AMS 132, Winter 2017: Classical and Bayesian Inference

David Draper

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

draper@ucsc.edu
www.ams.ucsc.edu/∼draper

CLASS WEB PAGE:
ams132-winter17-01.courses.soe.ucsc.edu

Lecture Notes, Part 2 (Conjugate Modeling)

2.1 Probability as quantification of uncertainty about observables; binary outcomes

Case Study: *Hospital-specific prediction of mortality rates.* Suppose I'm 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; if this were Your problem You'd have a different hospital in mind.

As part of this I decide to examine the medical records of all patients treated at the DH in one particular time window, say **January 2006–December 2009**, for one particular medical condition for which there's 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 I'm interested in there will be about $n = 400$ **AMI patients** at the DH.

The Meaning of Probability

To keep things simple I'll 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 I'll ignore that initially too.)

From the vantage point of December 2005, say, **what may be said** about the roughly 400 1s and 0s I'll observe in 2006–09?

The meaning of probability. I'm definitely **uncertain** about the 0–1 death outcomes $Y_1, \ldots, Y_n$ before I 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; I'll focus here on the **frequentist** and **Bayesian** stories.

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 $(Y_1, \ldots, Y_n)$ 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's of direct interest to me.

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.

There are two main **flavors** of **frequentist modeling**: **confidence intervals** (**Neyman**) and **likelihood inference** (**Fisher**); the **diagram** on the next page summarizes the **essence** of the **Neyman frequentist approach**.

The Basic Frequentist Model Diagram

(see sketch presented in class)

Neyman Frequentist Modeling

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

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

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 μ 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** σ 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).

In the diagram on page 5 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**.

The Repeated-Sampling Data Set

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're effectively dealing with **IID sampling**.

Intuitively both **SRS** and **IID** should be "good" — **representative** — **sampling methods** (meaning that the **sampled** and **unsampled** elements of the **population** should be **similar**), 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 **repeated-sampling data set**, the third part of the diagram on page 5: we imagine **all possible** $\hat{p}$ values based on an **IID sample** of size n from a **population** with $100p\%$ 1s and $100(1-p)\%$ 0s.

Expected Value and Standard Error

Let M be the **number of hypothetical repetitions** in the **repeated-sampling 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 **repeated-sampling 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 **frequencies** 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 distribution** (histogram) of the **imaginary** $\hat{p}$ values on the **density scale** is called the **(probability) density** of the random variable $\hat{p}$.

Central Limit Theorem

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

$$E_{\text{IID}}(\hat{p}) = p \quad \text{and} \quad SE_{\text{IID}}(\hat{p}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}, \tag{3}$$

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)**).

Suppose the sample of size $n = 400$ had **72** 1s and **328** 0s; then $\hat{p} = \frac{72}{400} = \mathbf{0.18}$.

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 $\hat{p}$:

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} = \sqrt{\frac{(0.18)(0.82)}{400}} \doteq 0.019. \tag{4}$$

Neyman (1923): Confidence Intervals

In other words, we 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.,

$$\begin{aligned} 0.95 &\doteq P_F\left[p - 1.96\,\widehat{SE}(\hat{p}) \le \hat{p} \le p + 1.96\,\widehat{SE}(\hat{p})\right] \\ &= P_F\left[\hat{p} - 1.96\,\widehat{SE}(\hat{p}) \le p \le \hat{p} + 1.96\,\widehat{SE}(\hat{p})\right]. \end{aligned} \tag{5}$$

Thus a 95% (frequentist) **confidence interval** for p runs from $\hat{p} - 1.96\,\widehat{SE}(\hat{p})$ to $\hat{p} + 1.96\,\widehat{SE}(\hat{p})$, 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., we're **"95% confident that p is between about 14% and 22%"**; but **what does this mean**?

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

$$P_F(0.142 \le p \le 0.218) \doteq 0.95, \tag{6}$$

Meaning of Confidence Intervals; Calibration

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?

(see sketch presented in class)

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**.

Random Variable Shorthand

The **diagram** on page 5 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** $0 < p < 1$ (a **parameter** is just a **fixed unknown constant**) — this is summarized by saying $\boxed{Y \sim \text{Bernoulli}(p)}$ — if Y takes on only the values 1 and 0 and

$$P(Y = y) = \left\{ \begin{array}{ll} p & \text{if } y = 1 \\ 1 - p & \text{if } y = 0 \end{array} \right\} = p^y \, (1 - p)^{1-y}. \tag{7}$$

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

Evidently what the population and sample parts of the diagram on page 5 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 θ.

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

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

What's the Population?

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

As a **frequentist** (recall page 4), 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 (hypothetical or actual); what are some **possibilities** for this population?

- (Fisher: **hypothetical**) All AMI patients who **might have** come to the DH in 2006–09 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 2006–09 is **like** a random sample of size 400 from the population of all AMI patients at the DH from (say) 2000–2015; 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 2006–2009.

What's the Population? (continued)

Cluster samples tend to be less informative than simple random samples (SRSs) 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; but if You're willing to pretend the data are like a sample from some population, interest would then focus on inference about the **parameter** θ, the "underlying death rate" in this larger collection of patients to which You feel comfortable generalizing the 400 1s and 0s: if θ were unusually high, that would be **prima facie** evidence of a possible quality of care problem.

Suppose (**as above**) that the frequentist model is

$$Y_i \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \quad i = 1, \ldots, n \quad \text{for some } 0 < \theta < 1. \tag{9}$$

Fisher Frequentist Modeling

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 $P(Y_1 = y_1), \ldots, P(Y_n = y_n)$:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n) &= P(Y_1 = y_1) \cdots P(Y_n = y_n) \\ &= \prod_{i=1}^{n} P(Y_i = y_i) \,. \end{aligned} \tag{10}$$

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

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

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 θ — 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(θ) distribution.

The Likelihood Function

In 1921 Fisher had the following idea (Laplace (1774) had it first): **after** the data have arrived it makes more sense to interpret (11) as a function of θ for fixed y — this is the **likelihood function** for θ in the Bernoulli(θ) model:

$$\begin{aligned} l(\theta|y) &= l(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \qquad (12) \\ &= P(Y_1 = y_1, \ldots, Y_n = y_n) \text{ but } \mathbf{interpreted} \\ &\quad \mathbf{as\ a\ function\ of}\ \theta\ \mathbf{for\ fixed}\ y. \end{aligned}$$

Fisher tried to create a theory of **inference** about θ **based only on this function** — this turns out to be an important ingredient, **but not the only important ingredient**, in inference from the Bayesian viewpoint.

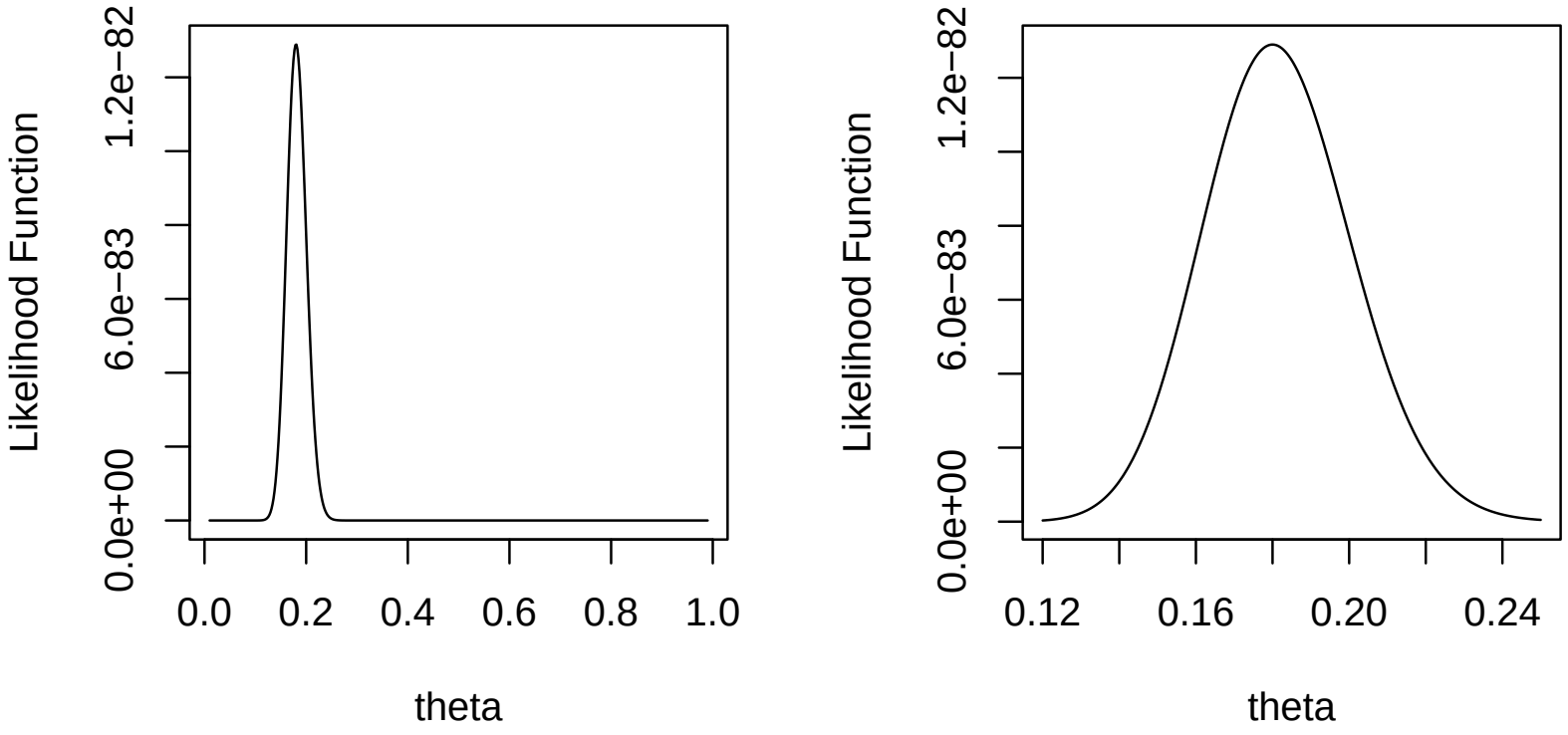
The Bernoulli(θ) likelihood function can be **simplified** as follows:

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

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

The Likelihood Function (continued)

What does this function **look like**, e.g., with $n = 400$ and $s = 72$ (this is similar to data You would get from the DH: a **30-day mortality rate** from AMI of **18%**)?

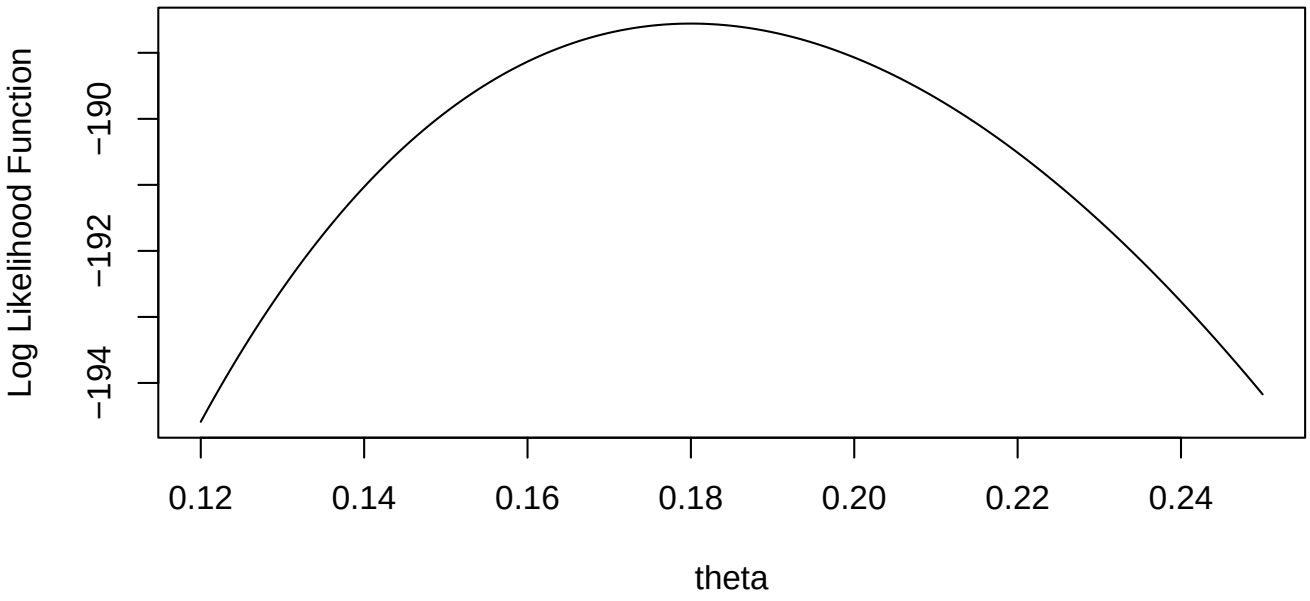


This looks a lot like a **Gaussian distribution** (not yet density-normalized) for θ, which is the **Bayesian** way to **interpret** the likelihood function (see below).

Likelihood and Log Likelihood

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 **assumed sampling model**).

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



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

Maximizing the Likelihood Function

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

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 θ to 0 and solving; here You get the familiar result

$$\hat{\theta}_{\text{MLE}} = \frac{s}{n} = \bar{y}.$$

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

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**) $c > 0$.

Calibrating the MLE

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

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

Evidently in the frequentist approach You want to compute the **variance** or **standard error** of $\hat{\theta}_{\text{MLE}}$ in **repeated sampling**, or estimated versions of these quantities — I'll focus on the estimated variance $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = \left[\widehat{SE}\left(\hat{\theta}_{\text{MLE}}\right)\right]^2$.

Fisher (1922) also proposed an **approximation** to $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right)$ 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**), it's easy to show that

$$V\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{14}$$

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

Examining the Log Likelihood Function

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**?

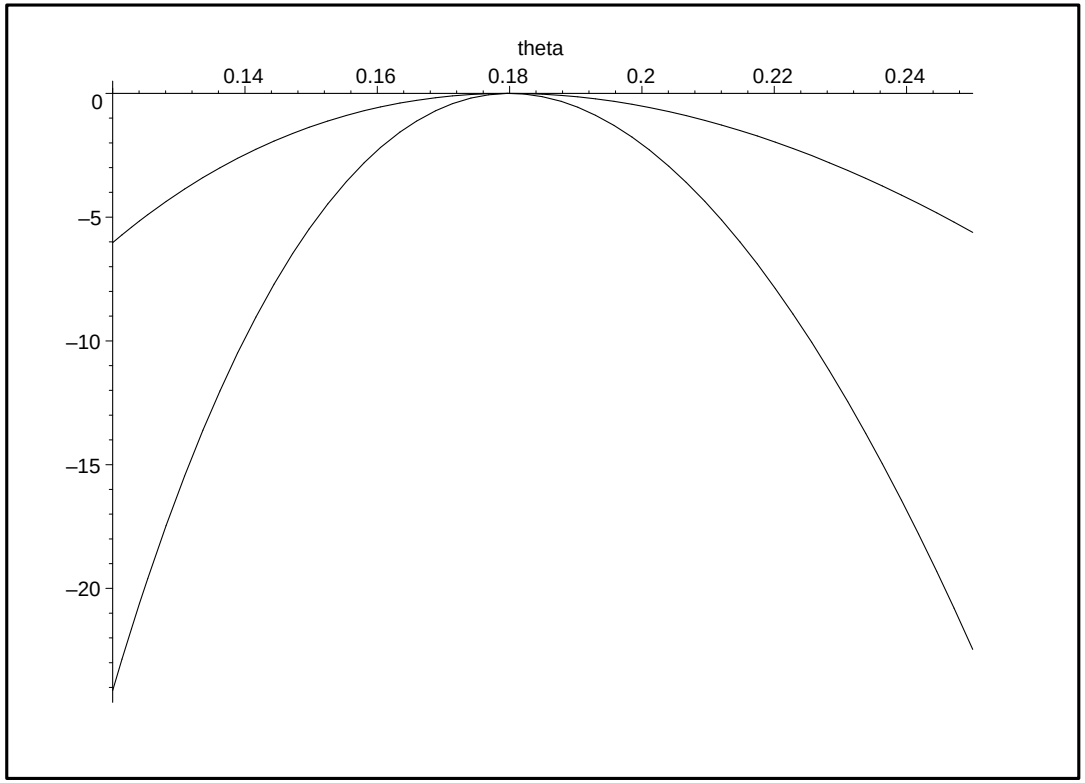
To answer this question, recall (page 20) that as far as maximizing the likelihood function is concerned it's equally good to work with **any (positive) constant multiple** of it, which is equivalent to saying that we can **add any constant** we want to the log likelihood function without harming anything.

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

```
sauternes 235> maple
    |\^/|     Maple 9.5 (SUN SPARC SOLARIS)
._|\|   |/|_. Copyright (c) Maplesoft, Waterloo Maple Inc. 2004
 \  MAPLE  /  All rights reserved. Maple is a trademark of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.
```

Examining the Log Likelihood Function (continued)

```
> ll := ( theta, s, n ) -> s * log( theta ) + ( n - s ) * log( 1 - theta );
> plotsetup( x11 );
> 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 );
```



Fisher Information

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** $\hat{\theta}_{\text{MLE}}$ (a negative number) has **increased** in size.

This led Fisher to define a quantity he called the **information** in the sample about θ — in his honor it's now called the (observed) **Fisher information**:

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \left[-\frac{\partial^2}{\partial\theta^2}\log l(\theta|y)\right]_{\theta=\hat{\theta}_{\text{MLE}}}. \tag{15}$$

This quantity **increases** as n goes up, whereas our uncertainty about θ based on the sample, as measured by $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right)$, 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:

$$\hat{V}\left(\hat{\theta}_{\text{MLE}}\right) \doteq \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right). \tag{16}$$

Likelihood-Based Large-Sample Confidence Intervals

In this case study the **Fisher information** and **repeated-sampling variance** come out

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{n}{\hat{\theta}(1-\hat{\theta})} \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{17}$$

(check this) which matches what You already know is **correct** in this case.

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

$$E\left(\hat{\theta}_{\text{MLE}}\right) \doteq \theta, \tag{18}$$

and (b) the sampling distribution of the MLE is approximately **Gaussian** with mean θ and estimated variance given by (16):

$$\hat{\theta}_{\text{MLE}} \dot{\sim} \text{Gaussian}\left[\theta, \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)\right]. \tag{19}$$

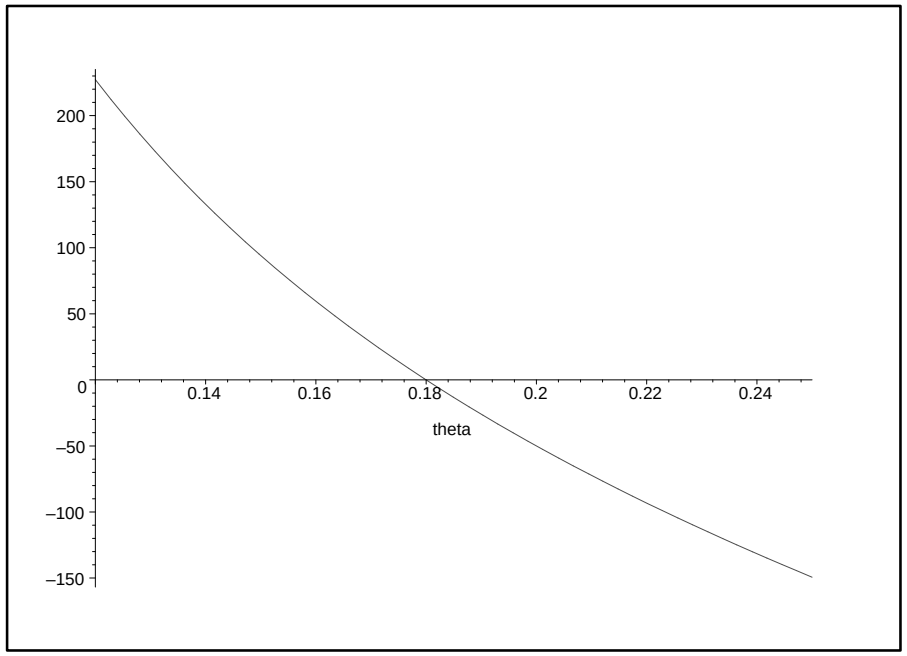
Thus for large n an **approximate 95% confidence interval** for θ is given by

$$\hat{\theta}_{\text{MLE}} \pm 1.96\sqrt{\hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)}.$$

Fisher Information Computation

You can **differentiate** to compute the Fisher 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 );
                                    s - n theta
                                - -----------------
                                  theta (-1 + theta)
> plot( score( theta, 72, 400 ), theta = 0.12 .. 0.25 );
```



Fisher Information Computation (continued)

```
> 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 );
                                    2
                          -n theta  - s + 2 s theta
                          -------------------------
                                2                 2
                              theta  (-1 + theta)

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

> information( s, n );
                                      3
                                     n
                                - ----------
                                  s (-n + s)
```

Fisher Information Computation (continued)

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

$$\texttt{variance} := (s, n) \rightarrow \frac{1}{\texttt{information}(s, n)}$$

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

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

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

$$\hat{V}\left(\hat{\theta}_{\text{MLE}}\right) \doteq \frac{\frac{s}{n}\left(1-\frac{s}{n}\right)}{n} = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{20}$$

which **coincides** with (17).

Repeated-Sampling Asymptotic Optimality of MLE

In the above expression for **Fisher information** in this problem,

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{n}{\hat{\theta}(1-\hat{\theta})},$$

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 θ on the basis of $(y_1, \ldots, y_n)$ 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):

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = O(n) \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = O\left(n^{-1}\right), \tag{21}$$

as $n \to \infty$, 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 θ 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_{MLE} achieves the lowest possible value: the MLE is **efficient**).

Bayesian Modeling

As a Bayesian in this situation, my job is to quantify my uncertainty about the 400 binary **observables** I'll get to see starting in 2006, i.e., my initial modeling task is **predictive** rather than inferential.

There is no samples-and-populations story in this approach, but probability and random variables arise in a different way: quantifying my uncertainty (for the purpose of betting with someone about some aspect of the 1s and 0s, say) requires **eliciting** from myself a joint **predictive** distribution that **accurately** captures my judgments about what I'll see: $\boxed{P_{B:\text{me}}(Y_1 = y_1, \ldots, Y_n = y_n)}$.

Notice that in the frequentist approach the random variables describe the **process** of observing a repeatable event (the "random sampling" appealed to here), whereas in the Bayesian approach You use random variables to quantify **my uncertainty about observables You haven't seen yet**.

I'll argue later that the concept of probabilistic **accuracy** has two components: You want Your uncertainty assessments to be both **internally** and **externally** consistent, which corresponds to the Bayesian and frequentist ideas of **coherence/logical consistency** and **calibration**, respectively.

2.3 Exchangeability as a Bayesian concept parallel to frequentist independence

Eliciting a 400-dimensional distribution doesn't sound easy; major **simplification** is evidently needed.

In this case, and many others, this is provided by **exchangeability** considerations.

If (as in the frequentist approach) You have no relevant information that distinguishes one AMI patient from another, Your uncertainty about the 400 1s and 0s is **symmetric**, in the sense that a random permutation of the **order** in which the 1s and 0s were labeled from 1 to 400 would leave Your uncertainty about them unchanged.

de Finetti (1930, 1964) called random variables with this property **exchangeable**:

> $\{Y_i, i = 1, \ldots, n\}$ are **exchangeable** if the distributions of $(Y_1, \ldots, Y_n)$ and $(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$ are the same for all permutations $(\pi(1), \ldots, \pi(n))$.

Exchangeability (continued)

NB **Exchangeability** and **IID** are **not the same**: IID implies exchangeability, and exchangeable Y_i do have identical marginal distributions, but they're not independent (if You're expecting **a priori** about 15% 1s, say (that's the 30-day death rate for AMI with average-quality care), the knowledge that in the first 50 outcomes at the DH 20 of them were deaths would certainly change Your prediction of the 51st).

de Finetti also defined **partial** or **conditional exchangeability** (e.g., Draper et al., 1993): if, e.g., the gender X of the AMI patients were available, and if there were evidence from the medical literature that 1s tended to be noticeably more likely for men than women, then You would probably want to assume **conditional** exchangeability of the Y_i given X (meaning that the male and female 1s and 0s, viewed as separate collections of random variables, are each unconditionally exchangeable).

This is related to Fisher's (1956) idea of **recognizable subpopulations**.

The judgment of exchangeability still seems to leave the joint distribution of the Y_i quite **imprecisely specified**.

de Finetti's Theorem For Binary Outcomes

After defining the concept of exchangeability, however, de Finetti went on to prove a **remarkable result**: if You're willing to regard the $\{Y_i, i = 1, \ldots, n\}$ as part (for instance, the beginning) of an **infinite** exchangeable sequence of 1s and 0s (meaning that every finite subsequence is exchangeable), then there's a simple way to characterize Your joint predictive distribution, if it's to be **coherent** (e.g., de Finetti, 1975; Bernardo and Smith, 1994).

(**Finite** versions of the theorem have since been proven, which say that the longer the exchangeable sequence into which You're willing to embed $\{Y_i, i = 1, \ldots, n\}$, the harder it becomes to achieve coherence/logical consistency with any probability specification that's far removed from the one below.)

de Finetti's Representation Theorem. If You're willing to regard $(Y_1, \ldots, Y_n)$ as the first n terms in an infinitely exchangeable binary sequence $(Y_1, Y_2, \ldots)$; then, with $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

- $\theta = \lim_{n\to\infty} \bar{Y}_n$ must exist, and the **marginal distribution** (given θ) for each of the Y_i must be $P(Y_i = y_i|\theta) = \text{Bernoulli}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$,

de Finetti's Theorem (continued)

where P is my **joint probability distribution** on $(Y_1, Y_2, \ldots)$;

- $H(t) = \lim_{n\to\infty} P(\bar{Y}_n \le t)$, the **limiting cumulative distribution function** (CDF) of the $\bar{Y}_n$ values, must also exist for all t and must be a valid CDF, and
- $P(Y_1, \ldots, Y_n)$ can be expressed as

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, dH(\theta). \tag{22}$$

When (as will essentially always be the case in realistic applications) my joint distribution P is sufficiently regular that H possesses a **density** (with respect to Lebesgue measure), $dH(\theta) = p(\theta)\, d\theta$, (22) can be written in a more accessible way as

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \int_0^1 \theta^s (1-\theta)^{n-s} \, p(\theta) \, d\theta, \tag{23}$$

where $s = \sum_{i=1}^{n} y_i = n\,\bar{y}_n$.

Generalizing Outward from the Observables

Important digression 1. Some **awkwardness** arose above in the **frequentist** approach to modeling the AMI mortality data, because it was not clear what population $\mathcal{P}$ the data could be regarded as **like a random sample from**.

This awkwardness also arises in **Bayesian** modeling: even though in practice You're only going to observe $(y_1, \ldots, y_n)$, de Finetti's representation theorem requires You to **extend Your judgment of finite exchangeability** to the **countably-infinite collective** $(y_1, y_2, \ldots)$,

$\rightarrow$ **and this is precisely like viewing** $(y_1, \ldots, y_n)$ **as a random sample from** $\mathcal{P} = (y_1, y_2, \ldots)$.

The key point is that the **difficulty** arising from **lack of clarity** about the **scope of valid generalizability** from a given set of observational data is a **fundamental scientific problem** that emerges whenever purely **observational** data are viewed through an **inferential** or **predictive** lens, whether the statistical methods You use are **frequentist** or **Bayesian**.

The Law of Total Probability

Important digression 2. It's a general fact about **true-false propositions** D and A that

$$\begin{aligned} P(D) &= P(D \text{ and } A) + P[D \text{ and (not } A)] \\ &= P(A)\,P(D|A) + P(\text{not } A)\,P(D|\text{not } A). \end{aligned} \tag{24}$$

This is a special case of the **Law of Total Probability** (LTP).

A and (not A) divide, or **partition**, the collection of all possible outcomes into two non-overlapping (**mutually exclusive**) and **exhaustive** possibilities.

Let $A_1, \ldots, A_k$ be any **finite partition**, i.e., $P(A_i \text{ and } A_j) = 0$ (mutually exclusive) and $\sum_{i=1}^{k} P(A_i) = 1$ (exhaustive); then a **more general** version of the LTP gives that

$$\begin{aligned} P(D) &= P(D \text{ and } A_1) + \ldots + P(D \text{ and } A_k) \\ &= P(A_1)\,P(D|A_1) + \ldots + P(A_k)\,P(D|A_k) \\ &= \sum_{i=1}^{k} P(A_i)\,P(D|A_i). \end{aligned} \tag{25}$$

Hierarchical (Mixture) Modeling

There is a **continuous** version of the LTP: by analogy with (25), if X and Y are real-valued random variables

$$p(y) = \int_{-\infty}^{\infty} p(x)\, p(y|x)\, dx. \tag{26}$$

$p(x)$ in this expression can be thought of as a **mixing distribution**.

Intuitively (26) says that the overall probability behavior $p(y)$ of Y is a mixture (**weighted average**) of the conditional behavior $p(y|x)$ of Y given X, weighted by the behavior $p(x)$ of X.

Another way to put this is to say that You have a choice: You can either model the random behavior of Y **directly**, through $p(y)$, or **hierarchically**, by first modeling the random behavior of X, through $p(x)$, and then modeling the conditional behavior of Y given X, through $p(y|x)$.

Notice that X and Y are **completely general** in this discussion — in other words, given any quantity Y that You want to model stochastically, You're free to choose any X (upon which Y depends) and model Y **hierarchically** given X instead, if that's easier.

Symbolically

$$Y \quad \leftrightarrow \quad \left\{ \begin{array}{c} X \\ Y|X \end{array} \right\}. \tag{27}$$

The reason for bringing all of this up now is that (23) can be **interpreted** as follows, with θ playing the role of x:

$$\begin{aligned} p(y_1, \ldots, y_n) &= \int_0^1 p(y_1, \ldots, y_n|\theta)\, p(\theta)\, d\theta \\ &= \int_0^1 \theta^s (1-\theta)^{n-s}\, p(\theta)\, d\theta. \end{aligned} \tag{28}$$

(28) implies that in any **coherent/logically consistent** expression of uncertainty about **exchangeable** binary quantities $Y_1, \ldots, Y_n$,

$$p(y_1, \ldots, y_n|\theta) = \theta^s (1-\theta)^{n-s}. \tag{29}$$

But (a) the left side of (29), interpreted as a function of θ for fixed $y = (y_1, \ldots, y_n)$, is recognizable as the **likelihood function** for θ given y,

(b) the right side of (29) is recognizable as the likelihood function for θ in **IID Bernoulli sampling**, and (c) (29) says that these must be the **same**.

Thus, to summarize de Finetti's Theorem **intuitively**, the assumption of exchangeability in my uncertainty about binary observables $Y_1, \ldots, Y_n$ amounts to behaving **as if**

- there is a quantity called θ, interpretable as either the **long-run relative frequency of 1s** or the marginal probability that any of the Y_i is 1,
- You need to treat θ as a **random** quantity with density $p(\theta)$, and
- **conditional** on this θ the Y_i are IID Bernoulli(θ).

In yet other words, for a Bayesian whose uncertainty about binary Y_i is exchangeable, the model may effectively be taken to have the simple **mixture** or **hierarchical** representation

$$\left\{ \begin{array}{ccc} \theta & \sim & p(\theta) \\ (Y_i|\theta) & \overset{\text{IID}}{\sim} & \text{Bernoulli}(\theta),\ i = 1, \ldots, n \end{array} \right\}. \tag{30}$$

Exchangeability and Conditional Independence

This is the **link** between frequentist and Bayesian modeling of binary outcomes: exchangeability implies that You should behave like a frequentist vis à vis the **likelihood function** (taking the Y_i to be IID Bernoulli(θ)), but a frequentist who treats θ as a random variable with a **mixing distribution** $p(\theta)$.

NB This is the first example of a general fact:

$$Y_i \text{ exchangeable} \leftrightarrow \left\{ \begin{array}{c} Y_i \textbf{ conditionally IID} \\ \text{given one or more parameters} \end{array} \right\}. \tag{31}$$

So **exchangeability** is a special kind of **conditional independence**: binary exchangeable y_i are not independent, but they become conditionally independent given θ.

(30) is an example of the simplest kind of **hierarchical model (HM)**: a model at the top level for the underlying death rate θ, and then a model below that for the binary mortality indicators Y_i conditional on θ (this is a basic instance of (27): it's **not easy** to model the **predictive** distribution for $(Y_1, \ldots, Y_n)$ directly, but it becomes a lot easier when θ is introduced at the **top level of a 2–level hierarchy**).

Mixing Distribution = Prior Distribution

To emphasize an important point mentioned above, to make sense of this in the Bayesian approach **You have to treat θ as a random variable**, even though logically it's a fixed unknown constant.

This is the main conceptual difference between the Bayesian and frequentist approaches: as a Bayesian You use the **machinery** of random variables to express Your uncertainty about unknown quantities.

Approach	Fixed	Random
Frequentist	θ	Y
Bayesian	y	θ

2.4 Prior, posterior, and predictive distributions

What's the **meaning** of the mixing distribution $p(\theta)$?

$p(\theta)$ doesn't involve $y = (y_1, \ldots, y_n)$, so it must represent my information about θ external to the data set y — as noted in Part 1, it has become traditional to call it my **prior distribution** for θ (I'll address how You might go about **specifying** this distribution below).

Bayes's Theorem

Q: If $p(\theta)$ represents my information external to θ, what represents this information **after** y has been observed?

A: It has to be $p(\theta|y)$, the **conditional** distribution for θ given how y came out.

It's **conventional** (again appealing to terms involving time) to call this the **posterior distribution** for θ given y.

Q: How do You get from $p(\theta)$ to $p(\theta|y)$, i.e., how do You **update** Your information about θ in light of the data?

A: Bayes's Theorem for **continuous** quantities:

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{p(y)}. \tag{32}$$

This requires some interpreting. As a Bayesian You're **conditioning on the data**, i.e., You're thinking of the left-hand side of (32) as a function of θ for fixed y, so that must also be true of the right-hand side; thus

(a) $p(y)$ is just a constant — in fact, You can think of it as the **normalizing constant**, put into the equation to make the product $p(\theta)\, p(y|\theta)$ integrate to 1;

and (b) $p(y|\theta)$ may look like the usual frequentist **sampling distribution** for y given θ (Bernoulli, in this case), but You have to think of it as a function of θ for fixed y.

We've already encountered this idea (page 17): $l(\theta|y) = c\,p(y|\theta)$ is the **likelihood function**.

So **Bayes's Theorem** becomes

$$\begin{aligned} p(\theta|y) &= c \cdot p(\theta) \cdot l(\theta|y) \qquad (33) \\ \text{posterior} &= \left(\begin{array}{c}\text{normalizing}\\ \text{constant}\end{array}\right) \cdot \text{prior} \cdot \text{likelihood}\ . \end{aligned}$$

You can also readily construct **predictive distributions** for the y_i before they're observed, or for future y_i once some of them are known.

For example, by the LTP, the **posterior predictive distribution** for $(y_{m+1}, \ldots, y_n)$ given $(y_1, \ldots, y_m)$ is

Predictive Distributions (continued)

$$p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m) = \tag{34}$$
$$\int_0^1 p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)\, p(\theta | y_1, \ldots, y_m)\, d\theta.$$

Consider $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$: if You **knew** θ, the information $y_1, \ldots, y_m$ about how the first m of the y_i came out would be **irrelevant** (imagine predicting the results of IID coin-tossing: if You somehow **knew** that the coin was perfectly fair, i.e., that $\theta = 0.5$, then getting (say) 6 heads in the first 10 tosses would be useless to You in quantifying the likely behavior of the next (say) 20 tosses — You'd just use the **known true value** of θ).

Thus $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$ is just $p(y_{m+1}, \ldots, y_n | \theta)$, which in turn is just the **sampling distribution** under IID B(θ) sampling for the binary observables $y_{m+1}, \ldots, y_n$, namely $\prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$.

And finally $p(\theta | y_1, \ldots, y_m)$ is recognizable as just the **posterior distribution** for θ given the first m of the binary outcomes.

Putting this all together gives

$$p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m) = \tag{35}$$
$$= \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, p(\theta | y_1, \ldots, y_m) \, d\theta$$

(we can't compute (35) yet because $p(\theta|y_1, \ldots, y_m)$ depends on $p(\theta)$, which we haven't **specified** so far).

This also brings up a key difference between a **parameter** like θ on the one hand and the Y_i, before You've observed any data, on the other: parameters are inherently **unobservable**.

This makes it harder to evaluate the **quality** of my uncertainty assessments about θ than to do so about the **observable** y_i: to see how well You're doing in predicting observables You can just compare Your predictive distributions for them with how they actually turn out, but of course this isn't possible with parameters like θ **that You'll never actually see**.

The de Finetti approach to modeling emphasizes the **prediction** of observables as a valuable adjunct to **inference** about unobservable parameters, for at least two reasons:

The Value of Predictive Thinking

- Key scientific questions are often **predictive** in nature: e.g., rather than asking "Is drug A better than B (on average across many patients) for lowering blood pressure?" (inference) the ultimate question is "How much more will drug A lower **this patient's** blood pressure than drug B?" (prediction); and
- Good **diagnostic checking** is predictive: An inference about an unobservable parameter can never be directly verified, but often You can reasonably conclude that inferences about the parameters of a model that produces poor predictions of observables are also **suspect**.

With the predictive approach parameters diminish in importance, especially those that have no physical meaning — such parameters (unlike θ above) are just **place-holders for a particular kind of uncertainty on my way to making good predictions**.

It's arguable (e.g., Draper, 1995) that the discipline of statistics, and particularly its applications in the social sciences, would be improved by a **greater emphasis on predictive feedback**.

Where Does the Prior Come From?

This is not to say that parametric thinking should be **abolished**.

As the calculations on the previous pages emphasized, parameters play an important simplifying role in forming modeling judgments: the single strongest simplifier of a joint distribution is **independence** of its components, and whereas, e.g., in the mortality example the Y_i are not themselves independent, they become so conditional on θ.

---

de Finetti's Theorem for 0–1 outcomes says informally that if You're trying to make **coherent/logically consistent** probability judgments about a series of 1s and 0s that You judge exchangeable, You may as well behave like a frequentist — IID Bernoulli(θ) — with a prior distribution $p(\theta)$; but **where does the prior come from**?

NB **Coherence/logical consistency** doesn't help in answering this question — it turns out that **any** prior $p(\theta)$ could be part of **somebody**'s coherent/logically consistent probability judgments.

Some people regard the need to answer this question in the Bayesian approach as a **drawback**, but it seems to me to be a **positive feature**, as follows.

Predictive Calibration

From Bayes's Theorem the prior is supposed to be a summary of what You know (and don't know) about θ external to the data set $(y_1, \ldots, y_n)$: from previous datasets of which You're aware, from the relevant literature, from expert opinion, ... from all "good" source(s), if any exist.

Such information is almost always present, and should presumably be used when available; the issue is how to do so "well."

> The goal is evidently to choose a prior that You'll **retrospectively** be **proud of**, in the sense that Your predictive distributions for the observables (a) are well-centered near the actual values and (b) have uncertainty bands that correspond well to the realized discrepancies between actual and predicted values; this is a form of **calibration** of Your probability assessments.

There is **no guaranteed way to do this**, just as there is no guaranteed way to arrive at a "good" frequentist model (see "Where does the likelihood come from?" below).

Some general comments on arriving at a **"good"** prior:

Choosing a "Good" Prior

- There is a growing literature on methodology for **elicitation** of prior information (e.g., Kadane et al., 1980; Craig et al., 1997; Kadane and Wolfson, 1997; O'Hagan, 1997), which brings together ideas from statistics and perceptual psychology (e.g., people turn out to be better at estimating **percentiles** of a distribution than they are at estimating **standard deviations** (SDs)).
- Bayes's Theorem on the **log scale** says (apart from the normalizing constant)

$$\log(\text{posterior}) = \log(\text{prior}) + \log(\text{likelihood}), \tag{36}$$

  i.e., (posterior information) = (data information) + (prior information). This means that **close attention should be paid to the information content of the prior** by, e.g., density-normalizing the likelihood and plotting it on the same scale as the prior: it's possible for small n for the **prior to swamp the data**, and in general You shouldn't let this happen without a good reason for doing so.

  Comfort can also be taken from the other side of this coin: with large n (in

many situations, at least) the **data will swamp the prior**, and specification errors become less important.

- When You notice You're quite uncertain about how to specify the prior, You can try **sensitivity** or **pre-posterior analysis**: exploring the mapping from prior to posterior, before the data are gathered, by (a) generating some possible values for the observables, (b) writing down several plausible forms for the prior, and (c) carrying these forward to posterior distributions — if the resulting distributions are similar (i.e., if "all reasonable roads lead to Rome"), You've uncovered a useful form of stability in Your results; if not You can try to capture the prior uncertainty **hierarchically**, by, e.g., adding another layer to a model like (30) above.
- Calibration can be estimated by a form of **cross-validation:** with a given prior You can (a) repeatedly divide the data at random into modeling and validation subsets, (b) update to posterior predictive distributions based on the modeling data, and (c) compare these distributions with the actual values in the validation data.

Note that calibration is **inherently frequentist** in spirit (e.g., "What percentage of the time do my 95% predictive intervals include the actual value?").

This leads to a useful **synthesis** of Bayesian and frequentist thinking:

> **Coherence/logical consistency** keeps me internally honest; **calibration** keeps me in good contact with the world.

2.6 Conjugate analysis; comparison with frequentist modeling

Example: Prior specification in the **AMI mortality case study**. Let's say

(a) I know (from the literature) that the 30-day AMI **mortality rate** given average care and average sickness at admission in the U.S. is about **15%**,

(b) I know **little** about **care** or **patient sickness** at the DH, but

(c) I'd be somewhat surprised if the "underlying rate" at the DH was much less than **5%** or more than **30%** (note the asymmetry).

To quantify these judgments I seek a **flexible family of densities** on $(0, 1)$,

The Beta Family of Densities on $(0, 1)$

one of whose members has mean **0.15** and (say)
95% central interval (0.05,0.30).

A convenient family for this purpose is the **Beta** distributions,

$$\text{Beta}(\theta|\alpha, \beta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{37}$$

defined for $(\alpha > 0, \beta > 0)$ and for $0 < \theta < 1$.

We can use `Maple` to evaluate the **normalizing constant** c.

```
sauternes 189> maple
    |\^/|     Maple 9.5 (SUN SPARC SOLARIS)
._|\|   |/|_. Copyright (c) Maplesoft, Waterloo Maple Inc. 2004
 \  MAPLE  /  All rights reserved. Maple is a trademark of
 <____ ____>  Waterloo Maple Inc.
      |       Type ? for help.

> assume( alpha > 0, beta > 0, theta > 0, theta < 1 );
> p1 := ( theta, alpha, beta ) -> theta^( alpha - 1 ) *
    ( 1 - theta )^( beta - 1 );
```

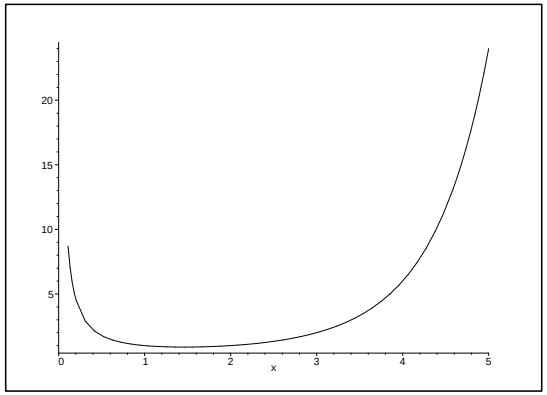
The Beta Distribution

```
                                                (alpha - 1)            (beta - 1)
     p1 := (theta, alpha, beta) -> theta                          (1 - theta)

> integrate( p1( theta, alpha, beta ), theta = 0 .. 1 );
                                  Beta(alpha~, beta~)
> help( Beta );
Beta - The Beta function
Calling Sequence:
     Beta( x, y )
Parameters:
     x - an expression
     y - an expression
Description:
- The Beta function is defined as follows:
      Beta( x, y ) = ( GAMMA( x ) * GAMMA( y ) ) / GAMMA( x + y )

> help( GAMMA );
GAMMA - The Gamma and Incomplete Gamma Functions
lnGAMMA - The log-Gamma function
```

The Beta Distribution (continued)

```
Calling Sequence:

     GAMMA( z )
     GAMMA( a, z )
     lnGAMMA( z )
Parameters:
     z - an expression
     a - an expression
Description:

- The Gamma function is defined for Re( z ) > 0 by
      GAMMA(z) =  int( exp( -t ) * t^( z - 1 ), t = 0 .. infinity )
  and is extended to the rest of the complex plane,
  less the non-positive integers, by analytic continuation.
  GAMMA has a simple pole at each of the points z = 0, -1, -2, ... .
- For positive real arguments z, the lnGAMMA function is defined by:
      lnGAMMA( z ) = ln( GAMMA( z ) )

> plotsetup( x11 );
> plot( GAMMA( x ), x = 0 .. 5, color = black );
```

The Beta Distribution (continued)



It turns out that $\Gamma(1) = 1, \Gamma(2) = 1, \Gamma(3) = 2, \Gamma(4) = 6$, and $\Gamma(5) = 24$ — the **pattern** here is that

$$\Gamma(n) = (n-1)! \quad \text{for integer } n. \tag{38}$$

Thus the **Gamma function** is a kind of **continuous generalization** of the **factorial** function.

What all of this has shown is that the **normalizing constant** in the Beta distribution is

$$c = \left[\int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} \, d\theta \right]^{-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}, \tag{39}$$

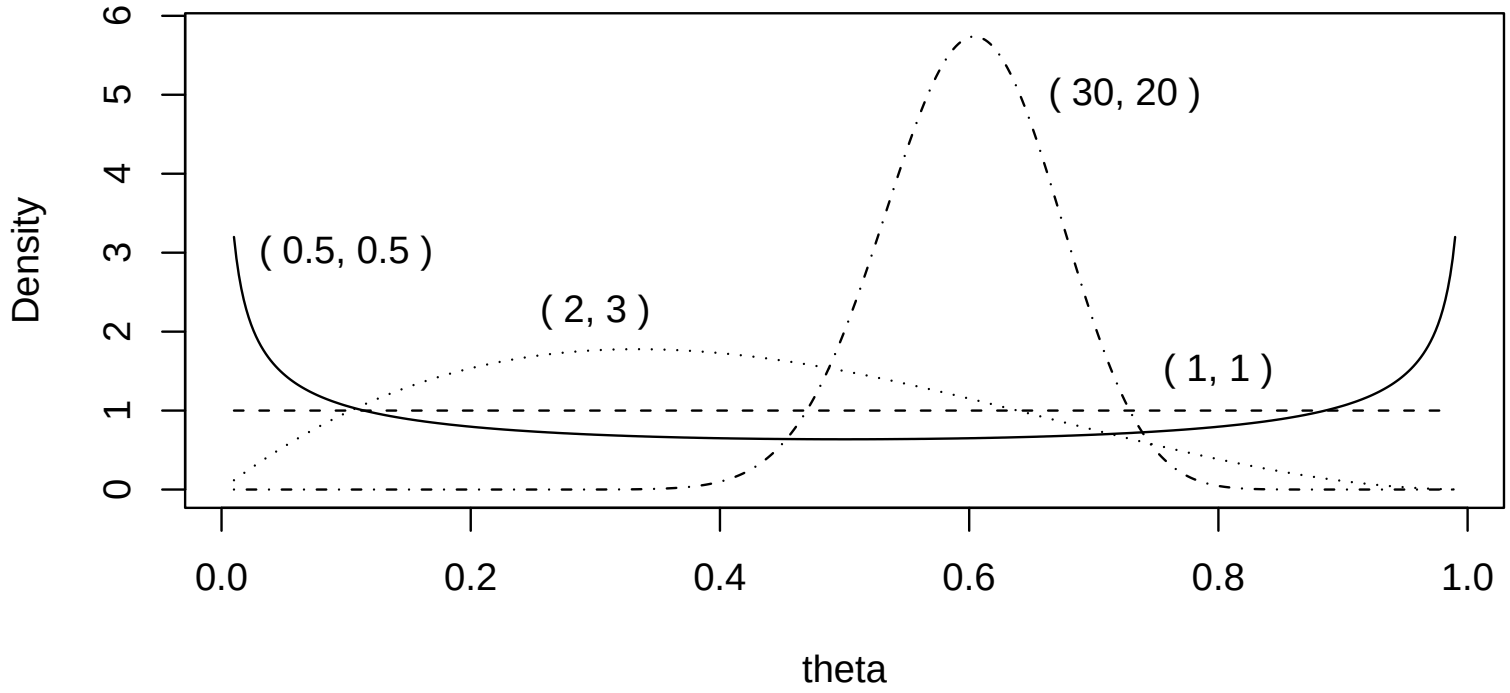
The Beta Distribution (continued)

so that the full definition of the **Beta distribution** is

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{40}$$

for $(\alpha > 0, \beta > 0)$ and for $0 < \theta < 1$.

The Beta family is **convenient** for two reasons: **(1)** It exhibits a wide variety of **distributional shapes**:



(2) As we saw above, the likelihood in this problem comes from the **Bernoulli** sampling distribution for the Y_i,

$$p(y_1, \ldots, y_n|\theta) = l(\theta|y) = \theta^s(1-\theta)^{n-s}, \tag{41}$$

where s is the **sum** of the y_i.

Now Bayes's Theorem says that to get the posterior distribution $p(\theta|y)$ You **multiply** the prior $p(\theta)$ and the likelihood — in this case $\theta^s(1-\theta)^{n-s}$ — and **renormalize** so that the product integrates to 1.

Rev. Bayes himself noticed back in the 1750s that if the prior is taken to be of the form $c\,\theta^u\,(1-\theta)^v$, the product of the prior and the likelihood **will also be of this form**, which makes the **computations** more straightforward.

The Beta family is said to be **conjugate** to the Bernoulli/Binomial likelihood.

Conjugacy of a family of **prior** distributions to a given **likelihood** is a bit hard to define precisely, but the basic idea — given a particular likelihood function — is to try to find a family of prior distributions so that the **product** of members of this family with the likelihood function will also be in the family.

The Beta Family (continued)

Conjugate analysis — finding conjugate priors for standard likelihoods and restricting attention to them on tractability grounds — is one of only two fairly general methods for getting closed-form answers in the Bayesian approach (the other is **asymptotic analysis**; see, e.g., Bernardo and Smith, 1994).

Suppose I restrict attention (for now) to members of the Beta family in trying to specify a **prior distribution** for θ in the AMI mortality example.

I want a member of this family that has **mean 0.15** and **95% central interval (0.05, 0.30)**.

```
> mean := integrate( theta * p( theta, alpha, beta ), theta = 0 .. 1 );
```

$$\text{mean} := \frac{\tilde{\alpha}}{\tilde{\alpha} + \tilde{\beta}}$$

```
> variance := simplify( integrate( ( theta - alpha /
    ( alpha + beta ) )^2 * p( theta, alpha, beta ), theta = 0 .. 1 ) );
```

Conjugate Analysis (continued)

```
                                   alpha~ beta~
      variance := ---------------------------------------
                                         2
                      (alpha~ + beta~)  (alpha~ + beta~ + 1)
```

As `Maple` has **demonstrated**, if $\theta \sim \text{Beta}(\alpha, \beta)$

$$E(\theta) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad V(\theta) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}. \tag{42}$$

```
> solve( mean = 15 / 100, beta );

                          17/3 alpha~

> solve( integrate( p( theta, alpha, 17 * alpha / 3 ),
    theta = 0.05 .. 0.30 ) = 0.95, alpha );

bytes used=3005456, alloc=1834672, time=0.82
bytes used=4006628, alloc=2293340, time=1.18
bytes used=5007408, alloc=2489912, time=1.58
>
```

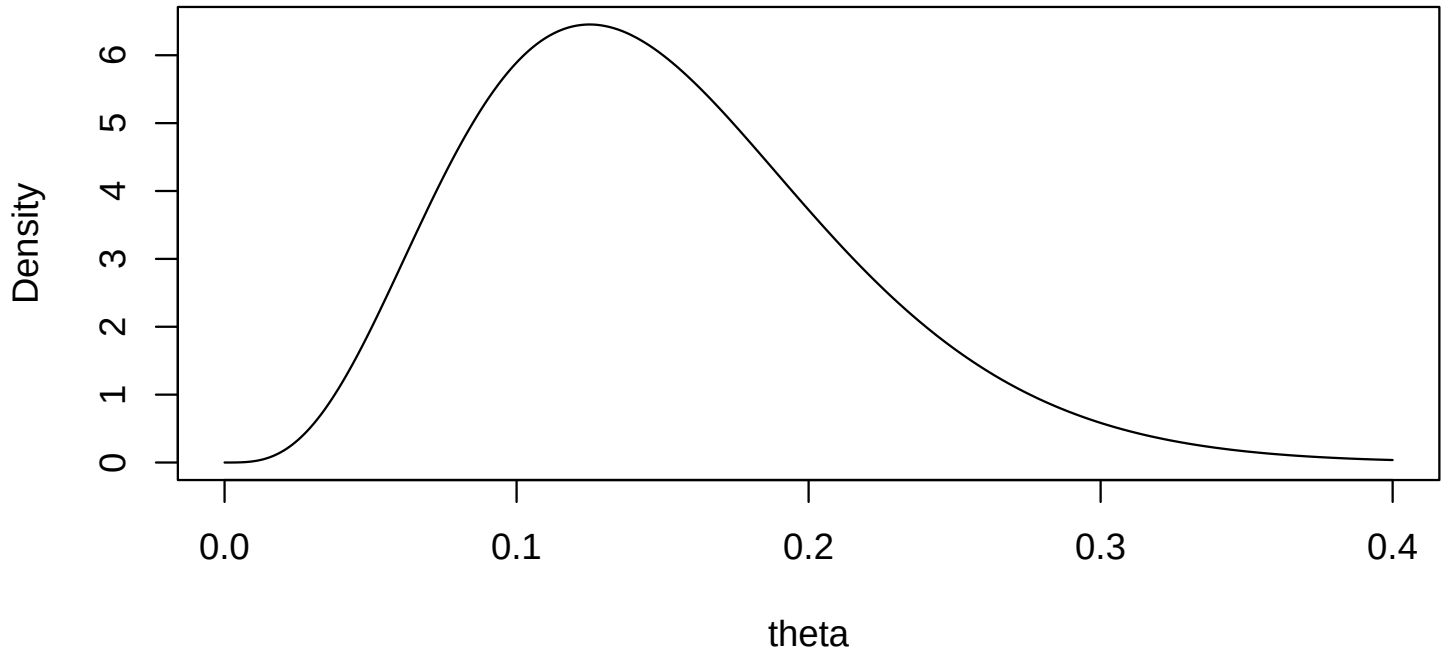
Conjugate Analysis (continued)

`Maple` can't solve this equation **symbolically** (and neither could you), but it can do so **numerically**; note how easy this is to do in `Maple`, by replacing `solve` with `fsolve`:

```
> fsolve( integrate( p( theta, alpha, 17 * alpha / 3 ),
    theta = 0.05 .. 0.30 ) = 0.95, alpha );

bytes used=7083468, alloc=2686484, time=2.50

  (output suppressed)

bytes used=27099104, alloc=3538296, time=11.99

                                  4.506062414

> 17 * 4.506062414 / 3;

                                  25.53435368
```

Conjugate Analysis (continued)

Thus the **Beta distribution** with $(\alpha, \beta) =$ **(4.5, 25.5)** meets my two prior specifications; it can readily be plotted in `Maple` or `R` (my choice here):



This prior distribution looks just like I want it to: it has a **long right-hand tail** and is **quite spread out**: the prior SD with this choice of (α, β) is $\sqrt{\frac{(4.5)(25.5)}{(4.5+25.5)^2(4.5+25.5+1)}} \doteq 0.064$, i.e., my prior says that I think the underlying AMI mortality rate at the DH is around **15%**, give or take about **6 or 7%**.

Hierarchical Model Expansion

In the usual jargon α and β are called **hyperparameters** since they're parameters of the prior distribution.

Written **hierarchically** the model we've arrived at is

$$\begin{array}{rcll} (\alpha, \beta) & = & (4.5, 25.5) & \text{(hyperparameters)} \\ (\theta|\alpha, \beta) & \sim & \text{Beta}(\alpha, \beta) & \text{(prior)} \\ (Y_1, \ldots, Y_n|\theta) & \overset{\text{IID}}{\sim} & \text{Bernoulli}(\theta) & \text{(likelihood)} \end{array} \tag{43}$$

(43) suggests what to do if You're not sure about the specifications that led to $(\alpha, \beta) = (4.5, 25.5)$: **hierarchically expand** the model by placing a distribution on (α, β) centered at $(4.5, 25.5)$.

This is an important Bayesian modeling tool: if the model is inadequate in some way, **expand it hierarchically** in directions suggested by the nature of its inadequacy (I'll give more examples of this later).

Q: Doesn't this set up the possibility of an **infinite regress**, i.e., how do You know **when to stop** adding layers to the hierarchy?

Conjugate Updating

A: (1) In practice people stop when they run out of (time, money), after having made sure that the final model passes **diagnostic checks**; and comfort may be taken from the empirical fact that (2) there tends to be a kind of **diminishing returns** principle: the farther a given layer in the hierarchy is from the likelihood (data) layer, the less it tends to affect the answer.

The conjugacy of the prior leads to a **simple closed form** for the posterior here: with y as the vector of observed $Y_i, i = 1, \ldots, n$ and s as the sum of the y_i (a **sufficient statistic** for θ, as noted above, with the Bernoulli likelihood),

$$\begin{aligned} p(\theta|y,\alpha,\beta) &= c\, l(\theta|y)\, p(\theta|\alpha,\beta) \\ &= c\, \theta^s\, (1-\theta)^{n-s}\, \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= c\, \theta^{(s+\alpha)-1} (1-\theta)^{(n-s+\beta)-1}, \end{aligned} \tag{44}$$

i.e., the **posterior** for θ is Beta$(\alpha + s, \beta + n - s)$.

This gives the hyperparameters a useful interpretation in terms of **effective information content of the prior**: it's as if the data (Beta$(s+1, n-s+1)$) were worth $(s+1) + (n-s+1) \doteq n$ observations and the prior (Beta(α, β)) were worth $(\alpha + \beta)$ observations.

The Prior Data Set

This can be used to judge whether the prior is **more informative than intended** — here it's equivalent to $(4.5 + 25.5) = \mathbf{30}$ binary observables with a mean of 0.15.

In **Bayesian inference** the **prior information** can always be thought of as **equivalent** to a **prior data set**, in the sense that if

(a) I were to **merge** the **prior data set** with the **sample data set** and do a **likelihood analysis** on the **merged data**, and

(b) You were to do a **Bayesian analysis** with the **same prior information** and **likelihood**,

we would get the **same answers**.

Conjugate analysis has the advantage that the prior sample size can be explicitly worked out: here, for example, the **prior data set** in effect consists of $\alpha = 4.5$ 1s and $\beta = 25.5$ 0s, with **prior sample size** $n_0 = (\alpha + \beta) \doteq 30$.

Even with **non-conjugate** Bayesian analyses, thinking of the **prior information** as equivalent to a **data set** is a **valuable heuristic**.

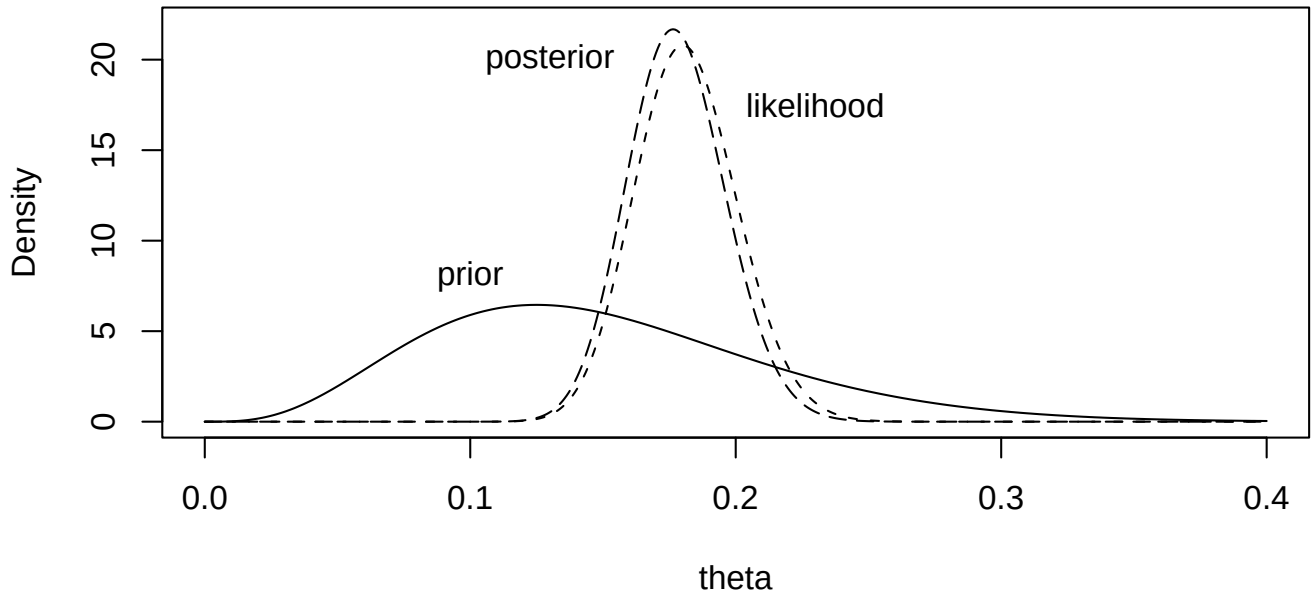
Prior-To-Posterior Updating

(44) can be **summarized** by saying

$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\theta|y) \sim \text{Beta}(\alpha + s, \beta + n - s), \qquad (45)$$

where $y = (y_1, \ldots, y_n)$ and $s = \sum_{i=1}^{n} y_i$.

Suppose the $n = 400$ **DH patients** include $s = 72$ **deaths** ($\frac{s}{n} = 0.18$).



Prior-To-Posterior Updating (continued)

Then the **prior** is Beta(4.5, 25.5), the **likelihood** is Beta(73, 329), the **posterior** for θ is Beta(76.5, 353.5), and the three densities plotted **on the same graph** are given above.

In this case the posterior and the likelihood nearly coincide, because the **data information** outweighs the **prior information** by $\frac{400}{30}$ = more than 13 to 1.

The mean of a Beta(α, β) distribution is $\frac{\alpha}{\alpha+\beta}$; with this in mind the posterior mean has an intuitive expression as a weighted average of the prior mean and data mean, with weights determined by the **effective sample size** of the prior, $(\alpha + \beta)$, and the **data sample size** n:

$$\frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{s}{n}\right)$$

$$\begin{array}{ccccccc} \text{posterior} & & \left(\begin{array}{c}\text{prior}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{prior}\\ \text{mean}\end{array}\right) & + & \left(\begin{array}{c}\text{data}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{data}\\ \text{mean}\end{array}\right) \\ \text{mean} & = & & & & & \\ .178 & = & (.070) & (.15) & + & (.93) & (.18) \end{array}$$

Comparison With Frequentist Modeling

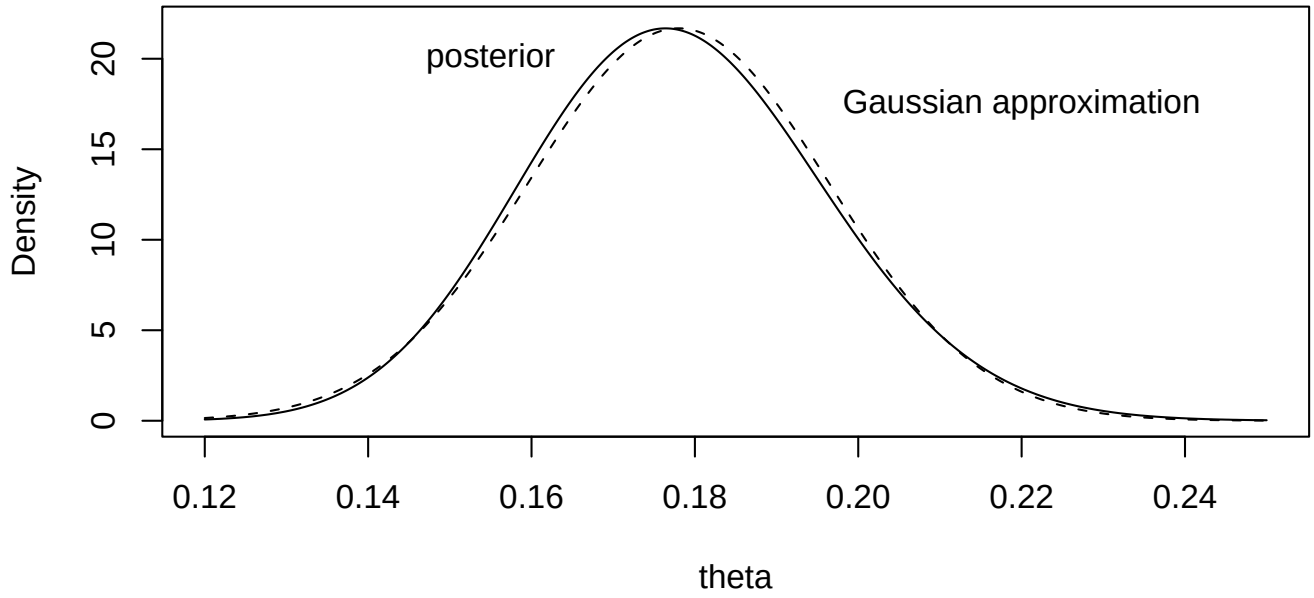
Another way to put this is that the data mean, $\bar{y} = \frac{s}{n} = \frac{72}{400} = .18$, has been **shrunken** toward the prior mean .15 by (in this case) a modest amount: the posterior mean is about .178, and the **shrinkage factor** is $\frac{30}{30+400}$ = about .07.

Comparison with frequentist modeling. To analyze these data as a frequentist You would appeal to the **Central Limit Theorem:** $n = 400$ is big enough so that the repeated-sampling distribution of $\bar{Y}$ is approximately $\mathrm{N}\left[\theta, \frac{\theta(1-\theta)}{n}\right]$, so (as we saw earlier) an approximate **95% confidence interval** for θ would be centered at $\hat{\theta} = \bar{y} = 0.18$, with an estimated standard error of $\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.0192$, and would run roughly from 0.142 to 0.218.

By contrast the posterior for θ is also **approximately Gaussian** (see the graph on the next page), with a mean of 0.178 and an SD of $\sqrt{\frac{\alpha^*\beta^*}{(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1)}} = 0.0184$, where α^* and β^* are the parameters of the Beta posterior distribution; a **95% central posterior interval** for θ would then run from about $0.178 - (1.96)(0.0184) = 0.142$ to $0.178 + (1.96)(0.0184) = 0.215$.

The two approaches (frequentist based only on the sample, Bayesian based on the sample and the prior You're using) give **almost the same** answers in this

Comparison With Frequentist Modeling (continued)



case, a result that's typical of situations with fairly large n and relatively **diffuse** prior information.

Note, however, that the **interpretation** of the two analyses differs:

- In the frequentist approach θ **is fixed but unknown and** $\bar{Y}$ **is random**, with the analysis based on imagining what would happen if the hypothetical random sampling were repeated, and appealing to the fact that across these repetitions $(\bar{Y} - \theta) \dot{\sim} \text{Gaussian}(0, .019^2)$; whereas

Comparison With Frequentist Modeling (continued)

- In the Bayesian approach $\bar{y}$ **is fixed at its observed value and** θ **is treated as random**, as a means of quantifying my posterior uncertainty about it: $(\theta - \bar{y}|\bar{y}) \dot{\sim} \text{Gaussian}(0, .018^2)$.

This means among other things that, while it's **not legitimate** with the frequentist approach to say that $P_F(.14 \leq \theta \leq .22) \doteq .95$, which is what many users of confidence intervals would like them to mean, the corresponding statement $P_B(.14 \leq \theta \leq .22|y, \text{diffuse prior information}) \doteq .95$ is a **natural consequence** of the Bayesian approach.

In the case of diffuse prior information and large n this justifies the fairly common informal practice of **computing inferential summaries in a frequentist way and then interpreting them in a Bayesian way**.

When **non-diffuse** prior information is available and You use it, Your answer will **differ** from a frequentist analysis based on the same likelihood.

If Your prior is retrospectively seen to have been **well-calibrated** You'll get a **better** answer than with the frequentist approach; if poorly calibrated, a **worse** answer (Samaniego and Reneau, 1994):

Comparison With Frequentist Modeling (continued)

"bad" Bayesian $\leq$ frequentist $\leq$ "good" Bayesian

What You make of this depends on Your **risk-aversion**: Is it better to try to land on the right in this box, running some risk of landing on the left, or to steer a middle course?

(**NB** I'll give several examples later in which a Bayesian analysis is better **even with diffuse prior information**: the point is that **likelihood methods don't always have good repeated-sampling properties with small samples**, and the **Bayesian approach** can **remedy** this problem.)

Bernoulli prediction. The **predictive distribution** for future Y_i in the Bernoulli model was shown back on page 45 (equation (35)) to be

$$p(Y_{m+1} = y_{m+1}, \ldots, Y_n = y_n | y_1, \ldots, y_m) = \tag{46}$$
$$= \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, p(\theta | y_1, \ldots, y_m) \, d\theta \ .$$

Bernoulli Prediction (continued)

It became clear earlier that if the **prior** is taken to be Beta(α, β) the **posterior** $p(\theta|y_1, \ldots, y_m)$ in this expression is Beta(α^*, β^*), where $\alpha^* = \alpha + s$ and $\beta^* = \beta + (n - s)$.

As an example of an **explicit calculation** with (46) in this case, suppose that You've observed n of the Y_i, obtaining data vector $y = (y_1, \ldots, y_n)$, and You want to predict Y_{n+1}.

Obviously $p(Y_{n+1} = y_{n+1}|y)$ has to be a **Bernoulli**(θ^*) distribution for some θ^*, and intuition says that θ^* should just be the **mean** $\frac{\alpha^*}{\alpha^*+\beta^*}$ of the posterior distribution for θ given y.

(46) in this case gives for $p(Y_{n+1} = y_{n+1}|y)$ the expression

$$\int_0^1 \theta^{y_{n+1}} (1-\theta)^{1-y_{n+1}} \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \theta^{\alpha^*-1} (1-\theta)^{\beta^*-1}\, d\theta$$

$$= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \int_0^1 \theta^{\alpha^*+y_{n+1}-1} (1-\theta)^{(\beta^*-y_{n+1}+1)-1}\, d\theta, \qquad (47)$$

and a **symbolic computing package** such as `Maple` (or examination of the

Bernoulli Prediction (continued)

logic leading to the **normalizing constant** of the **Beta distribution**) then yields that $p(Y_{n+1} = y_{n+1}|y)$ is

$$\left[\frac{\Gamma(\alpha^* + y_{n+1})}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\beta^* - y_{n+1} + 1)}{\Gamma(\beta^*)}\right]\left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right]. \tag{48}$$

Recalling that $\frac{\Gamma(x+1)}{\Gamma(x)} = x$ for any real number x leads to **simple expressions** that **match intuition**; in the case $y_{n+1} = 1$, for instance, (48) becomes

$$\begin{aligned} p(Y_{n+1} = 1|y) &= \left[\frac{\Gamma(\alpha^* + 1)}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right] \\ &= \frac{\alpha^*}{\alpha^* + \beta^*}. \end{aligned} \tag{49}$$

For example, with $(\alpha, \beta) = (4.5, 25.5)$ and $n = 400$ with $s = 72$, we saw earlier that the **posterior** for θ was Beta$(76.5, 353.5)$, and this posterior distribution has mean $\frac{\alpha^*}{\alpha^* + \beta^*} = 0.178$.

In this situation You would expect the next AMI patient who comes along to die within 30 days of admission with probability **0.178**, so the predictive distribution above **makes good sense**.

The Binomial Distribution

It became clear above that the **sum** $s = \sum_{i=1}^{n} y_i$ of the 1s and 0s is a **sufficient statistic** for θ with the Bernoulli likelihood.

This means that if You buy into the model $(Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$, **You don't care** whether You observe the entire data vector $Y = (Y_1, \ldots, Y_n)$ or its sum $S = \sum_{i=1}^{n} Y_i$.

The distribution of S in repeated sampling has a **familiar form**: it's just the **binomial** distribution $\text{Binomial}(n, \theta)$, which counts the number of successes in a series of IID success/failure trials.

Recall that if $S \sim \text{Binomial}(n, \theta)$ then S has **discrete density**

$$p(S = s|\theta) = \left\{ \begin{array}{cc} \left(\begin{array}{c} n \\ s \end{array} \right) \theta^s (1-\theta)^{n-s} & \text{if } s = 0, \ldots, n \\ 0 & \text{otherwise} \end{array} \right\}.$$

This gives **another conjugate updating rule** in simple Bayesian modeling for free: if the data set just consists of a single draw S from a binomial distribution, then the conjugate prior for the success probability θ is $\text{Beta}(\alpha, \beta)$, and the updating rule, which follows directly from (45), is

Two Important General Points

$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (S|\theta) \sim \text{Binomial}(n, \theta) \end{array} \right\} \rightarrow (\theta|s) \sim \text{Beta}(\alpha + s, \beta + n - s). \qquad (50)$$

---

1 (the **sequential** nature of **Bayesian learning**) Suppose You and I are observing data $(y_1, \ldots, y_n)$ to **learn** about a **parameter** θ, and we have no reason throughout this observation process to **change** (the sampling distribution/likelihood part of) our **model**.

We both start with the **same prior** $p_1(\theta)$ before any of the data arrive, but we adopt what appear to be **different analytic strategies**:

- You wait until the whole data set $(y_1, \ldots, y_n)$ has been observed and **update** $p_1(\theta)$ **directly** to the posterior distribution $p(\theta|y_1, \ldots, y_n)$, whereas
- I **stop** after seeing $(y_1, \ldots, y_m)$ for some $m < n$, update $p_1(\theta)$ to an **intermediate** posterior distribution $p(\theta|y_1, \ldots, y_m)$, and then I go on from there, observing $(y_{m+1}, \ldots, y_n)$ and finally updating to a posterior on θ that takes account of the **whole data set** $(y_1, \ldots, y_n)$.

Two Important General Points (continued)

Q_1 What should I use for my **intermediate prior distribution** $p_2(\theta)$?

A_1 Naturally enough, the **right thing to do** is to set $p_2(\theta) = p(\theta|y_1, \ldots, y_m)$.

The informal way people refer to this is to say that **yesterday's posterior distribution is today's prior distribution**.

Q_2 If I use the posterior in **A_1**, do You and I get the **same answer** for $p(\theta|y_1, \ldots, y_n)$ in the end?

A_2 **Yes** (You can check this).

2 (the generality of **conjugate analysis**) Having seen **conjugate priors** used with binary outcomes, it's clear that **conjugate analysis** has a variety of **advantages**:

- It's **mathematically straightforward**;
- The **posterior mean** turns out to be a **weighted average** of the **prior** and **data means**; and
- The **prior** is nicely **interpretable** as an information source that's **equivalent to a data set**, and it's easy to figure out the **prior sample size**.

Two Important General Points (continued)

It's natural to wonder, though, what's **lost** in addition to what's **gained** by adopting a conjugate prior.

The main **disadvantage** of conjugate priors is that in their simplest form they're **not flexible enough** to express **all possible forms** of prior information.

For example, in the AMI mortality case study, what if You wanted to combine a **bimodal** prior distribution with the Bernoulli likelihood?

This isn't possible when using a **single member** of the Beta(α, β) family.

However, it's possible to **prove** the following:

Theorem (Diaconis and Ylvisaker 1985). Given a likelihood that's a member of the **exponential family** (more about this later), any prior distribution can be expressed as a **mixture** of priors that are conjugate to that likelihood.

For example, in the **AMI case study** the model could be

$$\begin{aligned} J &\sim p(J) \\ (\theta|J) &\sim \text{Beta}(\alpha_J, \beta_J) \end{aligned} \tag{51}$$

2.8 Integer-Valued Outcomes

$$(Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta),\ i = 1, \ldots, n,$$

for some distribution $p(J)$ on the positive integers — this is **completely general** but loses some of the advantages of simple conjugate analysis (e.g., **closed-form computations** are no longer possible).

Case Study: *Hospital length of stay for birth of premature babies.* As a small part of a study I worked on at the Rand Corporation in the late 1980s, we obtained data on a random sample of $n = 14$ women who came to a hospital in Santa Monica, CA, in 1988 to **give birth to premature babies**.

One (**integer-valued**) outcome of interest was

$y =$ **length of hospital stay** (LOS).

Here's a preliminary look at the data in `R`:

```
> y
[1] 1 2 1 1 1 2 2 4 3 6 2 1 3 0
> sort( y )
[1] 0 1 1 1 1 1 2 2 2 2 3 3 4 6
```

Integer-Valued Outcomes (continued)

```
> table( y )
0 1 2 3 4 6
1 5 4 2 1 1
> stem( y, scale = 2 )
The decimal point is at the |
  0 | 0
  1 | 00000
  2 | 0000
  3 | 00
  4 | 0
  5 |
  6 | 0
> mean( y )
[1] 2.071429
> sd( y )
[1] 1.54244
> var( y )
[1] 2.37912
```

Poisson Modeling

One possible model for non-negative integer-valued outcomes is the

Poisson distribution

$$P(Y_i = y_i|\lambda) = \left\{ \begin{array}{cc} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} & \text{for } y_i = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right\}, \tag{52}$$

for some $\lambda > 0$.

As usual `Maple` can be used to work out the **mean** and **variance** of this distribution:

```
> assume( lambda > 0 );
> p := ( y, lambda ) -> lambda^y * exp( - lambda ) / y!;
                                                  y
                                          lambda  exp(-lambda)
                    p := (y, lambda) -> --------------------
                                                 y!
> simplify( sum( p( y, lambda ), y = 0 .. infinity ) );
                                          1
> simplify( sum( y * p( y, lambda ), y = 0 .. infinity ) );
                                       lambda~
```

Informal Model-Checking

```
> simplify( sum( ( y - lambda )^2 * p( y, lambda ),
    y = 0 .. infinity ) );
                                        lambda~
```

Thus if $(Y|\lambda) \sim \text{Poisson}(\lambda)$, $E(Y) = V(Y) = \lambda$, which people sometimes express by saying that the **variance-to-mean ratio** (VTMR) for the Poisson is 1.

R can be used to check informally whether the Poisson is a **good fit** to the LOS data:

```
> dpois( 0:7, mean( y ) )
[1] 0.126005645 0.261011693 0.270333539 0.186658872 0.096662630
[6] 0.040045947 0.013825386 0.004091186
> print( n <- length( y ) )
[1] 14
> table( y ) / n
         0          1          2          3          4          6
0.07142857 0.35714286 0.28571429 0.14285714 0.07142857 0.07142857
```

Informal Model-Checking (continued)

```
> cbind( c( dpois( 0:6, mean( y ) ),
    1 - sum( dpois( 0:6, mean( y ) ) ) ),
    apply( outer( y, 0:7, '==' ), 2, sum ) / n )

             [,1]       [,2]
[1,] 0.126005645 0.07142857
[2,] 0.261011693 0.35714286
[3,] 0.270333539 0.28571429
[4,] 0.186658872 0.14285714
[5,] 0.096662630 0.07142857
[6,] 0.040045947 0.00000000
[7,] 0.013825386 0.07142857
[8,] 0.005456286 0.00000000
```

The second column in the above table records the values of the **Poisson probabilities** for $\lambda = 2.07$, the mean of the y_i, and the third column is the **empirical relative frequencies**; informally the fit is reasonably good.

Another informal check comes from the fact that the sample mean and variance are 2.07 and $1.542^2 \doteq 2.38$, which are reasonably close.

Does Exchangeability → Sampling Distribution Here? (No.)

Exchangeability. As with the AMI mortality case study, before the data arrive I recognize that my uncertainty about the Y_i is exchangeable, and you would expect from a generalization of the binary-outcomes version of de Finetti's Theorem that the structure of a **plausible Bayesian model** for the data would then be

$$\begin{array}{rcll} \theta & \sim & p(\theta) & \text{(prior)} \\ (Y_i|\theta) & \overset{\text{IID}}{\sim} & F(\theta) & \text{(likelihood)}, \end{array} \tag{53}$$

where θ is some parameter (vector) and $F(\theta)$ is some **parametric family of distributions** on the non-negative integers indexed by θ.

Thus, in view of the preliminary examination of the data above, a **plausible Bayesian model** for these data is

$$\begin{array}{rcll} \lambda & \sim & p(\lambda) & \text{(prior)} \\ (Y_i|\lambda) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) & \text{(likelihood)}, \end{array} \tag{54}$$

where λ is a **positive real number**.

Model Uncertainty

NB (1) This approach to model-building involves a form of **cheating**, because we've **used the data twice**: once to choose the model, and again to draw conclusions conditional on the chosen model.

The result in general can be a failure to **assess** and **propagate** **model uncertainty** (e.g., Draper 1995).

(2) **Frequentist** modeling often employs this **same kind of cheating** in specifying the likelihood function.

(3) There are two Bayesian ways out of this dilemma: **cross-validation** and **Bayesian non-parametric/semi-parametric** methods.

The **latter** is beyond the scope of this course; I'll give examples of the **former** later.

To get more practice with Bayesian calculations I'm going to **ignore the model uncertainty problem for now** and pretend that somehow we knew that the Poisson was a good choice.

The likelihood function in model (54) is

Poisson Likelihood

$$
\begin{aligned}
l(\lambda|y) &= c\, p_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n|\lambda) \\
&= c \prod_{i=1}^{n} p_{Y_i}(y_i|\lambda) \qquad (55) \\
&= c \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = c\,\lambda^s\, e^{-n\lambda},
\end{aligned}
$$

where $y = (y_1, \ldots, y_n)$ and $s = \sum_{i=1}^n y_i$; here $\left(\prod_{i=1}^n y_i!\right)^{-1}$ can be **absorbed** into the generic positive c because it doesn't involve λ.

Thus (as was true in the Bernoulli model) $s = \sum_{i=1}^n y_i$ is **sufficient** for λ in the Poisson model, and we can write $l(\lambda|s)$ instead of $l(\lambda|y)$ if we want.

If a **conjugate** prior $p(\lambda)$ for λ exists it must be such that the product $p(\lambda)\, l(\lambda|s)$ has the same mathematical form as $p(\lambda)$.

Examination of (55) reveals that the same trick works here as with Bernoulli data, namely **taking the prior to be of the same form as the likelihood**:

$$
p(\lambda) = c\,\lambda^{\alpha-1} e^{-\beta\lambda} \qquad (56)
$$

The Gamma Distribution

for some $\alpha > 0, \beta > 0$ — this is the **Gamma distribution** $\lambda \sim \Gamma(\alpha, \beta)$ for $\lambda > 0$ (see GCSR, Appendix A).

As usual `Maple` can work out the **normalizing constant**:

```
> assume( lambda > 0, alpha > 0, beta > 0 );
> p1 := ( lambda, alpha, beta ) -> lambda^( alpha - 1 ) *
    exp( - beta * lambda );
                                          (alpha - 1)
     p1 := (lambda, alpha, beta) -> lambda            exp(-beta lambda)
> simplify( integrate( p1( lambda, alpha, beta ),
    lambda = 0 .. infinity ) );
                                 (-alpha~)
                          beta~           GAMMA(alpha~)
```

Thus $c^{-1} = \beta^{-\alpha}\,\Gamma(\alpha)$ and the **proper definition** of the Gamma distribution is

$$\text{If } \lambda \sim \Gamma(\alpha, \beta) \text{ then } p(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\,\lambda^{\alpha-1}\,e^{-\beta\,\lambda} \tag{57}$$

for $\alpha > 0, \beta > 0$.

The R Implementation of the Gamma Distribution

As usual R can also be used to explore the behavior of this family of distributions **as a function of its inputs** α and β, but you need to watch out — there are **two different parameterizations** in common usage:

```
> help( dgamma )
GammaDist                package:stats                R Documentation
The Gamma Distribution
Description:
     Density, distribution function, quantile function and random
     generation for the Gamma distribution with parameters 'shape' and
     'scale'.
Usage:
     dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
     pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
            log.p = FALSE)
     qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE,
            log.p = FALSE)
     rgamma(n, shape, rate = 1, scale = 1/rate)
Arguments:
    x, q: vector of quantiles.
```

The Gamma Distribution in R (continued)

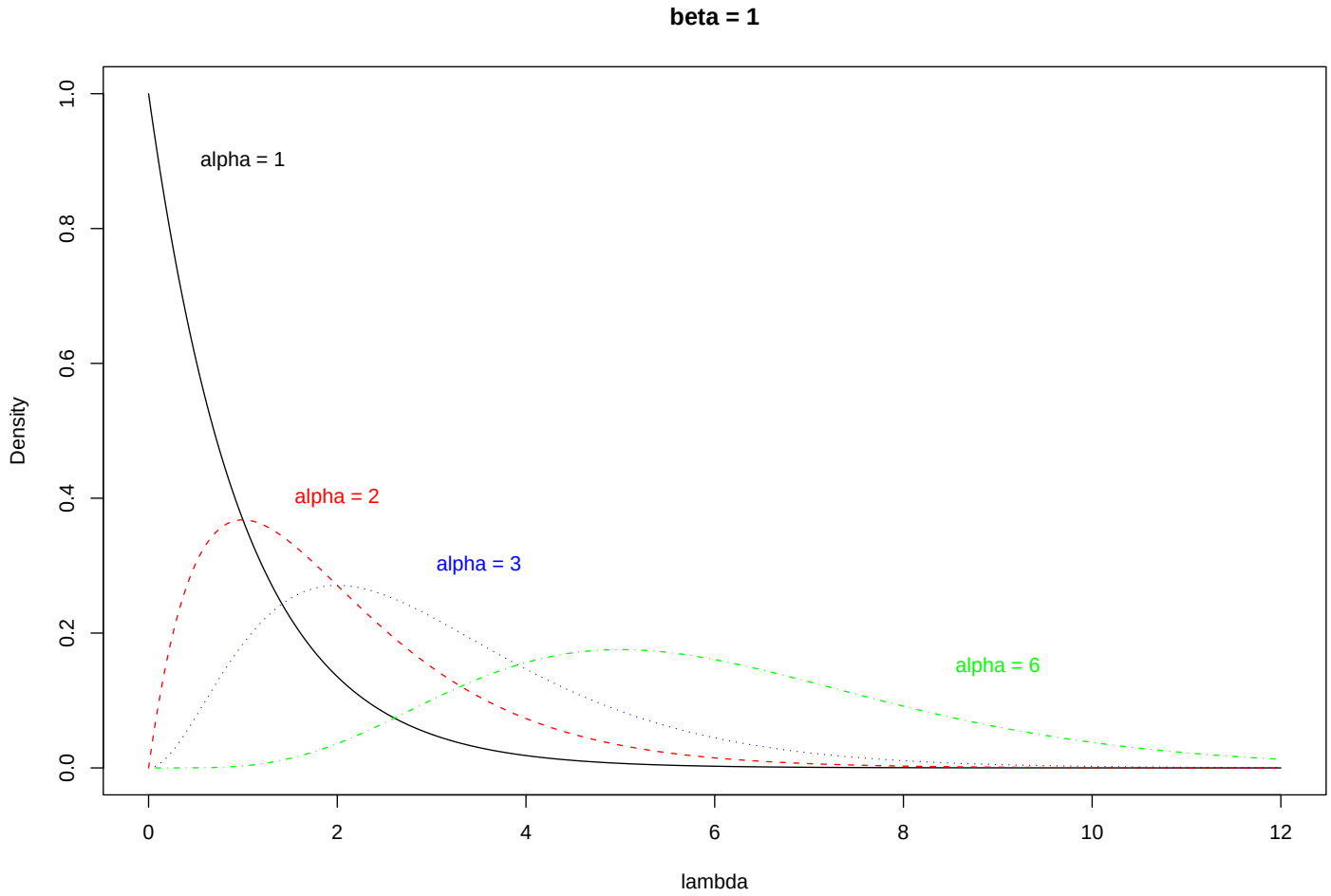
```
       p: vector of probabilities.
       n: number of observations. If 'length(n) > 1', the length is
          taken to be the number required.
    rate: an alternative way to specify the scale.
shape, scale: shape and scale parameters.  Must be positive, 'scale'
          strictly.
log, log.p: logical; if 'TRUE', probabilities/densities p are returned
          as log(p).
lower.tail: logical; if TRUE (default), probabilities are P[X <= x],
          otherwise, P[X > x].
Details:
     If 'scale' is omitted, it assumes the default value of '1'.
     The Gamma distribution with parameters 'shape' = a and 'scale' = s
     has density
                f(x)= 1/(s^a Gamma(a)) x^(a-1) e^-(x/s)
     for x >= 0, a > 0 and s > 0. (Here Gamma(a) is the function
     implemented by R's 'gamma()' and defined in its help.  Note that
     a=0 corresponds to the trivial distribution with all mass at point
     0.)
```

The Gamma Distribution in R (continued)

The quantity in R corresponding to our α is evidently **shape**, but notice that what R calls **scale** is $\frac{1}{\beta}$ for us; the name for β in R is rate, the reciprocal of **scale**:

```
lambda.grid.1 <- seq( 0, 12, length = 500 )
postscript( "gamma-beta-equals-1.ps" )
plot( lambda.grid.1, dgamma( lambda.grid.1, shape = 1, rate = 1 ),
  xlab = 'lambda', ylab = 'Density', type = 'l', main = 'beta = 1' )
text( 1, 0.9, 'alpha = 1' )
lines( lambda.grid.1, dgamma( lambda.grid.1, shape = 2, rate = 1 ),
  lty = 2, col = 'red' )
text( 2, 0.4, 'alpha = 2', col = 'red' )
lines( lambda.grid.1, dgamma( lambda.grid.1, shape = 3, rate = 1 ),
  lty = 3, col = 'blue' )
text( 3.5, 0.3, 'alpha = 3', col = 'blue' )
lines( lambda.grid.1, dgamma( lambda.grid.1, shape = 6, rate = 1 ),
  lty = 4, col = 'green' )
text( 9, 0.15, 'alpha = 6', col = 'green' )
dev.off( )
```

α Controls Shape in the Gamma Family



The `R` name for α is a **good choice**: α evidently controls the **shape** of the Gamma family.

What **distributional shape** does the Gamma approach as $\alpha \to \infty$?

The Exponential Distribution

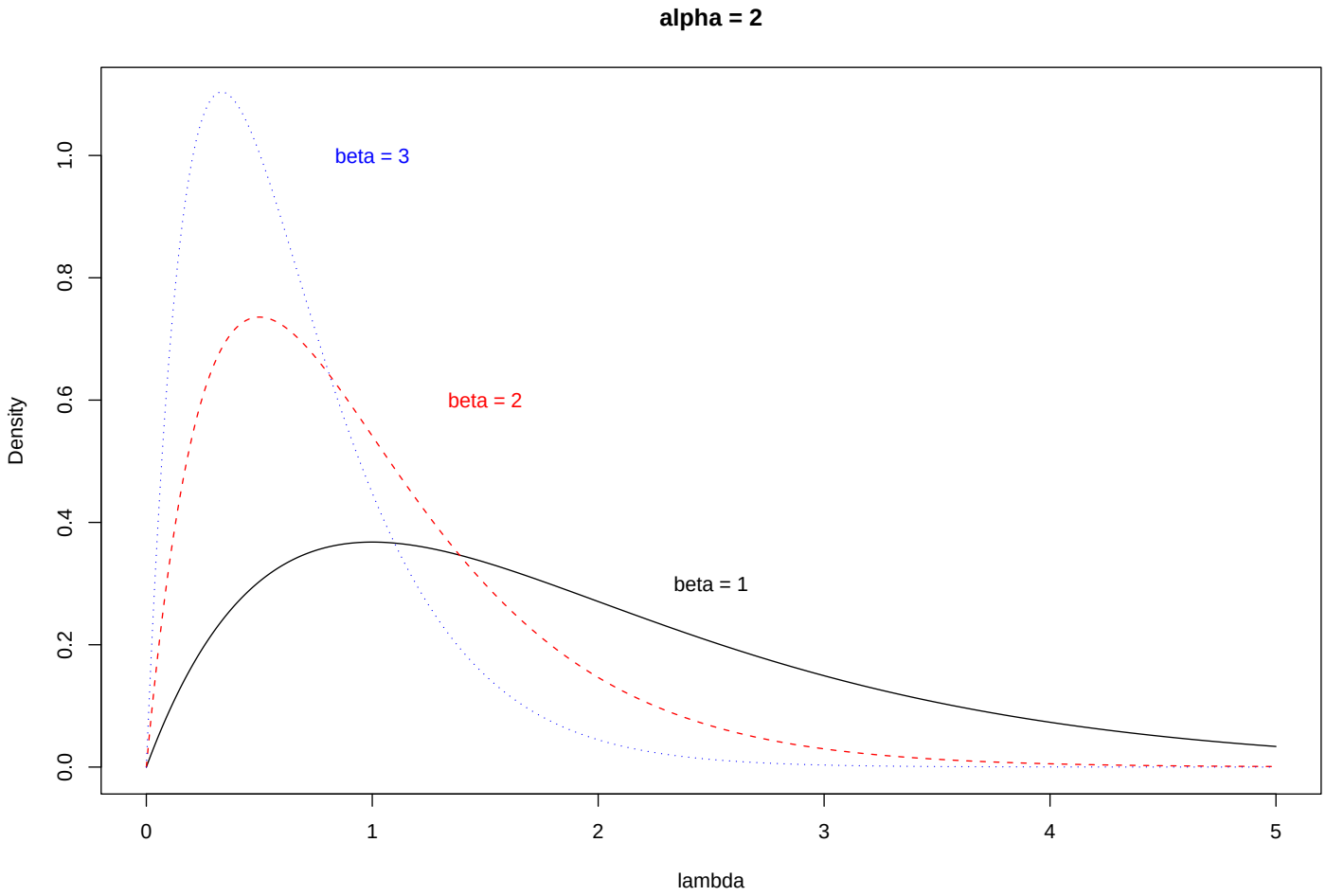
When $\alpha = 1$ the Gamma distributions have a special form that you'll probably recognize — they're the **exponential** distributions $\mathcal{E}(\beta)$: for $\beta > 0$

$$\text{If } \lambda \sim \mathcal{E}(\beta) \text{ then } p(\lambda) = \left\{ \begin{array}{cc} \beta\, e^{-\beta\,\lambda} & \text{for } \lambda > 0 \\ 0 & \text{otherwise} \end{array} \right\}. \tag{58}$$

What about the **effect** of β, or its reciprocal, on the distribution?

```
lambda.grid.2 <- seq( 0, 5, length = 500 )
plot( lambda.grid.2, dgamma( lambda.grid.2, shape = 2, rate = 1 ),
  xlab = 'lambda', ylab = 'Density', type = 'l', main = 'alpha = 2',
  ylim = c( 0, 1.1 ) )
text( 2.5, 0.3, 'beta = 1' )
lines( lambda.grid.2, dgamma( lambda.grid.2, shape = 2, rate = 2 ),
  lty = 2, col = 'red' )
text( 1.5, 0.6, 'beta = 2', col = 'red' )
lines( lambda.grid.2, dgamma( lambda.grid.2, shape = 2, rate = 3 ),
  lty = 3, col = 'blue' )
text( 1, 1, 'beta = 3', col = 'blue' )
```

β (and $\frac{1}{\beta}$) Control the Spread



In the Gamma family the parameter β controls the **spread** of the distribution, but $\frac{1}{\beta}$ controls the **scale**, in the sense that as $\frac{1}{\beta}$ increases the distribution becomes **more spread out**.

$\frac{1}{\beta}$ Is a Scale Parameter For the Gamma Distribution

Definition Given a random quantity y whose density $p(y|\sigma)$ depends on a parameter $\sigma > 0$, if it's possible to express $p(y|\sigma)$ in the form $\frac{1}{\sigma} f(\frac{y}{\sigma})$, where $f(\cdot)$ is a function which does not depend on y or σ, then σ is called a **scale parameter** for the parametric family p.

Letting $f(t) = e^{-t}$ and taking $\sigma = \frac{1}{\beta}$, You can see that the Gamma family can be expressed in this way, so $\frac{1}{\beta}$ is a **scale parameter** for the Gamma distribution.

As usual `Maple` can also work out the **mean** and **variance** of this family:

```
> p := ( lambda, alpha, beta ) -> beta^alpha * lambda^( alpha - 1 ) *
    exp( - beta * lambda ) / GAMMA( alpha );
                                      alpha       (alpha - 1)
                                beta      lambda              exp(-beta lambda)
p := (lambda, alpha, beta) -> ----------------------------------------
                                                GAMMA(alpha)
> simplify( integrate( p( lambda, alpha, beta ),
    lambda = 0 .. infinity ) );
                                        1
```

Conjugate Updating With the Poisson Likelihood

```
> simplify( integrate( lambda * p( lambda, alpha, beta ),
    lambda = 0 .. infinity ) );
                                   alpha~
                                   ------
                                   beta~
> simplify( integrate( ( lambda - alpha / beta )^2 *
    p( lambda, alpha, beta ), lambda = 0 .. infinity ) );
                                   alpha~
                                   ------
                                        2
                                   beta~
```

Thus if $\lambda \sim \Gamma(\alpha, \beta)$ then $E(\lambda) = \frac{\alpha}{\beta}$ and $V(\lambda) = \frac{\alpha}{\beta^2}$, and **conjugate updating** is now **straightforward**: with $y = (y_1, \ldots, y_n)$ and $s = \sum_{i=1}^{n} y_i$, by Bayes's Theorem

$$\begin{aligned} p(\lambda|y) &= c\, p(\lambda)\, l(\lambda|y) \\ &= c \left(c\, \lambda^{\alpha-1}\, e^{-\beta\lambda} \right) \left(c\, \lambda^{s}\, e^{-n\lambda} \right) \\ &= c\, \lambda^{(\alpha+s)-1}\, e^{-(\beta+n)\lambda}, \end{aligned} \tag{59}$$

and the **resulting distribution** is just $\Gamma(\alpha + s, \beta + n)$.

Conjugate Poisson Analysis

This can be **summarized** as follows:

$$\left\{ \begin{array}{c} (\lambda|\alpha,\beta) \sim \Gamma(\alpha,\beta) \\ (Y_i|\lambda) \overset{\text{IID}}{\sim} \text{Poisson}(\lambda), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\lambda|s) \sim \Gamma(\alpha^*, \beta^*), \tag{60}$$

where $(\alpha^*, \beta^*) = (\alpha + s, \beta + n)$ and $s = \sum_{i=1}^{n} y_i$ is a **sufficient statistic** for λ in this model.

The posterior mean of λ here is evidently $\frac{\alpha^*}{\beta^*} = \frac{\alpha+s}{\beta+n}$, and the prior and data means are $\frac{\alpha}{\beta}$ and $\bar{y} = \frac{s}{n}$, so (as was the case in the Bernoulli model) the posterior mean can be written as a **weighted average** of the prior and data means:

$$\frac{\alpha + s}{\beta + n} = \left(\frac{\beta}{\beta + n}\right)\left(\frac{\alpha}{\beta}\right) + \left(\frac{n}{\beta + n}\right)\left(\frac{s}{n}\right). \tag{61}$$

Thus the **prior sample size** n_0 in this model is just β (which makes sense given that $\frac{1}{\beta}$ is the scale parameter for the Gamma distribution), and the prior acts like a **dataset** consisting of β observations with mean $\frac{\alpha}{\beta}$.

The $\Gamma(\epsilon, \epsilon)$ Prior

LOS data analysis. Suppose that, before the current data set is scheduled to arrive, I know **little** about the mean length of hospital stay of women giving birth to premature babies.

Then for my prior on λ I'd like to specify a member of the $\Gamma(\alpha, \beta)$ family which is relatively **flat in the region in which the likelihood function is appreciable**.

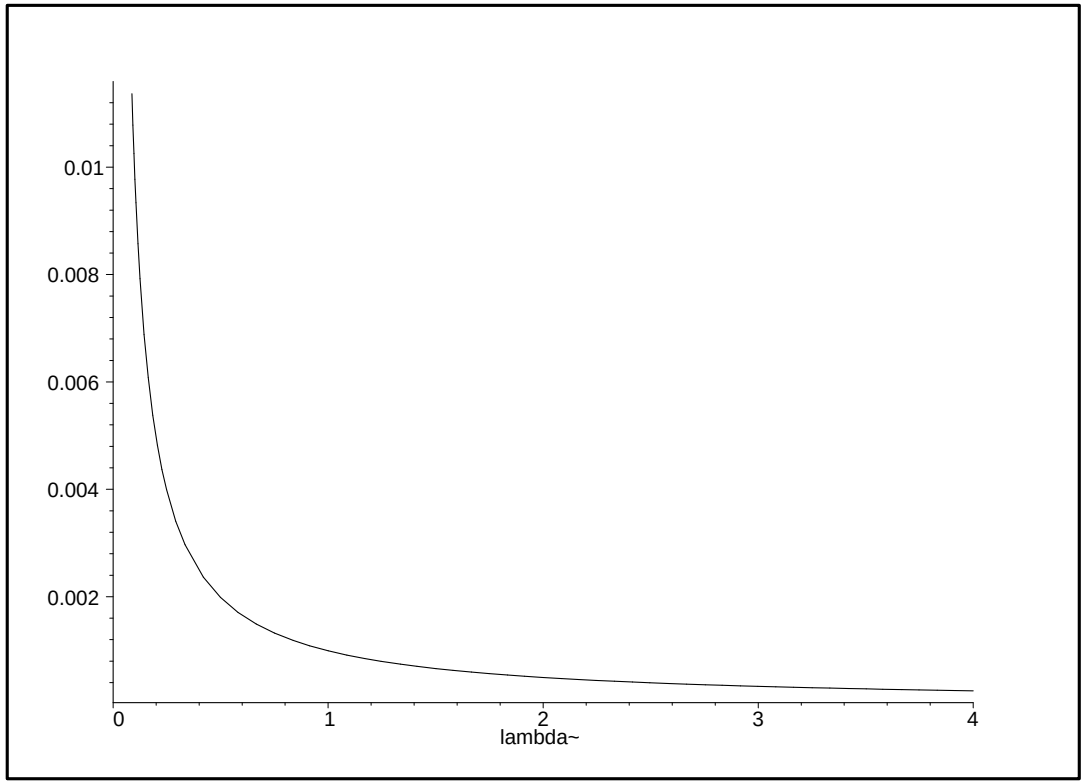
A **convenient** and **fairly all-purpose default choice** of this type is $\Gamma(\epsilon, \epsilon)$ for some small ϵ like 0.001.

When used as a prior this distribution has **prior sample size** ϵ; it also has mean 1, but that usually doesn't matter when ϵ is **tiny**.

```
> plot( p( lambda, 0.001, 0.001 ), lambda = 0 .. 4, color = black );
```

As the **graph** on the next page shows, this **distribution** is **rather flat** over the **entire region** $(1, \infty)$; it has an **unpleasant spike near 0**, but this is only a **potential problem** when the **likelihood density** is **concentrated near 0** (in that case You may be inserting **stronger prior information** than You **intended**).

The Empirical Rule



With the LOS data $s = 29$ and $n = 14$, so the **likelihood** for λ is like a $\Gamma(30, 14)$ density, which has mean $\frac{30}{14} \doteq 2.14$ and SD $\sqrt{\frac{30}{14^2}} \doteq 0.39$.

Thus by the **Empirical Rule** the likelihood is appreciable in the range $(\text{mean} \pm 3\,\text{SD}) \doteq (2.14 \pm 1.17) \doteq (1.0, 3.3)$, and you can see from the plot above that the prior is indeed **relatively flat** in this region.

LOS Data Analysis (continued)

From the **Bayesian updating** in (60), with a $\Gamma(0.001, 0.001)$ prior the **posterior** is $\Gamma(29.001, 14.001)$.

It's useful, in summarizing the **updating** from prior through likelihood to posterior, to make a table that records measures of **center** and **spread** at each point along the way.

For example, the $\Gamma(0.001, 0.001)$ **prior**, when regarded (as usual) as a **density** for λ, has mean 1.000 and SD $\sqrt{1000} \doteq 31.6$ (i.e., informally, as far as we're concerned, before the data arrive λ could be **anywhere between 0 and (say) 100**).

And the $\Gamma(29.001, 14.001)$ **posterior** has mean $\frac{29.001}{14.001} \doteq 2.071$ and SD $\sqrt{\frac{29.001}{14.001^2}} \doteq 0.385$, so after the data have arrived we know **quite a bit more than before**.

There are two main ways to summarize the **likelihood** — Fisher's approach based on **maximizing** it, and the Bayesian approach based on regarding it as a density and **integrating** over it — and it's instructive to compute them both and **compare**.

Likelihood Calculations

The **likelihood-integrating** approach (which, at least in one-parameter problems, is essentially equivalent to Fisher's (1935) attempt at **fiducial** inference) treats the $\Gamma(30, 14)$ likelihood as a density for λ, with mean $\frac{30}{14} \doteq 2.143$ and SD $\sqrt{\frac{30}{14^2}} \doteq 0.391$.

As for the **likelihood-maximizing** approach, from (55) the log likelihood function is

$$ll(\lambda|y) = ll(\lambda|s) = \log\left(c\,\lambda^s e^{-n\lambda}\right) = c + s\log\lambda - n\lambda, \tag{62}$$

and this is **maximized** as usual (check that it's the max) by setting the **derivative** equal to 0 and solving:

$$\frac{\partial}{\partial\lambda} ll(\lambda|s) = \frac{s}{\lambda} - n = 0 \quad \text{iff} \quad \lambda = \hat{\lambda}_{\text{MLE}} = \frac{s}{n} = \bar{y}. \tag{63}$$

Since the MLE $\hat{\lambda}_{\text{MLE}}$ turns out to be our old friend the **sample mean** $\bar{y}$, you might be tempted to conclude immediately that $\widehat{SE}\left(\hat{\lambda}_{\text{MLE}}\right) = \frac{\hat{\sigma}}{\sqrt{n}}$, where $\hat{\sigma} = 1.54$ is the sample SD, and indeed it's true in repeated sampling that $V\left(\bar{Y}\right) = \frac{V(Y_1)}{n}$; but the **Poisson distribution** has variance $V(Y_1) = \lambda$, so that

Calibrating the MLE

$\sqrt{V(\bar{Y})} = \frac{\sqrt{\lambda}}{\sqrt{n}}$, and there's no guarantee in the Poisson model that the best way to estimate $\sqrt{\lambda}$ in this standard error calculation is with the sample SD $\hat{\sigma}$ (in fact we have a **strong hint** from the above MLE calculation that the sample variance is **irrelevant** to the estimation of λ in the Poisson model, since the sample variance does not arise in the Poisson likelihood).

The right (large-sample) likelihood-based **standard error** for $\hat{\lambda}_{\text{MLE}}$, using the **Fisher information** logic we examined earlier, is obtained from the following calculation:

$$\begin{aligned} \frac{\partial^2}{\partial \lambda^2} \log l(\lambda|y) &= -\frac{s}{\lambda^2}, \quad \text{so} \qquad (64) \\ \hat{I}\left(\hat{\lambda}_{\text{MLE}}\right) &= \left[-\frac{\partial^2}{\partial \lambda^2} \log l(\lambda|y)\right]_{\lambda=\hat{\lambda}_{\text{MLE}}} \\ &= \left(\frac{s}{\lambda^2}\right)_{\lambda=\bar{y}} = \frac{s}{\bar{y}^2} = \frac{n}{\bar{y}}, \quad \text{and} \\ \hat{V}\left(\hat{\lambda}_{\text{MLE}}\right) &= \hat{I}^{-1}\left(\hat{\lambda}_{\text{MLE}}\right) = \frac{\bar{y}}{n} = \frac{\hat{\lambda}_{\text{MLE}}}{n}. \end{aligned}$$

Prior-Likelihood-Posterior Summaries

So in this case study Fisher's **likelihood-maximizing** approach would **estimate** λ by $\hat{\lambda}_{\text{MLE}} = \bar{y} = \frac{29}{14} \doteq 2.071$, with a **give-or-take** of

$$\widehat{SE}\left(\hat{\lambda}_{\text{MLE}}\right) = \frac{\sqrt{\hat{\lambda}_{\text{MLE}}}}{\sqrt{n}} = \frac{1.44}{\sqrt{14}} \doteq 0.385.$$

All of this may be **summarized** in the following table:

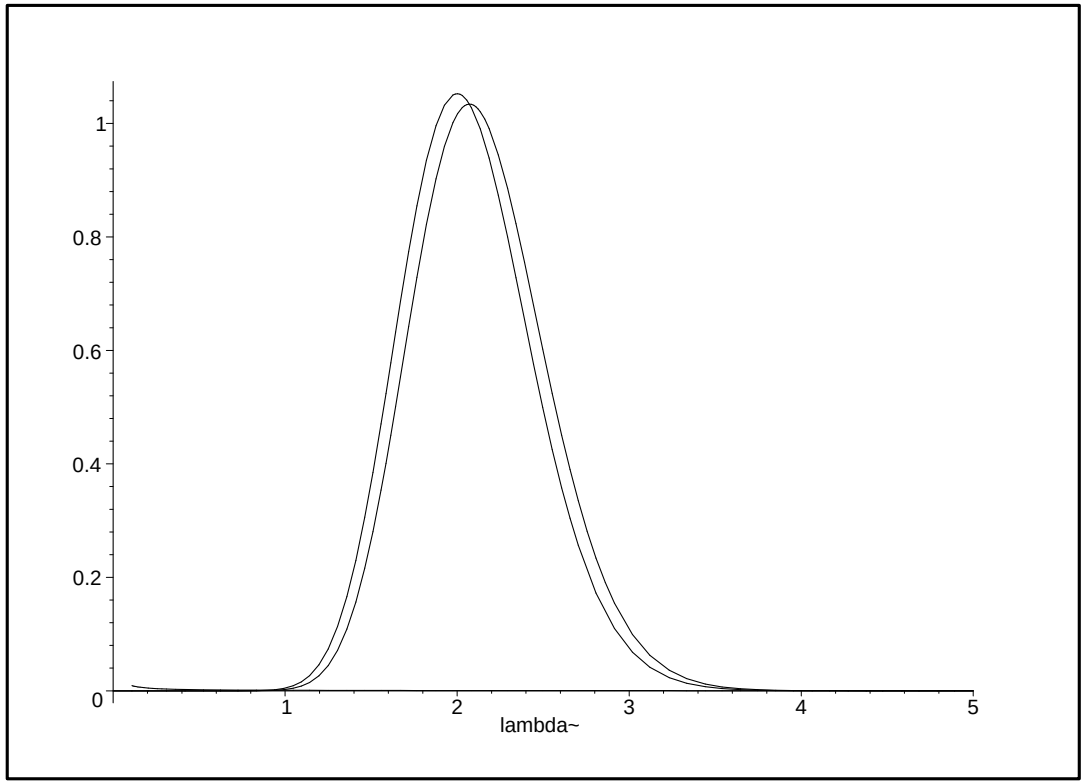
		Likelihood		
	Prior	Maximizing	Integrating	Posterior
Mean/Estimate	1.00	2.071	2.143	2.071
SD/SE	31.6	0.385	0.391	0.385

The **discrepancies** between the likelihood-maximizing and likelihood-integrating columns in this table would be smaller with a larger sample size and would **tend to 0** as $n \to \infty$.

The **prior-likelihood-posterior plot** comes out like this:

```
> plot( { p( lambda, 0.001, 0.001 ), p( lambda, 30, 14 ),
    p( lambda, 29.001, 14.001 ) }, lambda = 0 .. 5, color = black );
```

Prior-Likelihood-Posterior Summaries (continued)



For **interval estimation** in the maximum-likelihood approach the best we could do, using the technology I've described to you so far, would be to appeal to the **CLT** (even though n is only 14) and use $\hat{\lambda}_{\text{MLE}} \pm 1.96\,\widehat{SE}(\hat{\lambda}_{\text{MLE}}) \doteq 2.071 \pm (1.96)(0.385) \doteq (1.316, 2.826)$ as an **approximate 95% confidence interval** for λ.

The Interval Should Be Asymmetric in This Problem

You can see from the previous plot that the likelihood function is **asymmetric**, so a more careful method (e.g., the **bootstrap**; Efron 1979) would be needed to create a better interval estimate from the likelihood point of view.

Some trial and error with `Maple` can be used to find the lower and upper limits of the **central 95% posterior interval** for λ:

```
> evalf( Int( p( lambda, 29.001, 14.001 ), lambda = 0 .. 1.316 ) );
                                .01365067305
> evalf( Int( p( lambda, 29.001, 14.001 ), lambda = 0 .. 1.4 ) );
                                .02764660367
> evalf( Int( p( lambda, 29.001, 14.001 ), lambda = 0 .. 1.387 ) );
                                .02495470339
> evalf( Int( p( lambda, 29.001, 14.001 ), lambda = 2.826 .. infinity ) );
                                .03403487851
> evalf( Int( p( lambda, 29.001, 14.001 ), lambda = 2.890 .. infinity ) );
                                .02505307631
```

Thus a **95% (central) posterior interval** for λ, given a diffuse prior, runs from **1.387** to **2.890**, and is (correctly) **asymmetric** around the posterior mean of 2.071.

The R Solution

R can be used to work out the **limits of this interval** even more readily:

```
> help( qgamma )
GammaDist                  package:base                  R Documentation
The Gamma Distribution
Description:
     Density, distribution function, quantile function and random
     generation for the Gamma distribution with parameters 'shape' and
     'scale'.
Usage:
     dgamma(x, shape, scale=1, log = FALSE)
     pgamma(q, shape, scale=1, lower.tail = TRUE, log.p = FALSE)
     qgamma(p, shape, scale=1, lower.tail = TRUE, log.p = FALSE)
     rgamma(n, shape, scale=1)
Arguments:
    x, q: vector of quantiles.
       p: vector of probabilities.
       n: number of observations.
shape, scale: shape and scale parameters.
log, log.p: logical; if TRUE, probabilities p are given as log(p).
lower.tail: logical; if TRUE (default), probabilities are P[X <= x],
```

The R Solution (continued)

```
          otherwise, P[X > x].
Value:
     'dgamma' gives the density, 'pgamma' gives the distribution
     function 'qgamma' gives the quantile function, and 'rgamma'
     generates random deviates.
See Also:
     'gamma' for the Gamma function, 'dbeta' for the Beta distribution
     and 'dchisq' for the chi-squared distribution which is a special
     case of the Gamma distribution.
> qgamma( 0.025, 29.001, 1 / 14.001 )
[1] 1.387228
> qgamma( 0.975, 29.001, 1 / 14.001 )
[1] 2.890435
```

Maple or R can also be used to obtain the **probability content**, according to the posterior distribution, of the approximate 95% (large-sample) likelihood-based interval:

```
> evalf( Int( p( lambda, 29.001, 14.001 ), lambda = 1.316 .. 2.826 ) );
                              .9523144484
```

Predictive Distributions

So the **maximization** approach has led to **decent approximations** here (later I'll give examples where maximum likelihood doesn't do well in small samples).

Predictive distributions in this model can be computed by `Maple` **in the usual way**: e.g., to compute $p(y_{n+1}|y)$ for $y = (y_1, \ldots, y_n)$ we want to evaluate

$$\begin{aligned}
p(y_{n+1}|y) &= \int_0^\infty p(y_{n+1}, \lambda|y)\, d\lambda \\
&= \int_0^\infty p(y_{n+1}|\lambda, y)\, p(\lambda|y)\, d\lambda \qquad (65) \\
&= \int_0^\infty p(y_{n+1}|\lambda)\, p(\lambda|y)\, d\lambda \\
&= \int_0^\infty \frac{\lambda^{y_{n+1}} e^{-\lambda}}{y_{n+1}!} \frac{(\beta^*)^{\alpha^*}}{\Gamma(\alpha^*)} \lambda^{\alpha^*-1} e^{-\beta^*\lambda}\, d\lambda, \\
&= \frac{(\beta^*)^{\alpha^*}}{\Gamma(\alpha^*)\, y_{n+1}!} \int_0^\infty \lambda^{(\alpha^*+y_{n+1})-1} e^{-(\beta^*+1)\lambda}\, d\lambda,
\end{aligned}$$

where $\alpha^* = \alpha + s$ and $\beta^* = \beta + n$; in these expressions y_{n+1} is a **non-negative integer**.

Predictive Distributions (continued)

```
> assume( astar > 0, bstar > 0, yf > 0 );
> simplify( bstar^astar * int( lambda^( astar + yf - 1 ) *
    exp( - ( bstar + 1 ) * lambda ), lambda = 0 .. infinity ) /
    ( GAMMA( astar ) * yf! ) );
                astar~                  (-astar~ - yf~)
          bstar~       (bstar~ + 1)                    GAMMA(astar~ + yf~)
          ------------------------------------------------------------
                          GAMMA(astar~) GAMMA(yf~ + 1)
```

A bit of **rearranging** then gives that for $y_{n+1} = 0, 1, \ldots,$

$$p(y_{n+1}|y) = \frac{\Gamma(\alpha^* + y_{n+1})}{\Gamma(\alpha^*)\,\Gamma(y_{n+1}+1)} \left(\frac{\beta^*}{\beta^*+1}\right)^{\alpha^*} \left(\frac{1}{\beta^*+1}\right)^{y_{n+1}}. \qquad (66)$$

This is called the **Poisson-Gamma** distribution, because (65) is asking us to take a **mixture** (weighted average) of Poisson distributions, using probabilities from a Gamma distribution as the mixing weights.

(66) is a generalization of the **negative binomial** distribution (e.g., Johnson and Kotz 1994), which you may have encountered in your earlier probability study.

The Poisson-Gamma Distribution

Maple can try to get simple expressions for the **mean** and **variance** of this distribution:

```
> assume( alpha > 0, beta > 0 );
> pg := ( y, alpha, beta ) -> GAMMA( alpha + y ) *
     ( beta / ( beta + 1 ) )^alpha * ( 1 / ( beta + 1 ) )^y /
     ( GAMMA( alpha ) * GAMMA( y + 1 ) );
                                                  /  beta  \alpha /   1    \y
                                 GAMMA(alpha + y) |-------|      |--------|
                                                  \beta + 1/      \beta + 1/
    pg := (y, alpha, beta) -> ----------------------------------------
                                          GAMMA(alpha) GAMMA(y + 1)
> simplify( sum( pg( y, alpha, beta ), y = 0 .. infinity ) );
                                          1
> simplify( sum( y * pg( y, alpha, beta ), y = 0 .. infinity ) );
                                        alpha~
                                        ------
                                        beta~
```

So the **mean** of the Poisson-Gamma(α^*, β^*) distribution is $E(y_{n+1}|y) = \frac{\alpha^*}{\beta^*}$.

```
> simplify( sum( ( y - alpha / beta )^2 * pg( y, alpha, beta ),
    y = 0 .. infinity ) );
                                  alpha~ (beta~ + 1)
                                  ------------------
                                                2
                                       beta~
```

And the variance of the Poisson-Gamma(α^*, β^*) distribution is

$$V(y_{n+1}|y) = \frac{\alpha^*}{\beta^*}\left(1 + \frac{1}{\beta^*}\right). \tag{67}$$

This provides an interesting **contrast between inference and prediction**: we've already seen in this model that the posterior mean and variance of λ are $\frac{\alpha^*}{\beta^*} = \frac{\alpha+s}{\beta+n}$ and $\frac{\alpha^*}{(\beta^*)^2} = \frac{\alpha+s}{(\beta+n)^2}$, respectively.

Thus λ (the **inferential** objective) and y_{n+1} (the **predictive** objective) have the same posterior mean, but the posterior variance of y_{n+1} is **much larger**, as can be seen by the following argument.

Contrasting Inference and Prediction (continued)

	Posterior	
Quantity	Mean	Variance
λ	$\frac{\alpha+s}{\beta+n}$	$\frac{\alpha+s}{(\beta+n)^2} = \frac{\alpha+s}{\beta+n}\left(0+\frac{1}{\beta+n}\right)$
y_{n+1}	$\frac{\alpha+s}{\beta+n}$	$\frac{\alpha+s}{\beta+n}\left(1+\frac{1}{\beta+n}\right)$

(1) Denoting by μ the mean of the **population** from which the Y_i are thought of as (like) a random sample, when n is large α and β will be **small** in relation to s and n, respectively, and the ratio $\bar{y} = \frac{s}{n}$ should **more and more closely approach** μ — thus for large n,

$$E(\lambda|y) = E(y_{n+1}|y) \doteq \mu. \tag{68}$$

(2) For the Poisson distribution the (population) mean μ and variance σ^2 are **equal**, meaning that for large n the ratio $\frac{\alpha+s}{\beta+n}$ will be close both to μ and to σ^2.

Thus for large n,

$$V(\lambda|y) \doteq \frac{\sigma^2}{n} \quad \text{but} \quad V(y_{n+1}|y) \doteq \sigma^2. \tag{69}$$

Predictive Model-Checking

An informal way to restate (69) is to say that accurate **prediction** of new data is an **order of magnitude harder** (in powers of n) than accurate **inference** about population parameters.

Bayesian model-checking with predictive distributions. One way to **check** a model like (60) is as follows — as i goes from 1 to n, do the following two things:

(1) Temporarily **set aside** observation y_i, obtaining a new dataset $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ with $(n-1)$ observations.

(2) Use the current Bayesian model applied to y_{-i} to **predict** y_i, and summarize the extent to which the actual value of y_i is **surprising** in view of this predictive distribution.

A simple measure of surprise is **predictive** z**–scores** (later, if there's time, I'll talk about a better measure):

$$z_i = \frac{y_i - E[y_i|y_{-i}]}{\sqrt{V[y_i|y_{-i}]}}. \tag{70}$$

Predictive Model-Checking (continued)

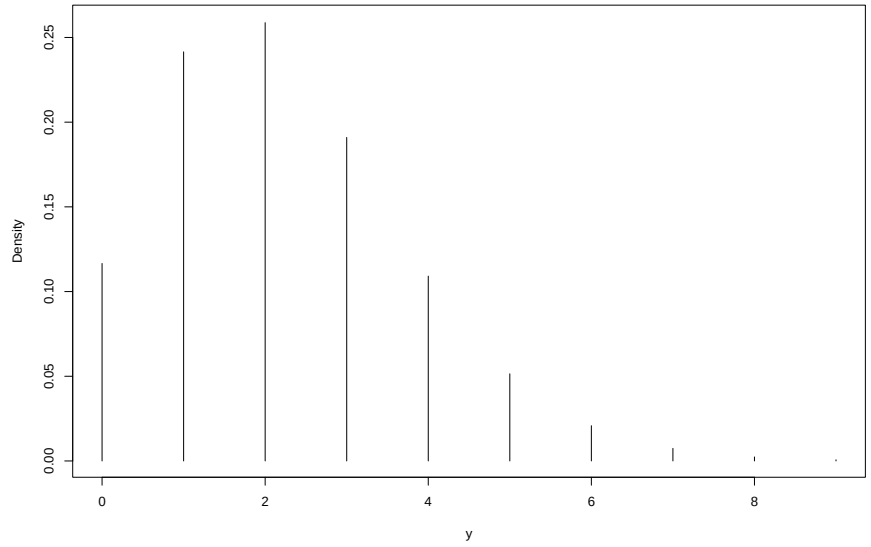
The idea is to compare the surprise measure with its **expected behavior** if the model had been **"correct"** (e.g., $z = (z_1, \ldots, z_n)$ should have mean 0 and SD 1, and a normal qqplot of the z_i values should be approximately linear).

Here's some R code to carry out this program on the **LOS data**.

```
> poisson.gamma <- function( y, alpha, beta ) {
    log.density <- lgamma( alpha + y ) + alpha *
      log( beta / ( beta + 1 ) ) - y * log( beta + 1 ) -
      lgamma( alpha ) - lgamma( y + 1 )
    return( exp( log.density ) )
  }
> print( y <- sort( y ) )
 [1] 0 1 1 1 1 1 2 2 2 2 3 3 4 6
> print( y.current <- y[ -1 ] )
 [1] 1 1 1 1 1 2 2 2 2 3 3 4 6
> print( n.current <- length( y.current ) )
[1] 13
> alpha <- beta <- 0.001
> print( s.current <- sum( y.current ) )
[1] 29
```

Predictive Model-Checking (continued)

```
> print( alpha.star <- alpha + s.current )
[1] 29.001
> print( beta.star <- beta + n.current )
[1] 13.001
> print( pg.current <- poisson.gamma( 0:9, alpha.star, beta.star ) )
 [1] 0.1165953406 0.2415099974 0.2587508547 0.1909752933 0.1091243547
 [6] 0.0514422231 0.0208209774 0.0074357447 0.0023899565 0.0007017815
> plot( 0:9, pg.current, type = 'n', xlab = 'y', ylab = 'Density' )
> for ( i in 0:9 ) {
    segments( i, 0, i, pg.current[ i + 1 ] )
  }
```



Predictive Model-Checking (continued)

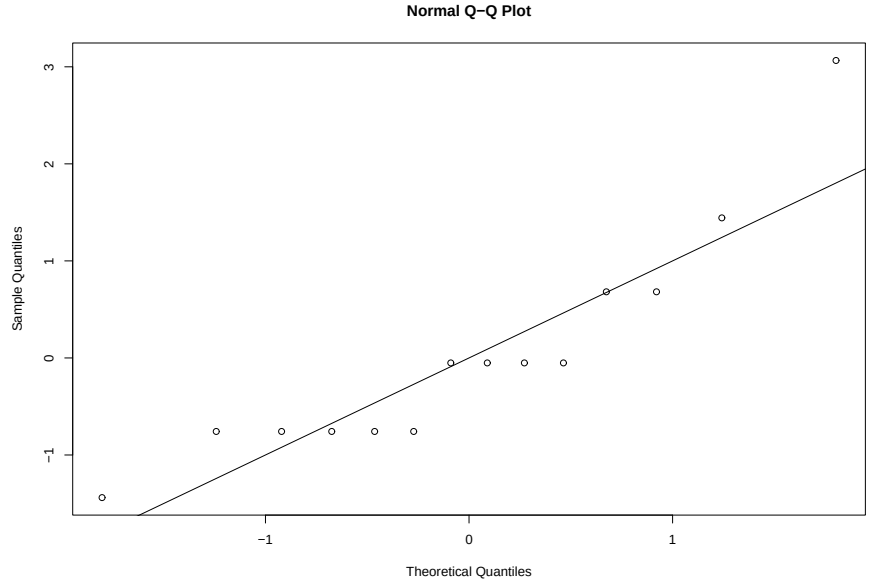
The omitted observed value of **0** is **not too unusual** in this predictive distribution.

The following `R` code **loops** through the whole dataset to get the **predictive z–scores**.

```
alpha <- beta <- 0.001
z <- rep( 0, n )
for ( i in 1:n ) {
  y.current <- y[ -i ]
  n.current <- length( y.current )
  s.current <- sum( y.current )
  alpha.star <- alpha + s.current
  beta.star <- beta + n.current
  predictive.mean.current <- alpha.star / beta.star
  predictive.SD.current <- sqrt( ( alpha.star / beta.star ) *
    ( 1 + 1 / beta.star ) )
  z[ i ] <- ( y[ i ] - predictive.mean.current ) /
    predictive.SD.current
}
```

Predictive Model-Checking (continued)

```
z
 [1] -1.43921925 -0.75757382 -0.75757382 -0.75757382 -0.75757382
 [6] -0.75757382 -0.05138023 -0.05138023 -0.05138023 -0.05138023
[11]  0.68145253  0.68145253  1.44329065  3.06513271
mean( z )
[1] 0.03133708
sqrt( var( z ) )
[1] 1.155077
qqnorm( z )
abline( 0, 1 )
```



Predictive Model-Checking (continued)

The 14 predictive z–scores have mean **0.03** (about right) and SD **1.16** (close enough to 1 when sampling variability is considered?), and the **normal qqplot** above shows that the only really surprising observation in the data, as far as the Poisson model was concerned, is the value of **6**, which has a z–score of **3.07**.

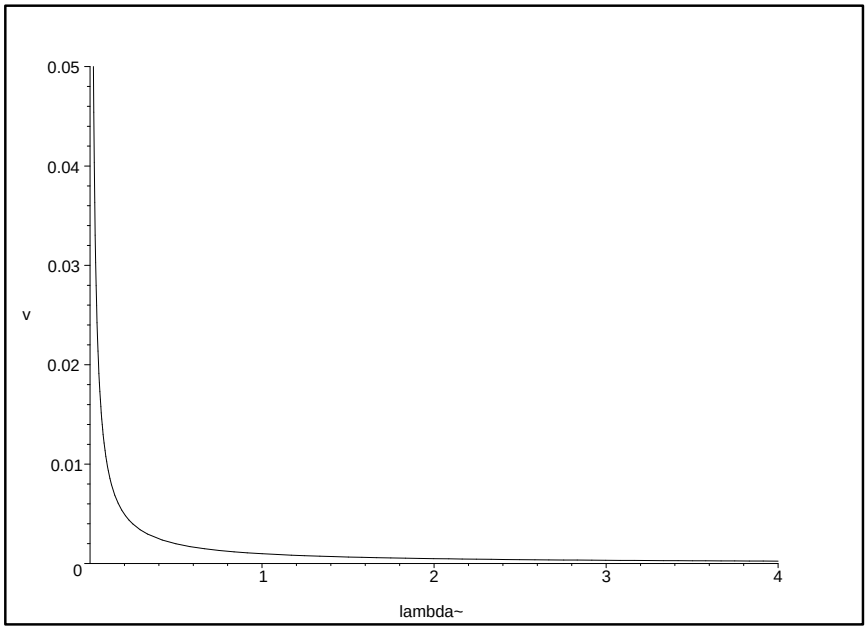
NB The figure above is only a **crude approximation** to the right qqplot, which would have to be created by **simulation**; even so it's enough to **suggest** how the model might be **improved**.

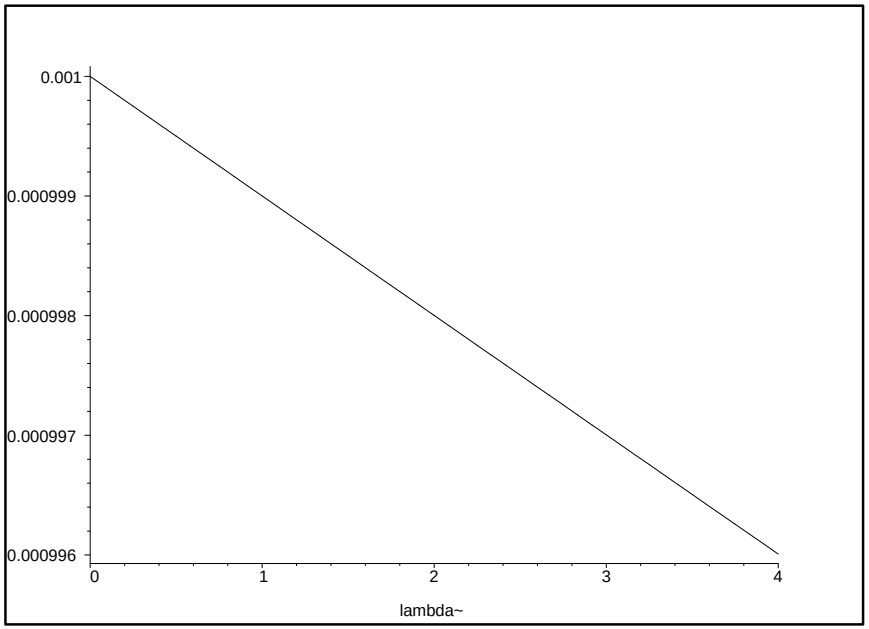
I would conclude **informally** (a) that the Poisson is a **decent** model for these data, but (b) if you wanted to expand the model in a direction suggested by this diagnostic you should look for a model with **extra-Poisson variation**: the sample VTMR in this dataset was about **1.15**.

Diffuse priors in the LOS case study. In specifying a **diffuse** prior for λ in the LOS case study, several **alternatives** to $\Gamma(\epsilon, \epsilon)$ might occur to you, including $\Gamma(1, \epsilon), \Gamma(\alpha, \beta)$ for some large α (like 20, to get a roughly **normal** prior) and small β (like 1, to have a **small prior sample size**), and $U(0, C)$ for some cutoff C (like 4) chosen to avoid **truncation** of the likelihood function, where $U(a, b)$ denotes the **uniform** distribution on (a, b).

Prior Specification: Sensitivity Analysis

```
> plot( p( lambda, 0.001, 0.001 ), lambda = 0 .. 4, v = 0 .. 0.05,
    color = black );
> plot( p( lambda, 1.0, 0.001 ), lambda = 0 .. 4, color = black );
```





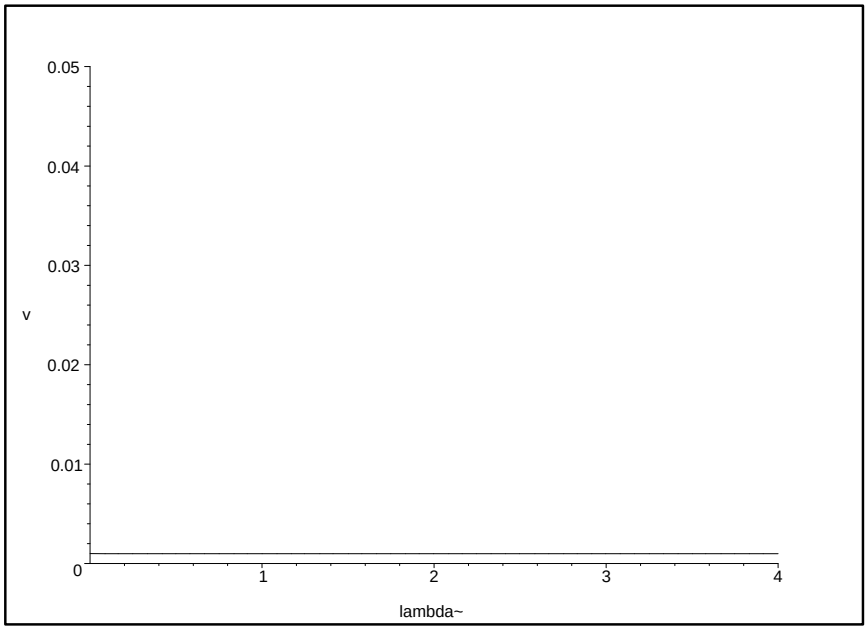
$\Gamma(1, \epsilon)$ doesn't look promising initially as a **flat** prior, but that's a consequence of `Maple`'s default choice of **vertical axis**:

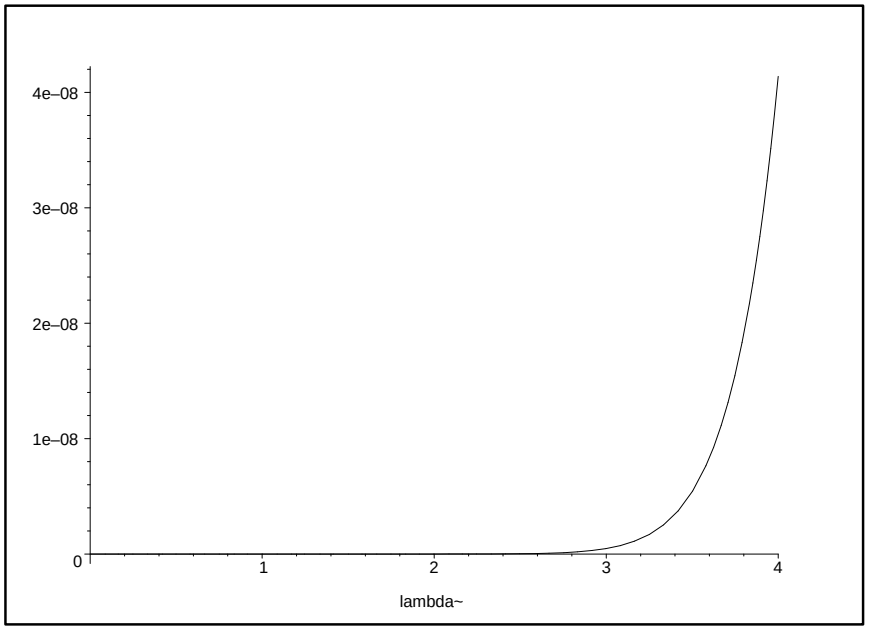
```
> plot( p( lambda, 1.0, 0.001 ), lambda = 0 .. 4, v = 0 .. 0.05,
    color = black );
```

Prior Specification: Sensitivity Analysis (continued)

```
> plot( p( lambda, 20, 1 ), lambda = 0 .. 4, color = black );
```

(Left: right-hand plot on previous page with **more sensible vertical scale**; right: $\Gamma(20, 1)$ with **not-so-sensible horizontal scale**)






```
> plot( p( lambda, 20, 1 ), lambda = 0 .. 40, color = black );
```

As is evident on the next page, $\Gamma(20, 1)$ does indeed look **not far from Gaussian**, and at first it may appear that it is indeed **relatively flat**

Prior Specification: Sensitivity Analysis (continued)

in the region where the likelihood is appreciable ($\lambda \in (1.0, 3.3)$), but we'll see below that it's actually **rather more informative** than we intend.

Recalling that the **mean** and **SD** of a $\Gamma(\alpha, \beta)$ random quantity are $\frac{\alpha}{\beta}$ and $\sqrt{\frac{\alpha}{\beta^2}}$, respectively, and that when used as a prior with the Poisson likelihood the $\Gamma(\alpha, \beta)$ distribution acts like a dataset with **prior sample size** β, you can construct the following table:

Prior				Posterior			
α	$\beta =$ Sample Size	Mean	SD	α^*	β^*	Mean	SD
0.001	0.001	1	31.6	29.001	14.001	2.071	0.385
1	0.001	1000	1000	30	14.001	2.143	0.391
20	1	20	4.47	49	15	3.267	0.467
20	0.001	20000	4472	49	14.001	3.500	0.500
$U(0, C)$ for $C > 4$		$\frac{C}{2}$	$\frac{C}{\sqrt{12}}$	30	14	2.143	0.391

Prior Specification: Sensitivity Analysis (continued)

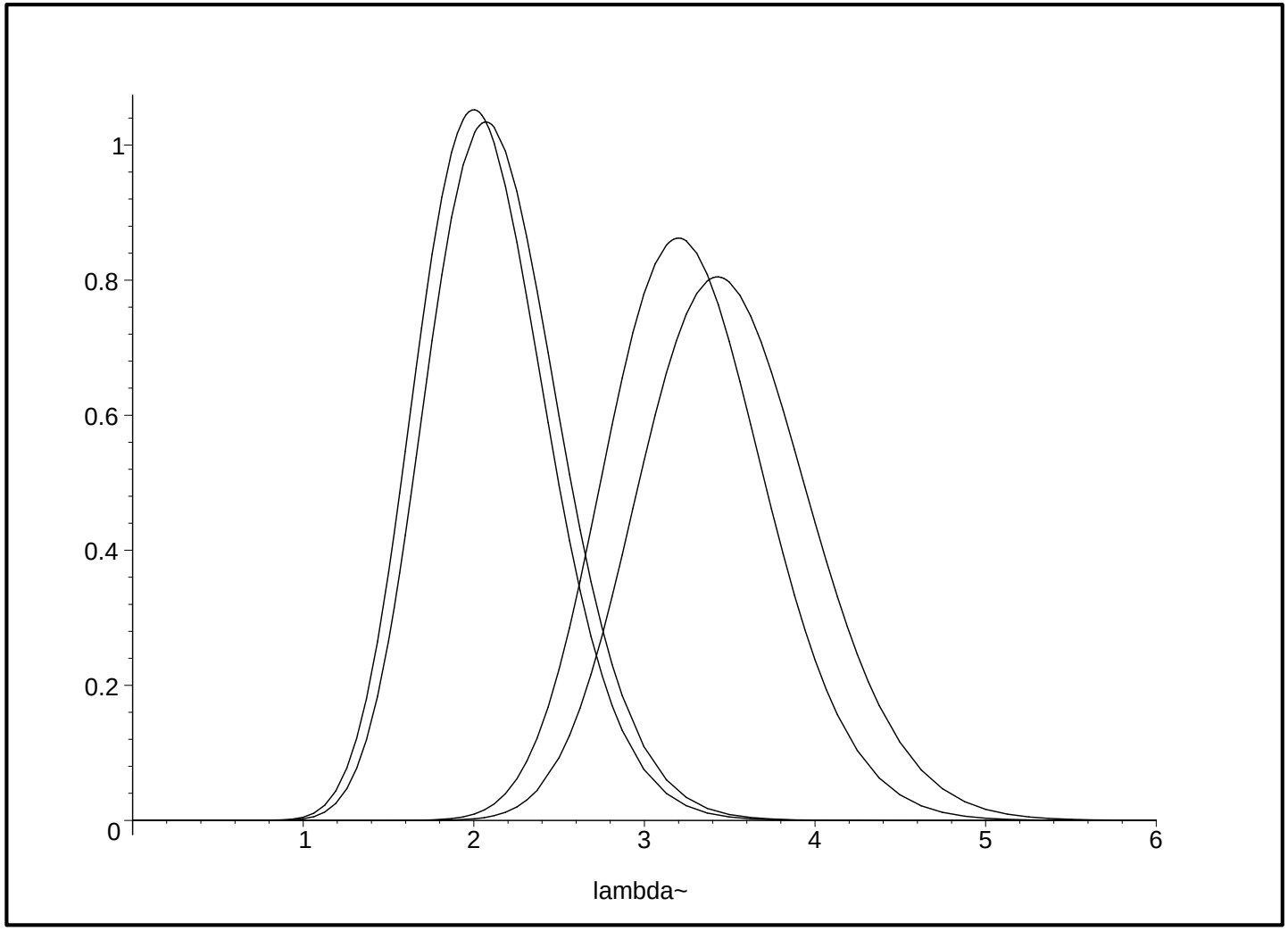
The $\Gamma(1, \epsilon)$ prior leads to an analysis that's **essentially equivalent** to the **integrated likelihood (fiducial)** approach back on page 100, and the $U(0, C)$ prior for $C > 4$ (say) produces similar results: $U(0, C)$ yields the $\Gamma(s + 1, n)$ posterior **truncated** to the right of C (and this truncation has **no effect** if you choose C big enough).

You might say that the $U(0, C)$ distribution has a **prior sample size of 0** in this analysis, and its prior mean $\frac{C}{2}$ and SD $\frac{C}{\sqrt{12}}$ (both of which can be made arbitrarily large by letting C grow without bound) are **irrelevant** (an example of how intuition can change when you depart from the class of **conjugate** priors).

```
> plot( { p( lambda, 29.001, 14.001 ), p( lambda, 30, 14.001 ),
    p( lambda, 49, 15 ), p( lambda, 49, 14.001 ) }, lambda = 0 .. 6,
    color = black );
```

The **moral** from the table above and the graph on the next page is that with only $n = 14$ observations, **some care is needed** (e.g., through **pre-posterior** analysis) to achieve a prior that **doesn't affect the posterior very much**, if that's the scientifically appropriate **information content** of the prior.

Prior Specification: Sensitivity Analysis (continued)



(Reading from left to right, **posteriors** with the following **priors**:
$\Gamma(0.001, 0.001)$, $\Gamma(1, 0.001)$, $\Gamma(20, 1)$, $\Gamma(20, 0.001)$)

2.9 Continuous Outcomes

For **continuous outcomes** there's an analogue of de Finetti's Theorem that's **equally central** to Bayesian model-building (e.g., Bernardo and Smith, 1994):

de Finetti's Theorem for Continuous Outcomes. If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of **real-valued** random quantities with probability measure p, there exists a probability measure Q over $\mathcal{D}$, the space of all distribution functions on R, such that the joint distribution function of $Y_1, \ldots, Y_n$ has the form

$$p(y_1, \ldots, y_n) = \int_{\mathcal{D}} \prod_{i=1}^{n} F(y_i)\, dQ(F), \tag{71}$$

where $Q(F) \stackrel{P}{=} \lim_{n\to\infty} p(F_n)$ and F_n is the **empirical distribution function** based on $Y_1, \ldots, Y_n$.

In other words, exchangeability of real-valued observables is **equivalent** to the hierarchical model

$$\begin{array}{rcll} F & \sim & p(F) & \text{(prior)} \\ (Y_i|F) & \stackrel{\text{IID}}{\sim} & F & \text{(likelihood)} \end{array} \tag{72}$$

Model Uncertainty

for some **prior distribution** p on the **set** $\mathcal{D}$ **of all possible CDFs**.

This prior makes the continuous form of de Finetti's Theorem **considerably harder to apply**: to take the elicitation task seriously is to try to specify a probability distribution on a **function space** (F is in effect an **infinite-dimensional** parameter).

(**NB** This task is not unique to Bayesians — it's equivalent to asking **"Where does the likelihood come from?"** in frequentist analyses of observational data.)

What people often do in practice is to appeal to considerations that narrow down the field, such as an *a priori* judgment that the Y_i ought to be **symmetrically** distributed about a measure of center μ, and then try to use a fairly **rich parametric family** satisfying (e.g.) the symmetry restriction as a substitute for all of $\mathcal{D}$.

Strictly speaking you're not supposed to look at the Y_i while specifying your prior on $\mathcal{D}$ — this can lead to a failure to fully assess and propagate **model uncertainty** — but not doing so can permit the data to surprise you in ways

that would make you want to go back and revise your prior (this is an example of **Cromwell's Rule** in action).

As mentioned earlier, in this course I'll suggest two potential ways out of this dilemma, based on **out-of-sample predictive validation** (the model-checking in the LOS data above was an example of this) and

Bayesian nonparametrics.

Case Study: *Measurement of physical constants.* What used to be called the National Bureau of Standards (NBS) in Washington, DC, conducts extremely high precision measurement of physical constants, such as the actual weight of so-called **check-weights** that are supposed to serve as reference standards (like the official kg).

In 1962–63, for example, $n = 100$ weighings (listed below) of a block of metal called **NB10**, which was supposed to weigh exactly 10g, were made under conditions **as close to IID as possible** (Freedman et al., 1998).

The data are on the next page, and give rise to (at least) the following questions: **Q:** (a) How much does NB10 **really weigh**? (b) How certain are you given the data that the true weight of NB10 is **less than** (say) 405.25?

And (c) How accurately can you **predict** the 101st measurement?

Value	375	392	393	397	398	399	400	401
Frequency	1	1	1	1	2	7	4	12
Value	402	403	404	405	406	407	408	409
Frequency	8	6	9	5	12	8	5	5
Value	410	411	412	413	415	418	423	437
Frequency	4	1	3	1	1	1	1	1

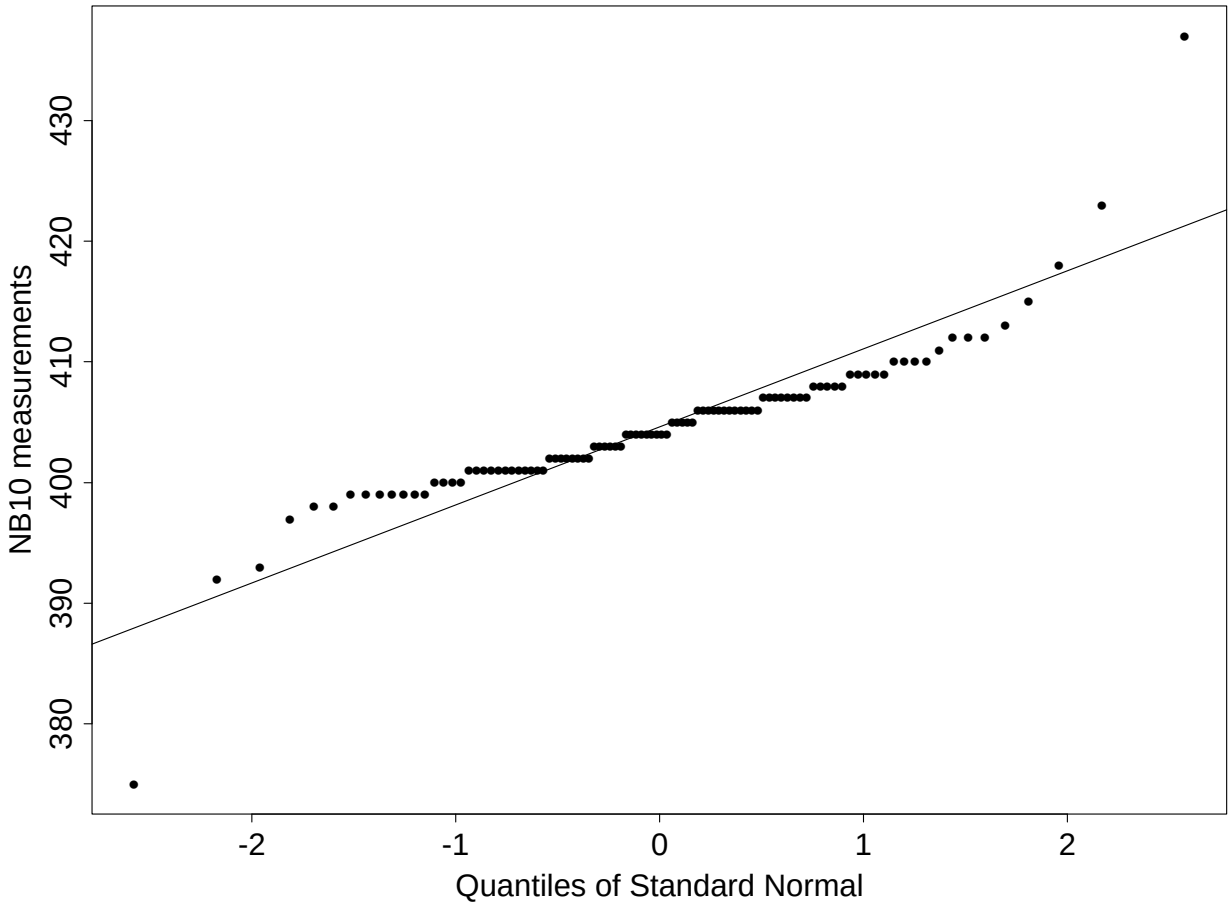
The graph below is a **normal qqplot** of the 100 measurements $y = (y_1, \ldots, y_n)$, which have a mean of $\bar{y} = 404.6$ (the units are **micrograms below 10g**) and an SD of $s = 6.5$.

Evidently it's plausible in answering these questions to assume **symmetry** of the "underlying distribution" F in de Finetti's Theorem.

One standard choice, for instance, is the **Gaussian:**

$$\begin{aligned} (\mu, \sigma^2) &\sim p(\mu, \sigma^2) \\ (Y_i | \mu, \sigma^2) &\overset{\text{IID}}{\sim} N(\mu, \sigma^2) \,. \end{aligned} \tag{73}$$

Diagnosing Non-Normality



Here $N\left(\mu, \sigma^2\right)$ is the familiar **normal density**

$$p(y_i|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \mu\right)^2\right]. \tag{74}$$

One-Parameter Gaussian Location Model

Even though you can see from the previous graph that (73) is **not a good model** for the NB10 data, I'm going to fit it anyway, for practice in working with the normal distribution from a Bayesian point of view (later we'll **improve** upon the Gaussian).

(73) is more **complicated** than the models in the AMI and LOS case studies because the parameter θ here is a **vector**: $\theta = (\mu, \sigma^2)$.

To warm up for this new complexity let's first consider a **cut-down version of the model** in which we pretend that σ is known to be $\sigma_0 = 6.5$ (the sample SD).

This **simpler model** is then

$$\left\{ \begin{array}{ccc} \mu & \sim & p(\mu) \\ (Y_i|\mu) & \overset{\text{IID}}{\sim} & N\left(\mu, \sigma_0^2\right) \end{array} \right\}. \tag{75}$$

The **likelihood function** in this model is

$$l(\mu|y) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_0^2}(y_i - \mu)^2\right]$$

Gaussian Inference For μ

$$
\begin{aligned}
&= c \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \qquad (76)\\
&= c \exp\left[-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n} y_i^2 - 2\,\mu\sum_{i=1}^{n} y_i + n\mu^2\right)\right]\\
&= c \exp\left[-\frac{1}{2\left(\frac{\sigma_0^2}{n}\right)}(\mu-\bar{y})^2\right].
\end{aligned}
$$

Thus the likelihood function, when thought of as a **density** for μ, is a **normal distribution** with mean $\bar{y}$ and SD $\frac{\sigma_0}{\sqrt{n}}$.

Notice that this SD is the same as the frequentist **standard error** for $\bar{Y}$ based on an IID sample of size n from the $N\left(\mu, \sigma_0^2\right)$ distribution.

(76) also shows that the sample mean $\bar{y}$ is a **sufficient statistic** for μ in model (75).

In finding the conjugate prior for μ it would be nice if the **product of two normal distributions is another normal distribution**, because that would demonstrate that the conjugate prior is normal.

Gaussian Inference For μ (continued)

Suppose therefore, to see where it leads, that the **prior for** μ is (say)

$$p(\mu) = N\left(\mu_0, \sigma_\mu^2\right).$$

Then **Bayes's Theorem** would give

$$\begin{aligned} p(\mu|y) &= c\, p(\mu)\, l(\mu|y) \qquad (77) \\ &= c \exp\left[-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2\right] \exp\left[-\frac{n}{2\sigma_0^2}(\mu-\bar{y})^2\right] \\ &= c\, \exp\left\{-\frac{1}{2}\left[\frac{(\mu-\mu_0)^2}{\sigma_\mu^2} + \frac{n(\mu-\bar{y})^2}{\sigma_0^2}\right]\right\}, \end{aligned}$$

and we want this to **be of the form**

$$\begin{aligned} p(\mu|y) &= c \exp\left\{-\frac{1}{2}\left[A(\mu-B)^2 + C\right]\right\} \\ &= c \exp\left\{-\frac{1}{2}\left[A\mu^2 - 2AB\mu + (AB^2 + C)\right]\right\} \qquad (78) \end{aligned}$$

for some B, C, and $A > 0$.

`Maple` can help **see if this works**:

Gaussian Inference For μ (continued)

```
> collect( ( mu - mu0 )^2 / sigmamu^2 +
    n * ( mu - ybar )^2 / sigma0^2, mu );
                                                                  2          2
/   1          n   \   2   /     mu0        n ybar \        mu0      n ybar
|-------- + -------| mu  + |-2 -------- - 2 -------| mu + -------- + -------
|       2         2|       |          2           2|              2         2
\sigmamu    sigma0 /       \   sigmamu      sigma0 /      sigmamu    sigma0
```

Matching coefficients for A and B (we don't really care about C) gives

$$A = \frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2} \quad \text{and} \quad B = \frac{\frac{\mu_0}{\sigma_\mu^2} + \frac{n\bar{y}}{\sigma_0^2}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}}. \tag{79}$$

Since $A > 0$ this demonstrates two things: (1) the **conjugate prior** for μ in model (75) is **normal**, and (2) the **conjugate updating rule** (when σ_0 is assumed known) is

$$\left\{ \begin{array}{c} \mu \sim N\left(\mu_0, \sigma_\mu^2\right) \\ \left(Y_i|\mu\right) \overset{\text{IID}}{\sim} N\left(\mu, \sigma_0^2\right), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\mu|y) = (\mu|\bar{y}) = N\left(\mu_*, \sigma_*^2\right), \tag{80}$$

where the **posterior mean and variance** are given by

$$\mu_* = B = \frac{\left(\frac{1}{\sigma_\mu^2}\right)\mu_0 + \left(\frac{n}{\sigma_0^2}\right)\bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} \quad \text{and} \quad \sigma_*^2 = A^{-1} = \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}}. \tag{81}$$

It becomes useful in understanding the meaning of these expressions to define the **precision** of a distribution, which is just the **reciprocal** of its variance: whereas the variance and SD scales measure **uncertainty**, the precision scale quantifies **information** about an unknown.

With this convention (81) has a series of **intuitive interpretations**, as follows:

- The **prior**, considered as an **information source**, is Gaussian with mean μ_0, variance σ_μ^2, and **precision** $\frac{1}{\sigma_\mu^2}$, and when viewed as a data set consists of n_0 (to be determined below) observations;

- The **likelihood**, considered as an **information source**, is Gaussian with mean $\bar{y}$, variance $\frac{\sigma_0^2}{n}$, and **precision** $\frac{n}{\sigma_0^2}$, and when viewed as a data set consists of n observations;

Conjugate Updating: Gaussian Inference About μ

• The **posterior**, considered as an **information source**, is Gaussian, and the posterior mean is a **weighted average** of the prior mean and data mean, with weights given by the **prior** and **data precisions**;

• The **posterior precision** (the reciprocal of the posterior variance) is just the **sum** of the prior and data precisions (this is why people invented the idea of precision — on this scale information about μ in model (75) is **additive**); and

• **Rewriting** μ_* as

$$\mu_* = \frac{\left(\frac{1}{\sigma_\mu^2}\right)\mu_0 + \left(\frac{n}{\sigma_0^2}\right)\bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} = \frac{\left(\frac{\sigma_0^2}{\sigma_\mu^2}\right)\mu_0 + n\bar{y}}{\frac{\sigma_0^2}{\sigma_\mu^2} + n}, \tag{82}$$

you can see that the **prior sample size** is

$$n_0 = \frac{\sigma_0^2}{\sigma_\mu^2} = \frac{1}{\left(\frac{\sigma_\mu}{\sigma_0}\right)^2}, \tag{83}$$

which makes sense: the **bigger** σ_μ is in relation to σ_0, the **less prior information** is being incorporated in the conjugate updating (82).

Bayesian inference with multivariate θ. Returning now to (73) with σ^2 unknown, (as mentioned above) this model has a $(k = 2)$-dimensional **parameter vector** $\theta = (\mu, \sigma^2)$.

When $k > 1$ you can still use Bayes' Theorem directly to obtain the **joint posterior distribution**,

$$\begin{aligned} p(\theta|y) &= p(\mu, \sigma^2|y) = c\, p(\theta)\, l(\theta|y) \\ &= c\, p(\mu, \sigma^2)\, l(\mu, \sigma^2|y), \end{aligned} \tag{84}$$

where $y = (y_1, \ldots, y_n)$, although making this calculation directly requires a k-dimensional **integration** to evaluate the normalizing constant c; for example, in this case

$$\begin{aligned} c &= [p(y)]^{-1} = \left(\iint p(\mu, \sigma^2, y)\, d\mu\, d\sigma^2 \right)^{-1} \\ &= \left(\iint p(\mu, \sigma^2)\, l(\mu, \sigma^2|y)\, d\mu\, d\sigma^2 \right)^{-1}. \end{aligned} \tag{85}$$

Usually, however, you'll be more interested in the **marginal posterior distributions**, in this case $p(\mu|y)$ and $p(\sigma^2|y)$.

Marginalization

Obtaining these requires k integrations, each of dimension $(k-1)$, a process that people refer to as **marginalization** or **integrating out the nuisance parameters** — for example,

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y)\, d\sigma^2 \ . \tag{86}$$

Predictive distributions also involve a k-dimensional integration: for example, with $y = (y_1, \ldots, y_n)$,

$$\begin{aligned} p(y_{n+1}|y) &= \iint p(y_{n+1}, \mu, \sigma^2|y)\, d\mu\, d\sigma^2 \\ &= \iint p(y_{n+1}|\mu, \sigma^2)\, p(\mu, \sigma^2|y)\, d\mu\, d\sigma^2. \end{aligned} \tag{87}$$

And, finally, if you're interested in a **function of the parameters**, you have some more hard integrations ahead of you.

For instance, suppose you wanted the posterior distribution for the **coefficient of variation** $\lambda = g_1(\mu, \sigma^2) = \frac{\sqrt{\sigma^2}}{\mu}$ in model (73).

Then one fairly direct way to get this posterior (e.g., Bernardo and Smith, 1994) is to (a) introduce a **second function** of the parameters, say

The Integration Challenge

$\eta = g_2(\mu, \sigma^2)$, such that the mapping $f = (g_1, g_2)$ from (μ, σ^2) to (λ, η) is **invertible**; (b) compute the joint posterior for (λ, η) through the usual **change-of-variables formula**

$$p(\lambda, \eta|y) = p_{\mu,\sigma^2}\left[f^{-1}(\lambda, \eta)|y\right] \left|J_{f^{-1}}(\lambda, \eta)\right|, \tag{88}$$

where $p_{\mu,\sigma^2}(\cdot, \cdot|y)$ is the joint posterior for μ and σ^2 and $\left|J_{f^{-1}}\right|$ is the **determinant** of the **Jacobian** of the inverse transformation; and (c) **marginalize** in λ by integrating out η in $p(\lambda, \eta|y)$, in a manner analogous to (86).

Here, for instance, $\eta = g_2(\mu, \sigma^2) = \mu$ would create an invertible f, with **inverse** defined by $(\mu = \eta, \sigma^2 = \lambda^2\eta^2)$; the **Jacobian determinant** comes out $2\lambda\eta^2$ and (94) becomes $p(\lambda, \eta|y) = 2\lambda\eta^2\, p_{\mu,\sigma^2}(\eta, \lambda^2\eta^2|y)$.

This process involves **two integrations**, one (of dimension k) to get the normalizing constant that defines (88) and one (of dimension $(k-1)$) to get rid of η.

You can see that when k is a lot bigger than 2 all these integrals may create **severe computational problems** — this has been the **big stumbling block** for applied Bayesian work for a long time.

Gaussian Modeling With Unknown μ and σ^2

More than 200 years ago **Laplace** (1774) — the second applied Bayesian in history (after Bayes himself) — developed, as one avenue of solution to this problem, what people now call **Laplace approximations** to high-dimensional integrals of the type arising in Bayesian calculations (see, e.g., Tierney and Kadane, 1986).

Starting in the next case study after this one, we'll use another, computationally intensive, **simulation-based** approach: **Markov chain Monte Carlo** (MCMC).

Back to model (73). The conjugate prior for $\theta = \left(\mu, \sigma^2\right)$ in this model (see GCSR) turns out to be most simply described **hierarchically**:

$$\begin{aligned} \sigma^2 &\sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \\ (\mu|\sigma^2) &\sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right). \end{aligned} \tag{89}$$

Here saying that $\sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2)$, where SI stands for **scaled inverse**, amounts to saying that the precision $\tau = \frac{1}{\sigma^2}$ follows a **scaled** χ^2 distribution with parameters ν_0 and σ_0^2.

Gaussian Modeling With Unknown μ and σ^2 (continued)

The scaling is chosen so that σ_0^2 can be interpreted as a **prior estimate** of σ^2, with ν_0 the **prior sample size** of this estimate (i.e., **think of a prior data set with ν_0 observations and sample SD σ_0**).

Since χ^2 is a special case of the Gamma distribution, SI-χ^2 must be a special case of the **inverse Gamma** family — its **density** (see GCSR, Appendix A) is

$$\sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \leftrightarrow \tag{90}$$

$$p(\sigma^2) = \frac{\left(\frac{1}{2}\nu_0\right)^{\frac{1}{2}\nu_0}}{\Gamma\left(\frac{1}{2}\nu_0\right)} \left(\sigma_0^2\right)^{\frac{1}{2}\nu_0} \left(\sigma^2\right)^{-\left(1+\frac{1}{2}\nu_0\right)} \exp\left(-\frac{\nu_0\,\sigma_0^2}{2\sigma^2}\right).$$

As may be verified with `Maple`, this distribution has **mean** (provided that $\nu_0 > 2$) and **variance** (provided that $\nu_0 > 4$) given by

$$E\left(\sigma^2\right) = \frac{\nu_0}{\nu_0 - 2}\sigma_0^2 \quad \text{and} \quad V\left(\sigma^2\right) = \frac{2\nu_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)}\sigma_0^4. \tag{91}$$

The parameters μ_0 and κ_0 in the second level of the prior model (89), $(\mu|\sigma^2) \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$, have **simple parallel interpretations** to those of σ_0^2 and ν_0: μ_0 is the **prior estimate** of μ, and κ_0 is the **prior effective sample size** of this estimate.

Bivariate Gaussian Likelihood

The **likelihood function** in model (73), with **both** μ and σ^2 **unknown**, is

$$\begin{aligned} l(\mu, \sigma^2|y) &= c \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\ &= c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \qquad (92) \\ &= c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\,\mu \sum_{i=1}^{n} y_i + n\mu^2\right)\right]. \end{aligned}$$

The **expression in brackets** in the last line of (92) is

$$\begin{aligned} [\ \cdot\ ] &= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 + n(\mu - \bar{y})^2 - n\bar{y}^2\right] \qquad (93) \\ &= -\frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right], \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the **sample variance**. Thus

$$l(\mu, \sigma^2|y) = c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right]\right\}, \qquad (94)$$

Bivariate Gaussian Likelihood (continued)

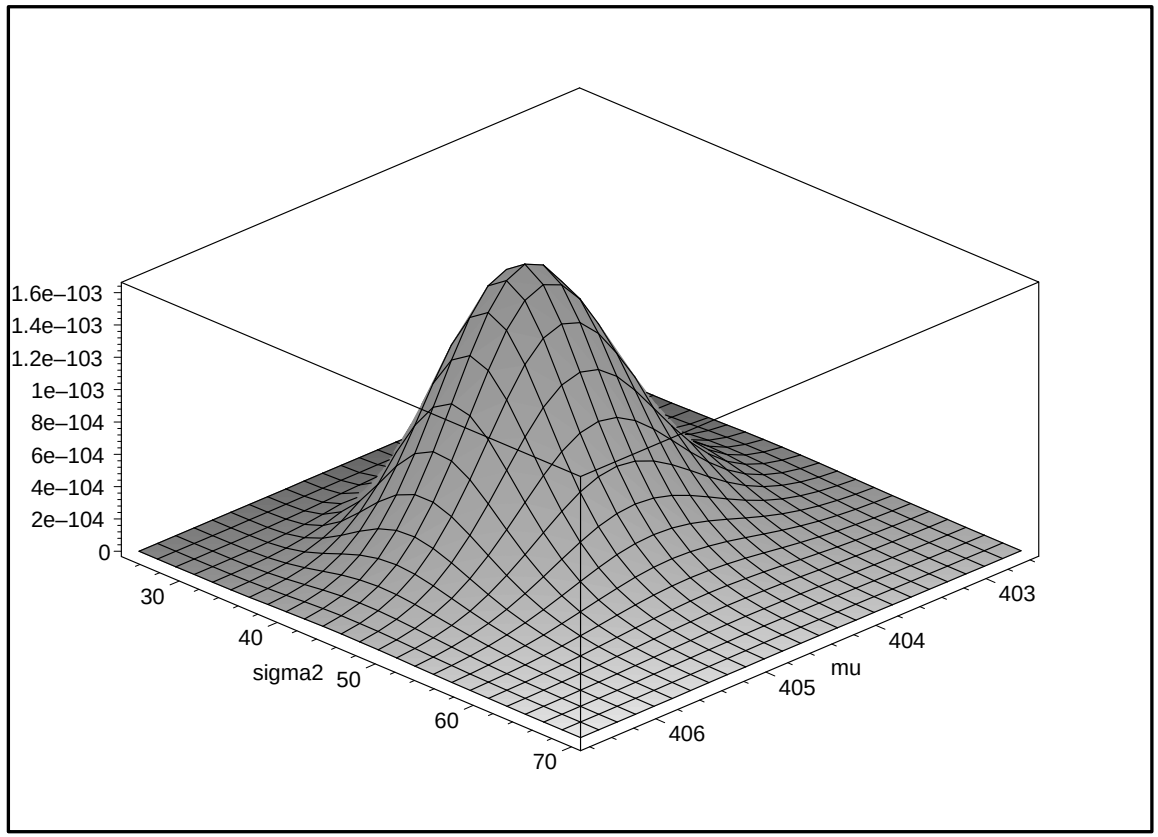
and it's clear that the **vector** $\left(\bar{y}, s^2\right)$ is **sufficient** for $\theta = \left(\mu, \sigma^2\right)$ in this model, i.e., $l(\mu, \sigma^2|y) = l(\mu, \sigma^2|\bar{y}, s^2)$.

`Maple` can be used to make **3D** and **contour plots** of this likelihood function with the NB10 data:

```
> l := ( mu, sigma2, ybar, s2, n ) -> sigma2^( - n / 2 ) *
    exp( - ( n * ( mu - ybar )^2 + ( n - 1 ) * s2 ) / ( 2 * sigma2 ) );

l := (mu, sigma2, ybar, s2, n) ->
                                                  2
          (- 1/2 n)                    n (mu - ybar)  + (n - 1) s2
    sigma2           exp(- 1/2 ---------------------------)
                                          sigma2

> plot3d( l( mu, sigma2, 404.6, 42.25, 100 ), mu = 402.6 .. 406.6,
    sigma2 = 25 .. 70 );
```

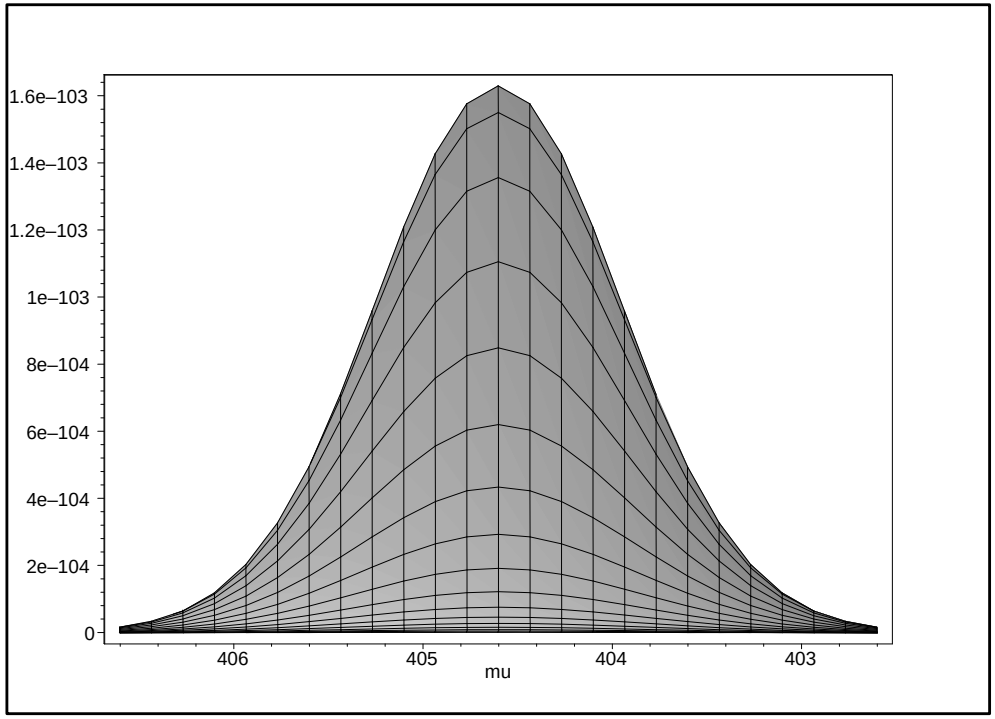
The result (next page) looks something like a **bivariate normal** density except that it's **skewed** along the σ^2 dimension:

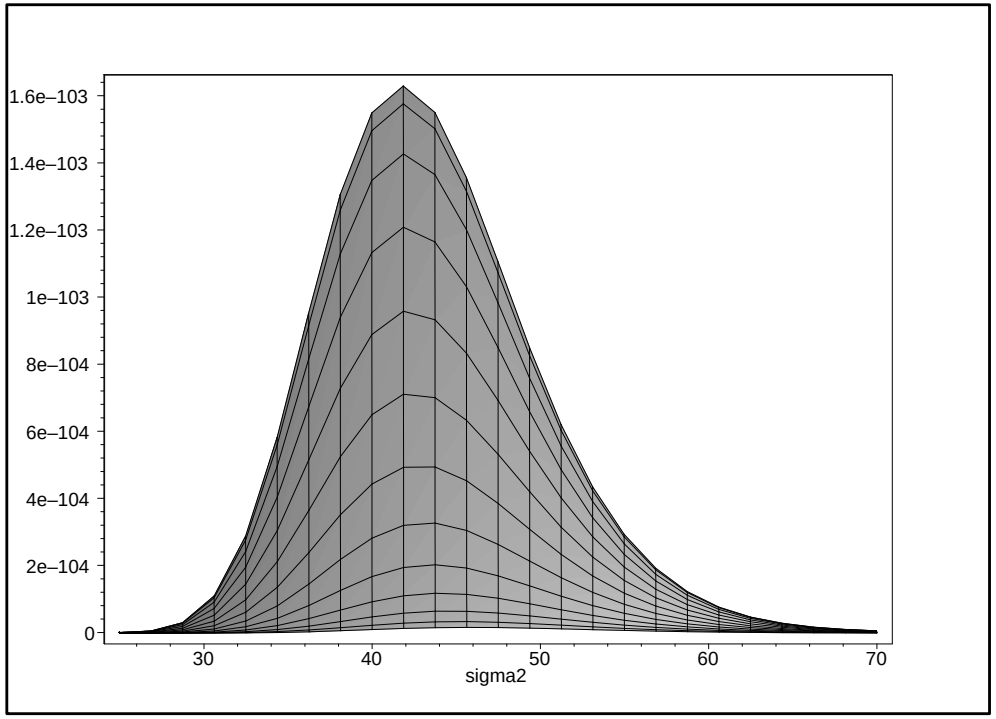
Bivariate Gaussian Likelihood (continued)



Bivariate Gaussian Likelihood (continued)

You can use the mouse to **rotate** 3D plots and get **other useful views** of them:

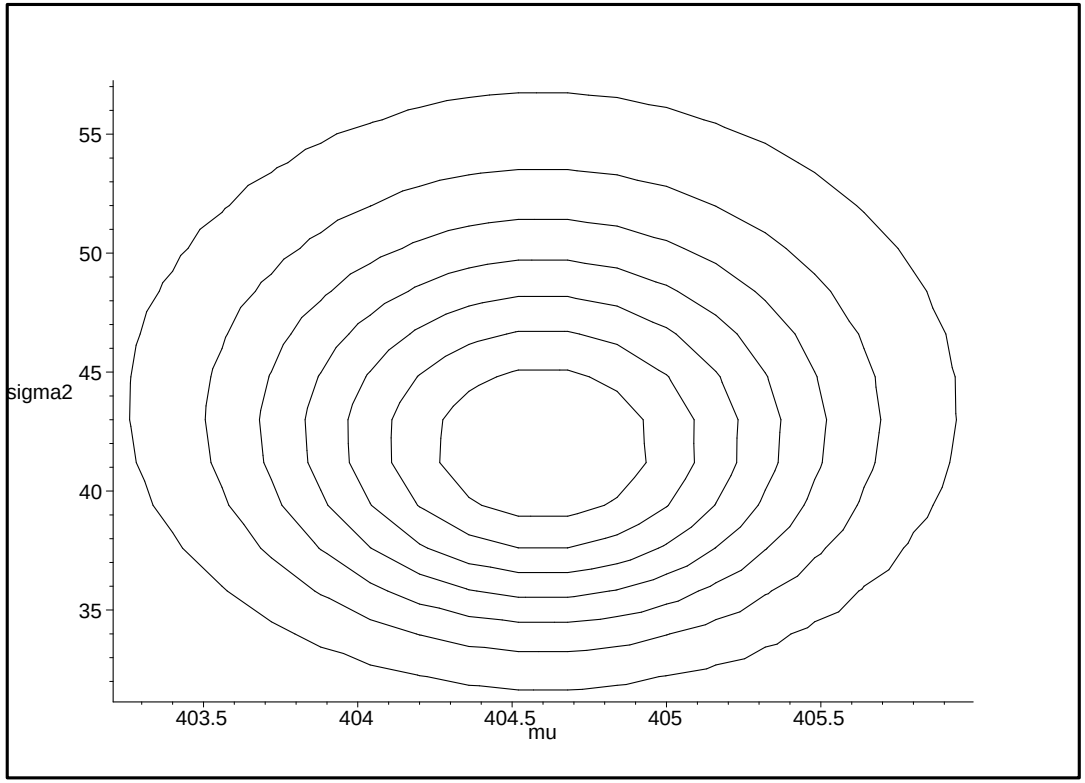




The **projection** or **shadow plot** of μ (left) looks a lot like a **normal** (or maybe a t) distribution, and the shadow plot of σ^2 (right) looks a lot like a **Gamma** (or maybe an **inverse Gamma**) distribution.

Bivariate Gaussian Likelihood (continued)

```
> plots[ contourplot ]( 10^100 * l( mu, sigma2, 404.6, 42.25, 100 ),
    mu = 402.6 .. 406.6, sigma2 = 25 .. 70, color = black );
```



The **contour plot** shows that μ and σ^2 are **uncorrelated** in the likelihood distribution, and the **skewness** of the marginal distribution of σ^2 is also evident.

Inferential analysis. Having adopted the **conjugate prior** (89), what I'd like next is simple expressions for the **marginal posterior distributions** $p(\mu|y)$ and $p(\sigma^2|y)$ and for **predictive distributions** like $p(y_{n+1}|y)$.

Fortunately, in model (73) all of the **integrations** (such as (86) and (87)) may be done **analytically** (see, e.g., Bernardo and Smith 1994), yielding the following results:

$$\begin{aligned} (\sigma^2|y,\mathcal{G}) &\sim \text{SI-}\chi^2(\nu_n,\sigma_n^2), \\ (\mu|y,\mathcal{G}) &\sim t_{\nu_n}\left(\mu_n,\frac{\sigma_n^2}{\kappa_n}\right), \quad \text{and} \\ (y_{n+1}|y,\mathcal{G}) &\sim t_{\nu_n}\left(\mu_n,\frac{\kappa_n+1}{\kappa_n}\sigma_n^2\right). \end{aligned} \tag{95}$$

In the above **expressions**

$$\begin{aligned} \nu_n &= \nu_0+n, \\ \sigma_n^2 &= \frac{1}{\nu_n}\left[\nu_0\sigma_0^2+(n-1)s^2+\frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2\right], \end{aligned}$$

Gaussian Inferential Analysis (continued)

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0 + n}\,\mu_0 + \frac{n}{\kappa_0 + n}\,\bar{y}, \quad \text{and} \\ \kappa_n &= \kappa_0 + n, \end{aligned} \tag{96}$$

$\bar{y}$ and s^2 are the usual **sample mean** and **variance** of y, and $\mathcal{G}$ denotes the assumption of the **Gaussian model**.

Here $t_\nu(\mu, \sigma^2)$ is a **location-scale** version of the usual t_ν distribution, i.e.,

$$W \sim t_\nu(\mu, \sigma^2) \iff \tfrac{W-\mu}{\sigma} \sim t_\nu.$$

This distribution (see GCSR, Appendix A) has **density**

$$\eta \sim t_\nu(\mu, \sigma^2) \leftrightarrow p(\eta) = \frac{\Gamma\left[\frac{1}{2}(\nu+1)\right]}{\Gamma\left(\frac{1}{2}\nu\right)\sqrt{\nu\pi\sigma^2}} \left[1 + \frac{1}{\nu\sigma^2}(\eta-\mu)^2\right]^{-\frac{1}{2}(\nu+1)}. \tag{97}$$

It turns out that $t_\nu(\mu, \sigma^2)$ has **mean** μ (as long as $\nu > 1$) and **variance** $\frac{\nu}{\nu-2}\sigma^2$ (as long as $\nu > 2$).

Notice that, as with all previous conjugate examples, the posterior mean is again a **weighted average** of the prior mean and data mean, with weights determined by the **prior sample size** and the **data sample size**:

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\,\mu_0 + \frac{n}{\kappa_0 + n}\,\bar{y}. \tag{98}$$

NB10 Gaussian Analysis. *Question (a):* I don't know anything about what NB10 is supposed to weigh (down to the nearest microgram) or about the accuracy of the NBS's measurement process, so I want to use a **diffuse prior** for μ and σ^2.

Considering the meaning of the **hyperparameters**, to provide little prior information I want to choose both ν_0 and κ_0 **close to 0**.

Making them exactly 0 would produce an **improper** prior distribution (which doesn't integrate to 1), but choosing positive values as close to 0 as you like yields a **proper and highly diffuse prior**.

You can see from (95, 96) that the result is then

$$(\mu|y, \mathcal{G}) \sim t_n\left[\bar{y}, \frac{(n-1)s^2}{n^2}\right] \doteq N\left(\bar{y}, \frac{s^2}{n}\right), \tag{99}$$

i.e., with diffuse prior information (as with the Bernoulli model in the AMI case study) the 95% central Bayesian interval **virtually coincides** with the usual

NB10 Gaussian Analysis (continued)

frequentist 95% confidence interval

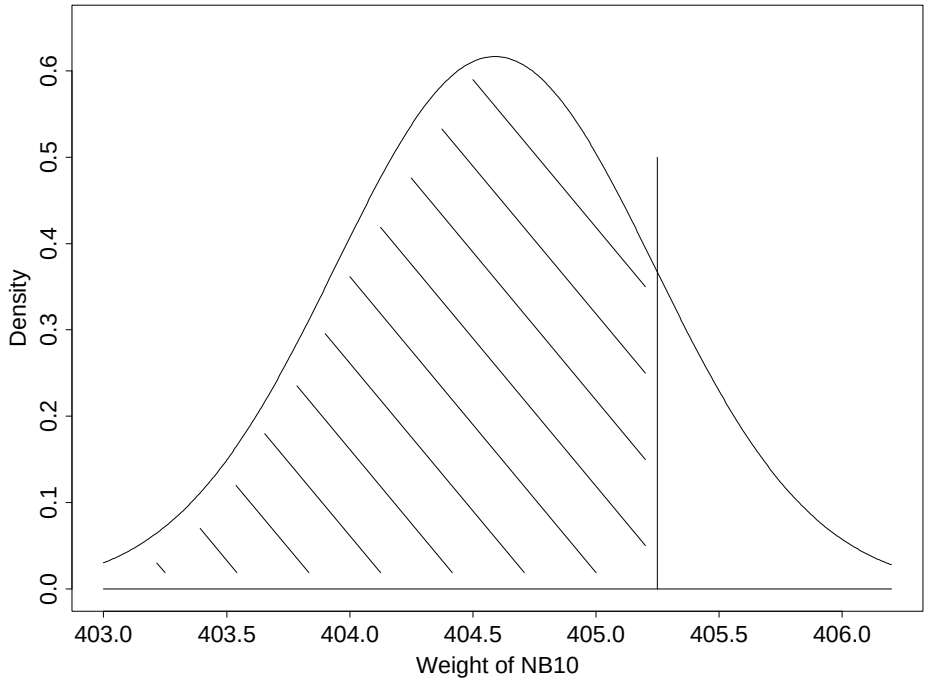
$$\bar{y} \pm t^{.975}_{n-1} \frac{s}{\sqrt{n}} = 404.6 \pm (1.98)(0.647) = (403.3, 405.9).$$

Thus both {frequentists who assume $\mathcal{G}$} and {Bayesians who assume $\mathcal{G}$ with a diffuse prior} conclude that **NB10 weighs about** 404.6μg **below 10g, give or take about** 0.65μg.

Question (b). If interest focuses on whether NB10 weighs **less than some value** like 405.25, when reasoning in a Bayesian way you can answer this question directly: the posterior distribution for μ is shown below, and $P_B(\mu < 405.25 | y, \mathcal{G}, \text{diffuse prior}) \doteq .85$, i.e., your **betting odds** in favor of the proposition that $\mu < 405.25$ are about 5.5 to 1 (see graph next page).

When reasoning in a frequentist way $P_F(\mu < 405.25)$ is **undefined**; about the best you can do is to test $H_0 : \mu < 405.25$, for which the p-value would (approximately) be $p = P_{F,\mu=405.25}(\bar{y} > 404.6) = .85$, i.e., **"insufficient evidence to reject** H_0 **at the usual significance levels"** (note the **connection** between the p-value and the posterior probability, which arises in this example because the null hypothesis is **one-sided**).

NB10 Gaussian Analysis (continued)



NB The significance test tries to answer a **different question**: in Bayesian language it looks at $P(\bar{y}|\mu)$ instead of $P(\mu|\bar{y})$.

Most people find the latter quantity **more interpretable**.

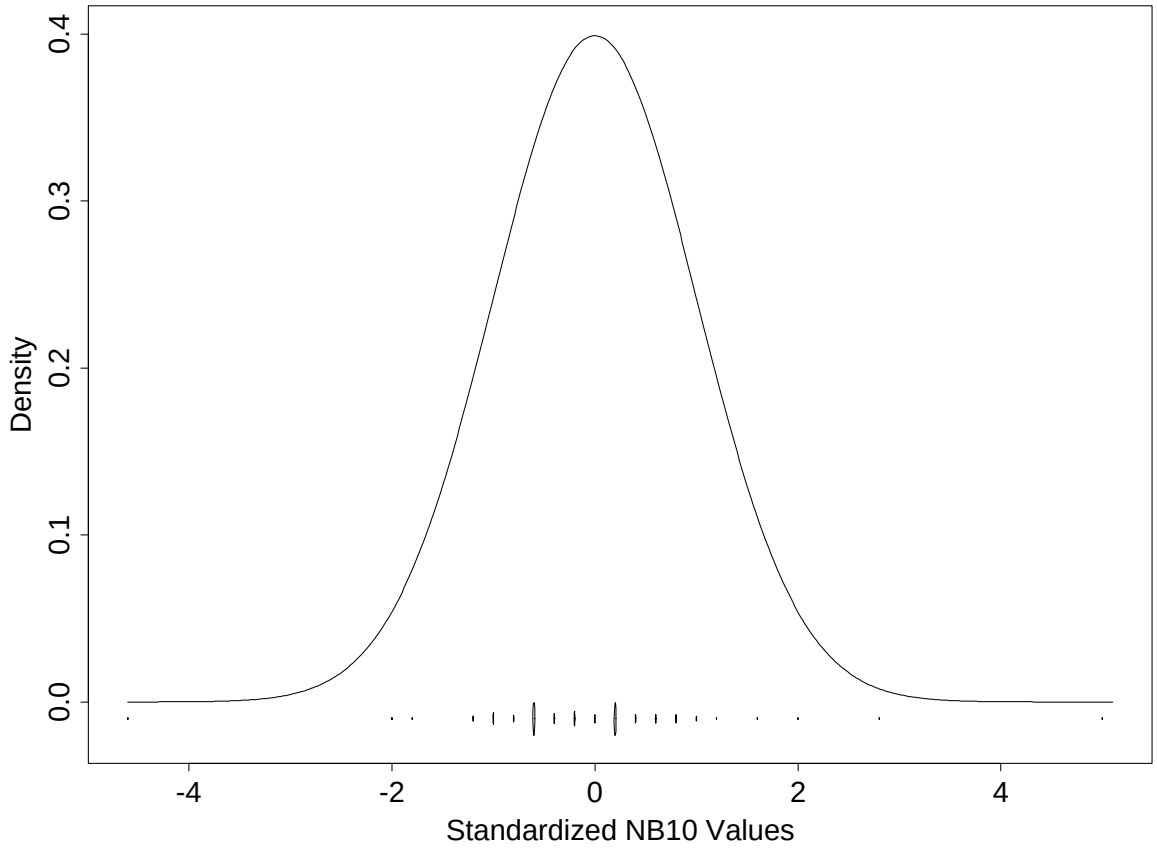
Question (c). We saw earlier that **in this model**

$$(y_{n+1}|y, \mathcal{G}) \sim t_{\nu_n}\left[\mu_n, \frac{\kappa_n + 1}{\kappa_n}\sigma_n^2\right], \tag{100}$$

NB10 Gaussian Analysis (continued)

and for n large and ν_0 and κ_0 close to 0 this is $(y_{n+1}|y, \mathcal{G}) \dot{\sim} N(\bar{y}, s^2)$, i.e., a **95% posterior predictive interval** for y_{n+1} is $(392, 418)$.

A **standardized version** of this predictive distribution is plotted below, with the standardized NB10 data values **superimposed**.



Predictive Diagnostics; Model Expansion

It's evident from this plot (and also from the normal qqplot given earlier) that the Gaussian model provides a **poor fit** for these data — the three most extreme points in the data set in standard units are $-4.6, 2.8$, and 5.0.

With the **symmetric heavy tails** indicated in these plots, in fact, the empirical CDF looks quite a bit like that of a t distribution with a rather small number of **degrees of freedom**.

This suggests revising the previous model by **expanding** it: **embedding** the Gaussian in the t family and adding a parameter ν for **tail-weight**.

This is an example of an important Bayesian idea — **model expansion**: (a) finding out how the current model M is inadequate and (b) embedding M in a richer class of models $\mathcal{M}$ of which M is a special case, where $\mathcal{M}$ is chosen to remedy the deficiencies of M revealed in (a).

Unfortunately there's no standard **closed-form conjugate** choice for the prior on ν.

A more **flexible** approach to computing is evidently needed — this is where **Markov chain Monte Carlo** methods (our next main topic) come in.

2.10 The Exponential Family

In our examples of **conjugate** analysis so far, we worked out the form of the **conjugate prior** just by **looking** at the likelihood function.

This works in **simple** problems, but it would be nice to have a **general** way of figuring out what the conjugate prior has to be (if it exists) once the likelihood is **specified**.

It was noticed a long time ago that many of the **standard sampling distributions** that you're likely to want to use in constructing likelihood functions in parametric Bayesian modeling have the **same general form**, which is referred to as the **exponential family**.

I bring this up here because there's a **simple theorem** that specifies the **conjugate prior** for likelihoods that belong to the **exponential family**.

Definition (e.g., Bernardo and Smith, 1994): Given data y_1 (a sample of size 1) and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (marginal) sampling distribution $p(y_1|\theta)$ belongs to the k**-dimensional exponential family** if it can be expressed in the form

$$p(y_1|\theta) = f_1(y_1)\, g_1(\theta)\, \exp\left[\sum_{j=1}^{k} \phi_j(\theta)\, h_j(y_1)\right] \tag{101}$$

Exponential Family (continued)

for $y_1 \in \mathcal{Y}$ and 0 otherwise; if $\mathcal{Y}$ doesn't depend on θ the family is called **regular**.

The vector $[\phi_1(\theta), \ldots, \phi_k(\theta)]$ in (101) is called the **natural parameterization** of the exponential family.

When any single observation (e.g., y_1) has a sampling distribution of the form (101), the **joint distribution** $p(y|\theta)$ of a **sample** $y = (y_1, \ldots, y_n)$ of size n that's conditionally IID from (101) (which also defines, as usual, the **likelihood function** $l(\theta|y)$) will be

$$\begin{aligned} p(y|\theta) &= l(\theta|y) = c \prod_{i=1}^{n} p(y_i|\theta) \\ &= c \left[\prod_{i=1}^{n} f_1(y_i) \right] [g_1(\theta)]^n \exp \left[\sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i) \right]. \end{aligned}$$

This leads to **another way** to define the exponential family: in (101) take $f(y) = \prod_{i=1}^{n} f_1(y_i)$ and $g(\theta) = [g_1(\theta)]^n$ to yield

Definition: Given data $y = (y_1, \ldots, y_n)$ (a conditionally IID sample of size n)

Exponential Family (continued)

and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (joint) sampling distribution $p(y|\theta)$ belongs to the k-**dimensional exponential family** if it can be expressed in the form

$$p(y|\theta) = f(y)\, g(\theta)\, \exp\left[\sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i)\right]. \tag{102}$$

Either way you can see that $\{\sum_{i=1}^n h_1(y_i), \ldots, \sum_{i=1}^n h_k(y_i)\}$ is a set of **sufficient** statistics for θ under this sampling model, because the likelihood $l(\theta|y)$ depends on y only through the values of $\{h_1, \ldots, h_k\}$.

Now here's the **theorem about the conjugate prior**: if the likelihood $l(\theta|y)$ is of the form (102), then in searching for a **conjugate** prior $p(\theta)$ — that is, a prior of the same functional form as the likelihood — you can see directly what will work:

$$p(\theta) = c\, g(\theta)^{\tau_0}\, \exp\left[\sum_{j=1}^{k} \phi_j(\theta)\, \tau_j\right], \tag{103}$$

for some $\tau = (\tau_0, \ldots, \tau_k)$.

Exponential Family (continued)

With this choice the **posterior** for θ will be

$$p(\theta|y) = c\, f(y)\, g(\theta)^{1+\tau_0} \exp\left\{\sum_{j=1}^{k} \phi_j(\theta)\left[\tau_j + \sum_{i=1}^{n} h_j(y)\right]\right\}, \tag{104}$$

which is indeed of the **same form** (in θ) as (102).

Example (1) With $s = \sum_{i=1}^{n} y_i$, recall that the Bernoulli/binomial likelihood can be written

$$\begin{aligned} l(\theta|y) &= c\,\theta^s (1-\theta)^{n-s} \\ &= c\,(1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^s \\ &= c\,(1-\theta)^n \exp\left[s \log\left(\frac{\theta}{1-\theta}\right)\right], \end{aligned} \tag{105}$$

which shows (a) that this sampling distribution is a member of the **exponential family** with $k = 1$, $g(\theta) = (1-\theta)^n$, $\phi_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ (**NB** the natural parameterization, and the basis of **logistic regression**), and $h_1(y_i) = y_i$, and (b) that $\sum_{i=1}^{n} h_1(y_i) = s$ is sufficient for θ.

Then (103) says that the **conjugate prior** for the Bernoulli/binomial likelihood is

$$\begin{aligned} p(\theta) &= c\,(1-\theta)^{n\tau_0} \exp\left[\tau_1 \log\left(\frac{\theta}{1-\theta}\right)\right] \\ &= c\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \text{Beta}(\alpha,\beta) \end{aligned} \tag{106}$$

for some α and β, as we've already seen is **true**.

Example (2) For a setting with $k > 1$, take $\theta = (\mu, \sigma^2)$ with the **Gaussian likelihood**:

$$\begin{aligned} l(\theta|y) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\ &= c\left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right)\right]. \end{aligned} \tag{107}$$

This is **of the form (102)** with $k = 2,\ f(y) = 1, g(\theta) = \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right), \phi_1(\theta) = -\frac{1}{2\sigma^2}, \phi_2(\theta) = \frac{\mu}{\sigma^2}, h_1(y_i) = y_i^2$, and $h_2(y_i) = y_i$, which shows that $[h_1(y) = \sum_{i=1}^n y_i^2, h_2(y) = \sum_{i=1}^n y_i]$ or equivalently $(\bar{y}, s^2)$ is

sufficient for θ.

Some **unpleasant algebra** then demonstrates that an application of the **conjugate prior** theorem (54) in the exponential family leads to (95) as the conjugate prior for the Gaussian likelihood when both μ and σ^2 are unknown.

Example (3) An example of a **non-regular exponential family**: suppose (as in the case study in homework 3 problem 2) that a reasonable model for the data is to take the observed values $(y_i|\theta)$ to be conditionally IID from the **uniform** distribution $U(0,\theta)$ on the interval $(0,\theta)$ for unknown θ:

$$p(y_1|\theta) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{for } 0 < y_1 < \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta} I(0,\theta), \tag{108}$$

where $I(A) = 1$ if A is true and 0 otherwise.

θ in this model is called a **range-restriction** parameter; such parameters are fundamentally different from **location** and **scale** parameters (like the mean μ and variance σ^2 in the $N(\mu,\sigma^2)$ model, respectively) or **shape** parameters (like the degrees of freedom ν in the t_ν model).

(108) is an **example of (102)** with $c = 1, f_1(y) = 1, g_1(\theta) = \frac{1}{\theta}, h_1(y) = 0$, and $\phi_1(\theta) =$ anything you want (e.g., 1), but only when the set $\mathcal{Y} = (0, \theta)$ is taken to depend on θ.

Truncated distributions with **unknown truncation point(s)** also lead to non-regular exponential families; an example would be if You needed to model Your data as like random draws from a (rescaled) $N(\mu, \sigma^2)$ distribution forced to live not on $(-\infty, \infty)$ but on (A, B) with at least one of A and B unknown.

As you'll see in homework 3, inference in non-regular exponential families is **similar** in some respects to the story when the exponential family is regular, but there are some **important differences** too.

2.11 Pros and cons of maximum likelihood.

Strength of **maximum likelihood** as an approach to **parametric inference**:

- Fisher's approach **extends** readily to situations in which the parameter θ is a **vector** of length $k > 1$:

— The **log likelihood** $ll(\theta|y)$ is now a function of the k values $(\theta_1, \ldots, \theta_k)$ for fixed data vector y; in **regular problems** (in which the maximum occurs in the

Pros and Cons of Maximum Likelihood (continued)

interior of the parameter space) the **MLE** can be found by

(a) creating a **system** of k equations in k unknowns, by setting the **(first) partial derivatives** $\frac{\partial}{\partial \theta_j} ll(\theta|y)$ equal to 0, and

(b) **solving** this system, either **analytically** or **numerically**.

Example: With a **conditionally IID Gaussian sampling distribution** for y_i in which the data-generating **mean** μ and **variance** σ^2 are both **unknown**, from equation (94) above the **likelihood** is

$$l(\mu, \sigma^2|y) = c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right]\right\}, \tag{109}$$

and the **log likelihood** is then evidently

$$ll(\mu, \sigma^2|y) = c - n \log \sigma - \frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right]. \tag{110}$$

Q: If I find the **MLE** of $\gamma = \sigma^2$ in this model and you find the **MLE** of $\eta = \sigma$, how should these two estimates be **related**?

It would be **nice** if $\hat{\gamma}_{\text{MLE}} = (\hat{\eta}_{\text{MLE}})^2$; is this **true**?

Pros and Cons of Maximum Likelihood (continued)

To write the **log likelihood** in the γ **parameterization**, I just put γ wherever I see σ^2:

$$ll(\mu, \gamma|y) = c - \frac{n}{2}\log\gamma - \frac{1}{2\gamma}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right] ; \tag{111}$$

similarly, to write the **log likelihood** in the η **parameterization**, I just put η wherever I see σ:

$$ll(\mu, \eta|y) = c - n\log\eta - \frac{1}{2\eta^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right] . \tag{112}$$

The **system** of $k = 2$ equations in μ and γ is then

$$\left\{ \begin{array}{rcl} \frac{\partial}{\partial\mu} ll(\mu, \gamma|y) = -\frac{1}{\gamma} n\,(\mu - \bar{y}) & = & 0 \\ \frac{\partial}{\partial\gamma} ll(\mu, \gamma|y) = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right] & = & 0 \end{array} \right\} , \tag{113}$$

and this has the **solution**

$$\left\{ \mu = \hat{\mu}_{\text{MLE}} = \bar{y}, \quad \gamma = \hat{\gamma}_{\text{MLE}} = \widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \right\} . \tag{114}$$

Pros and Cons of Maximum Likelihood (continued)

Similarly, the **system** of $k = 2$ equations in μ and η is then

$$\left\{ \begin{array}{rcl} \frac{\partial}{\partial \mu} ll(\mu, \eta | y) = -\frac{1}{\eta^2} n \left(\mu - \bar{y}\right) & = & 0 \\ \frac{\partial}{\partial \eta} ll(\mu, \eta | y) = -\frac{n}{\eta} + \frac{1}{\eta^3} \left[n(\mu - \bar{y})^2 + (n-1)s^2 \right] & = & 0 \end{array} \right\}, \tag{115}$$

and this has the **solution**

$$\left\{ \mu = \hat{\mu}_{\text{MLE}} = \bar{y}, \quad \eta = \hat{\eta}_{\text{MLE}} = \hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2} \right\}. \tag{116}$$

So the **answer** to the question above is **nice**: you get the **same estimate** of μ either way, and $\widehat{\sigma^2}_{\text{MLE}} = \left(\hat{\sigma}_{\text{MLE}}\right)^2$.

This is an example of a **general property**, of both the **MLE** and the **posterior mode**, called functional invariance: the **simplest special case** of this property to state says that if $g(\cdot)$ is **invertible** then $\widehat{g(\theta)}_{\text{MLE}} = g\left(\hat{\theta}_{\text{MLE}}\right)$ (there's a **more general version** of functional invariance of the MLE for **all** functions g (not just the invertible ones), but it requires the idea of **profile likelihood**, which is **not important** for Bayesian work).

Pros and Cons of Maximum Likelihood (continued)

— The one-dimensional large-sample result we looked at earlier,

$$\hat{\theta}_{\text{MLE}} \dot{\sim} N\left[\theta, \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)\right], \tag{117}$$

generalizes with $k \geq 1$ to

$$\hat{\theta}_{\text{MLE}} \dot{\sim} N_k\left[\theta, \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)\right], \tag{118}$$

in which $N_k(\mu, \Sigma)$ is the multivariate normal distribution, in k dimensions, with mean vector μ and covariance matrix Σ; here the analogue of $\hat{I}$ in the one-dimensional result is a **matrix** (minus the **Hessian** [matrix of second partial derivatives] of the **log likelihood**, evaluated at the MLE) and what used to be the reciprocal operation in $\hat{I}^{-1}$ when $k = 1$ is now matrix inversion.

Example (continued): Carrying on with the **parameterization** $\gamma = \sigma^2$, the **second partial derivatives** are

$$\begin{aligned} &\frac{\partial^2}{\partial \mu^2} ll(\mu, \gamma | y) = -\frac{n}{\gamma}, \; \frac{\partial^2}{\partial \mu \, \partial \gamma} ll(\mu, \gamma | y) = \frac{n}{\gamma^2}(\mu - \bar{y}) \text{ and} \\ &\frac{\partial^2}{\partial \gamma^2} ll(\mu, \gamma | y) = \frac{n}{2\,\gamma^2} - \frac{1}{\gamma^3}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right], \end{aligned} \tag{119}$$

Pros and Cons of Maximum Likelihood (continued)

so the **information matrix** is

$$\hat{I} = -\left[\begin{array}{cc} \frac{\partial^2}{\partial \mu^2} ll(\mu, \gamma | y) & \frac{\partial^2}{\partial \mu \, \partial \gamma} ll(\mu, \gamma | y) \\ \frac{\partial^2}{\partial \mu \, \partial \gamma} ll(\mu, \gamma | y) & \frac{\partial^2}{\partial \gamma^2} ll(\mu, \gamma | y) \end{array} \right]_{\mu = \hat{\mu}, \gamma = \hat{\gamma}}$$

$$= \left[\begin{array}{cc} \frac{n}{\hat{\gamma}} & 0 \\ 0 & \frac{n}{2 \, \hat{\gamma}^2} \end{array} \right] \tag{120}$$

and the **large-sample estimated covariance matrix** of the MLE vector $\hat{\theta} = (\hat{\mu}, \hat{\gamma})$ (in repeated sampling) is then

$$\hat{V}\left(\hat{\theta}\right) \doteq \hat{I}^{-1} = \left[\begin{array}{cc} \frac{\hat{\gamma}}{n} & 0 \\ 0 & \frac{2 \, \hat{\gamma}^2}{n} \end{array} \right] . \tag{121}$$

What this **means** is that

(a) the **repeated-sampling variance** of $\hat{\mu} = \bar{y}$ is **estimated** to be $\hat{V}(\hat{\mu}) = \hat{V}(\bar{y}) \doteq \frac{\hat{\gamma}}{n} = \frac{\hat{\sigma}^2}{n} = \frac{(n-1)s^2}{n^2}$ (where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the usual repeated-sampling-unbiased **sample variance**) — note that in fact $V(\bar{y}) = \frac{\sigma^2}{n}$

Pros and Cons of Maximum Likelihood (continued)

and $E\left[\frac{(n-1)s^2}{n^2}\right] = \frac{\sigma^2}{n} - \frac{\sigma^2}{n^2}$, so that the **bias** in Fisher's answer is **quite small** $\left(O\left(\frac{1}{n^2}\right)\right)$; Fisher's theory **does not lead to good interval estimates** for μ in **small samples** [it pretends that the **distribution** of $\frac{\bar{y}-\mu}{\hat{\sigma}}$ is **Gaussian** when it's actually **scaled** t], but the t **approaches** the Gaussian as n increases);

(b) the **repeated-sampling covariance** $\hat{C}(\hat{\mu}, \hat{\gamma})$ between $\hat{\mu}$ and $\hat{\sigma^2}$ is **estimated** to be **0** — this is also **approximately correct** (it turns out that $\bar{y}$ and s^2 are **independent**, and therefore **uncorrelated**, in the **Gaussian sampling model**); and

(c) the **repeated-sampling variance** of $\hat{\gamma} = \hat{\sigma}^2$ is **estimated** to be $\hat{V}(\hat{\gamma}) = \frac{2\,\hat{\gamma}^2}{n} = \frac{2\,\hat{\sigma}^4}{n}$ — this is also approximately correct (in repeated sampling in this model $\hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-1}}{n}$, so that $V\left(\hat{\sigma}^2\right) = \frac{2(n-1)\sigma^4}{n^2} \doteq \frac{2\sigma^4}{n}$; Fisher's theory **again does not lead to good interval estimates** for σ^2 in small samples [it pretends that the **distribution** of $\hat{\sigma}^2$ is **Gaussian** when it's actually **scaled** χ^2], but again the χ^2 **approaches** the Gaussian as n increases).

The bottom line is that **maximum likelihood** is a **successful general approach** to **parametric inference** when the sample size n is **large** and

Pros and Cons of Maximum Likelihood (continued)

little or no relevant information, about the unknown θ, external to the data set y is available (in this case **maximum likelihood** and **Bayesian inferential conclusions** will tend to be **similar**).

Disadvantages of maximum likelihood in relation to **Bayesian inference** (this will become clear as we go along):

- With small samples sizes, when the **likelihood function** $l(\theta|y)$ is **skewed** (e.g., often in **hierarchical models** [more on this later]), **maximization** over θ is **not the best technology** for learning about θ; the **Bayesian** approach, which treats the likelihood as if it were a **density**, substitutes **integration** for **maximization** over θ, and this has been found to have **better repeated-sampling properties** (with **diffuse priors**) when n is small.
- The **frequentist** approach encourages thinking of each data set **in isolation**; the **Bayesian** approach explicitly provides a mechanism for **combining information** from **multiple sources**.
- Prediction of **observables** — an activity of **central importance** in **science/statistics** for its role in **model-checking** — is **much easier** from the **Bayesian** point of view.

2.12 References

Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.

Craig PS, Goldstein M, Seheult AH, Smith JA (1997). Constructing partial prior specifications for models of complex physical systems. *The Statistician*, **46**, forthcoming.

Draper D (1995). Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115–147, 233–239.

Draper D, Hodges JS, Mallows CL, Pregibon D (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**, 9–37.

de Finetti B (1930). Funzione caratteristica di un fenomeno aleatorio. *Mem. Acad. Naz. Lincei*, **4**, 86–133.

de Finetti B (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability*, HE Kyburg, Jr., and HE Smokler, eds., New York: Wiley (1980), 93–158.

de Finetti B (1974/5). *Theory of Probability*, **1–2**. New York: Wiley.

Fisher RA (1922). On the mathematical foundations of theoretical statistics.

References (continued)

Philosophical Transactions of the Royal Society of London A, **222**, 309–368.

Fisher RA (1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.

Freedman D, Pisani R, Purves R, Adhikari A (1998). *Statistics*, third edition. New York: Norton.

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Bayesian Data Analysis*, second edition. London: Chapman & Hall.

Hacking I (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Johnson NL, Kotz S (1970). *Distributions in statistics: Continuous univariate distributions*, **1**. New York: Wiley.

Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**, 845–854.

Kadane JB, Wolfson LJ (1997). Experiences in elicitation. *The Statistician*, **46**, forthcoming.

Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990).

References (continued)

The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953–1955 (with editorial comment, 1995–1997).

Laplace PS (1774). Mémoire sur la probabilité des causes par les évenements. *Mémoires de l'Academie des Sciences de Paris*, **6**, 621–656. English translation in 1986 as "Memoir on the probability of the causes of events," with an introduction by SM Stigler, *Statistical Science*, **1**, 359–378.

O'Hagan A (1997). Eliciting expert beliefs in substantial practical applications. *The Statistician*, **46**, forthcoming.

Samaniego FJ, Reneau DM (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*, **89**, 947–957.

Tierney L, Kadane JB (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.