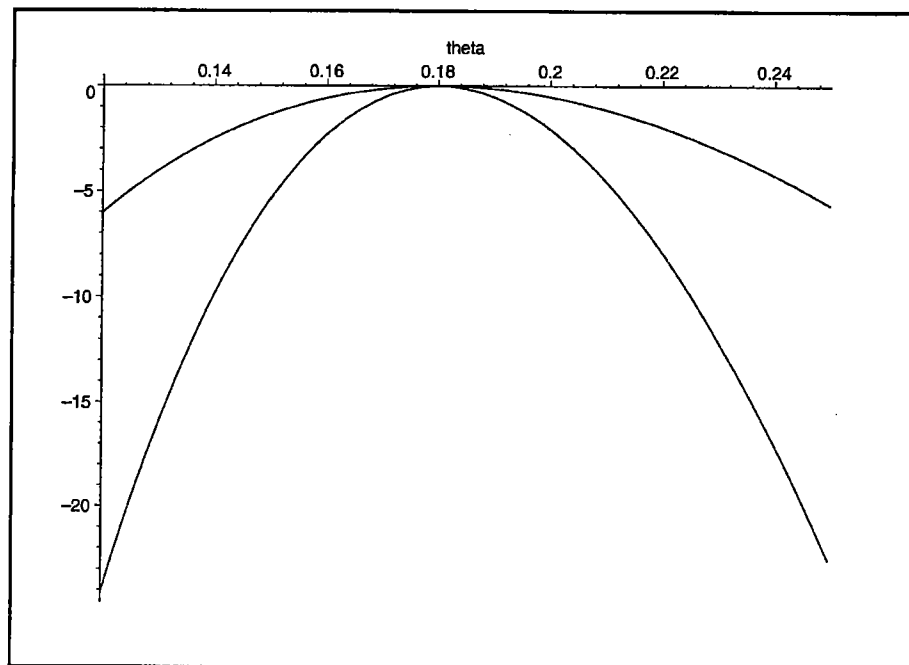
## Calibrating the MLE (continued)

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, observe first 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  $(\hat{\theta}_{MLE}, 0)$ :

```
> 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 );
```



## Calibrating the MLE (continued)

Notice that what's happened as  $n$  went from 400 to 1600 while holding the MLE constant at 18% mortality is that the **second derivative of the log likelihood function at  $\hat{\theta}_{MLE}$**  (a negative number) has **increased** in size.

This led Fisher (as you know) to define the **information** in the sample about  $\theta$ —in his honor (as you know) it's now called the (observed) **Fisher information**:

$$\hat{I}(\hat{\theta}_{MLE}) = \left[ -\frac{\partial^2}{\partial \theta^2} \log l(\theta|y) \right]_{\theta=\hat{\theta}_{MLE}} \quad (8)$$

This quantity **increases** as  $n$  goes up, whereas our uncertainty about  $\theta$  based on the sample, as measured by  $\hat{V}(\hat{\theta}_{MLE})$ , should go **down** with  $n$ .

Fisher conjectured and proved that the information and the estimated variance of the MLE in repeated sampling have the following simple **inverse relationship** when  $n$  is large:

$$\hat{V}(\hat{\theta}_{MLE}) \doteq \hat{I}^{-1}(\hat{\theta}_{MLE}). \quad (9)$$

He further proved that for large  $n$  (a) the MLE is approximately **unbiased**, meaning that in repeated sampling

$$E(\hat{\theta}_{MLE}) \doteq \theta, \quad (10)$$

and (b) the sampling distribution of the MLE is approximately **normal** with mean  $\theta$  and estimated variance given by (9):

$$\hat{\theta}_{MLE} \sim N[\theta, \hat{I}^{-1}(\hat{\theta}_{MLE})]. \quad (11)$$

Thus for large  $n$  an **approximate 95% confidence interval** for  $\theta$  is given by  $\hat{\theta}_{MLE} \pm 1.96\sqrt{\hat{I}^{-1}(\hat{\theta}_{MLE})}$ .

## Calibrating the MLE (continued)

You can **differentiate** to compute the information yourself in the AMI mortality case study, or you can use Maple to do it for you:

```
> score := ( theta, s, n ) -> simplify( diff( ll( theta, s, n ), theta ) );
```

```
score := (theta, s, n) -> simplify(diff(ll(theta, s, n), theta))
```

```
> score( theta, s, n );
```

$$\frac{s - n \theta}{\theta (-1 + \theta)}$$

```
> diff2 := ( theta, s, n ) -> simplify( diff( score( theta, s, n ),
theta ) );
```

```
diff2 := (theta, s, n) -> simplify(diff(score(theta, s, n), theta))
```

```
> diff2( theta, s, n );
```

$$\frac{-n \theta^2 - s + 2 s \theta}{\theta^2 (-1 + \theta)^2}$$

```
> information := ( s, n ) -> simplify( eval( - diff2( theta, s, n ),
theta = s / n ) );
```

```
> information( s, n );
```

$$\frac{n^3}{s (-n + s)}$$

```
> variance := ( s, n ) -> 1 / information( s, n );
```

```
variance := (s, n) -> 
$$\frac{1}{\text{information}(s, n)}$$

```

## Calibrating the MLE (continued)

> variance( s, n );

$$\frac{s(-n + s)}{n^3}$$

This expression can be **further simplified** to yield

$$\hat{V}(\hat{\theta}_{\text{MLE}}) = \frac{\frac{s}{n} \left(1 - \frac{s}{n}\right)}{n} = \frac{\hat{\theta}(1 - \hat{\theta})}{n}, \quad (12)$$

which **coincides** with (7).

From (12) **another expression** for the Fisher information in this problem is

$$\hat{I}(\hat{\theta}_{\text{MLE}}) = \frac{n}{\hat{\theta}(1 - \hat{\theta})}. \quad (13)$$

As  $n$  increases,  $\hat{\theta}(1 - \hat{\theta})$  will tend to the constant  $\theta(1 - \theta)$  (this is well-defined because we've assumed that  $0 < \theta < 1$ , since  $\theta = 0$  and  $1$  are probabilistically uninteresting), which means that information about  $\theta$  on the basis of  $(y_1, \dots, y_n)$  in the IID Bernoulli model **increases at a rate proportional to  $n$  as the sample size grows**.

This is **generally true** of the MLE (i.e., in **regular parametric problems**):

$$\hat{I}(\hat{\theta}_{\text{MLE}}) = O(n) \quad \text{and} \quad \hat{V}(\hat{\theta}_{\text{MLE}}) = O(n^{-1}), \quad (14)$$

as  $n \rightarrow \infty$ , where the notation  $a_n = O(b_n)$  (as usual) means that the ratio  $\left| \frac{a_n}{b_n} \right|$  is bounded as  $n$  grows.

Thus uncertainty about  $\theta$  on the basis of the MLE **goes down like  $\frac{c_{\text{MLE}}}{n}$  on the variance scale** with more and more data (in fact Fisher showed that  $c_{\text{MLE}}$  achieves the lowest possible value: the MLE is **efficient**).

