Bayesian Hierarchical Modeling

2: Optimal Prior Distribution Specification

David Draper

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

Short Course (Day 3)
University of Reading (UK)

An Example, to Fix Ideas

Case Study 1. (Krnjajić, Kottas, Draper [KKD] 2008): *In-home geriatric assessment (IHGA)*. In an **experiment** conducted in the **1980s** (Hendriksen et al. 1984), **572 elderly people**, **representative** of $\mathcal{P} =$ {all **non-institutionalized elderly people** in **Denmark**}, were **randomized**, **287** to a **control** (C) group (who received **standard health care**) and **285** to a **treatment** (T) group (who received **standard care plus IHGA**: a kind of **preventive medicine** in which each person's **medical** and **social needs** were assessed and acted upon **individually**).

One **important outcome** was the **number of hospitalizations** during the **two-year** life of the study:

	Number of Hospitalizations						
Group	0	1	...	k	n	Mean	SD
Control	n_{C0}	n_{C1}	...	n_{Ck}	$n_C = 287$	$\bar{y}_C$	s_C
Treatment	n_{T0}	n_{T1}	...	n_{Tk}	$n_T = 285$	$\bar{y}_T$	s_T

Let μ_C and μ_T be the **mean hospitalization rates** (per two years) in $\mathcal{P}$ under the C and T **conditions**, respectively.

Here are **four statistical questions** that **arose** from **this study**:

The Four Principal Statistical Activities

Q_1: Was the **mean number of hospitalizations per two years** in the IHGA group **different from** that in **control** by an **amount** that was **large** in **practical** terms? [**description** involving $\left(\frac{\bar{y}_T - \bar{y}_C}{\bar{y}_C}\right)$]

Q_2: Did IHGA **(causally) change** the **mean number of hospitalizations per two years** by an **amount** that was **large** in **statistical** terms? [**inference** about $\left(\frac{\mu_T - \mu_C}{\mu_C}\right)$]

Q_3: On the **basis** of **this study**, how **accurately** can You **predict** the **total decrease** in **hospitalizations** over a period of N years if **IHGA** were **implemented throughout Denmark**? **[prediction]**

Q_4: On the **basis** of **this study**, is the **decision** to **implement IHGA** throughout Denmark **optimal** from a **cost-benefit** point of view? **[decision-making]**

These questions **encompass** almost all of the **discipline** of **statistics**: **describing** a **data set** D, **generalizing outward inferentially** from D, **predicting new data** D^*, and helping people **make decisions** in the **presence** of **uncertainty** (I include **sampling/experimental design** under **decision-making**; **omitted**: data **quality assurance (QA)**, ...).

An Axiomatization of Statistics

1 **(definition) Statistics** is the study of **uncertainty**: how to **measure it well**, and how to **make good choices** in the face of it.

2 **(definition) Uncertainty** is a state of **incomplete information** about something of interest to **You** (Good, 1950: a **generic person** wishing to **reason sensibly** in the presence of **uncertainty**).

3 **(axiom) (Your uncertainty** about) **"Something of interest to You"** can always be **expressed** in terms of **propositions**: **true/false** statements $A, B, \ldots$

Examples: You may be **uncertain** about the **truth status** of

- $A =$ (**Hillary Clinton** will be **elected U.S. President** in **2016**), or
- $B =$ (the **in-hospital mortality rate** for patients at **hospital** H admitted in **calendar 2010** with a principal diagnosis of **heart attack** was **between 5% and 25%**).

4 **(implication)** It **follows** from 1–3 that **statistics** concerns **Your information** (**NOT** Your **beliefs**) about $A, B, \ldots$

Axiomatization (continued)

5 **(axiom)** But **Your information** cannot be **assessed** in a **vacuum**: all such **assessments** must be made **relative to (conditional on)** Your **background assumptions** and **judgments** about **how the world works** vis à vis $A, B, \ldots$.

6 **(axiom)** These **assumptions** and **judgments**, which are themselves a form of **information**, can always be **expressed** in a **set** $\mathcal{B}$ of **background propositions**, all of which **You believe** to be **true**.

Examples of $\mathcal{B}$:

- In the **IHGA study**, based on the **experimental design**, $\mathcal{B}$ would include the **propositions**

(**Subjects** were **representative of [like a random sample from]** $\mathcal{P}$),

(**Subjects** were **randomized** into one of two groups, **treatment** (**standard care + IHGA**) or **control** (**standard care**)).

7 **(definition)** Call the **"something of interest to You"** θ; in **applications** θ is often a **vector** (or **matrix**, or **array**) of **real numbers**, but **in principle** it could be **almost anything** (a **function**,

Axiomatization (continued)

an **image** of the **surface of Mars**, a **phylogenetic tree**, ...).

IHGA example: θ = **mean relative decrease** $\left(\frac{\mu_T - \mu_C}{\mu_C}\right)$ in **hospitalization rate** in $\mathcal{P}$.

8 **(axiom)** There will typically be an **information source (data set)** D that You judge to be **relevant** to **decreasing** Your uncertainty about θ; in **applications** D is often again a **vector** (or **matrix**, or **array**) of **real numbers**, but **in principle** it too could be **almost anything** (a **movie**, the **words** in a **book**, ...).

9 **(implication)** The **presence** of D creates a **dichotomy**:

- **Your information** about θ {**internal, external**} to D.

(People often talk about a **different dichotomy**: **Your information** about θ {**before, after**} D arrives **(prior, posterior)**, but **temporal considerations** are actually **irrelevant**.)

10 **(implication)** It **follows** from 1–9 that **statistics** concerns itself principally with **five things** (omitted: **description**, **data QA**, ...):

(1) **Quantifying Your information** about θ **internal** to D (given $\mathcal{B}$), and doing so **well** (this term is **not yet defined**);

Foundational Question

(2) **Quantifying Your information** about θ **external** to D (given $\mathcal{B}$), and doing so **well**;

(3) **Combining** these two **information sources** (and doing so **well**) to create a **summary** of **Your uncertainty** about θ (given $\mathcal{B}$) that includes **all available information** You judge to be **relevant** (this is **inference**);

and using **all Your information** about θ (given $\mathcal{B}$) to make

(4) **Predictions** about **future** data values D^* and

(5) **Decisions** about how to **act sensibly**, even though **Your information** about θ may be **incomplete**.

Foundational question: How should these tasks be **accomplished**?

This topic will be **continued** in **Day 3** of this **short course** in **much greater depth**; **for now**, let's **focus** on **task (2)** above — **optimal prior distribution specification** — **first in general**, and **then** with **particular application** to **hierarchical models**.

Optimal Prior Distribution Specification

Sometimes the **prior distribution** is **uniquely specified** by **problem context**; this **most often occurs** when **little or no information** about θ **external** to the **data set** D is **available**.

Example 1. The **simplest situation** is **when** the **unknown** θ **takes on** a **finite number** of **distinct values** $(v_1, \ldots, v_k)$; **these** can **live** in **any space they want**, but **they** can **always** be **mapped** onto k **distinct places** on the **real number line**.

If the **totality** of **Your information** about θ **external** to D is **captured** by the **single proposition**

$$B = \{\theta \text{ takes on a } \textbf{finite} \text{ number of } \textbf{distinct} \text{ values } (v_1, \ldots, v_k)\}, \quad (1)$$

from which it would **follow** that $B = \mathcal{B}$, then **intuition suggests** that the **only possible prior specification** is

$$p(\theta = v_j|\mathcal{B}) = \frac{1}{k} \text{ for all } j = 1, \ldots, k \text{ and 0 otherwise}. \quad (2)$$

Laplace used this **repeatedly** in the **late 1700s without thinking** it **needed any justification**; much **later Keynes** (1921) called it the **Principle of Indifference**.

The Principle of Indifference is Actually a Theorem

However, thinking of **probability** as an **expression** of **Your information** about θ, it's **not just a Principle**, it's a **Theorem**, which (in **light** of its **history**) should be **given** a **new name**; I **propose calling it** the

Uniform Prior Distribution Theorem: **If** the **totality** of **Your information** about θ **external** to D is **captured** by the **single proposition**

$$\mathcal{B} = \{\theta \text{ takes on a } \textbf{finite} \text{ number of } \textbf{distinct} \text{ values } (v_1, \ldots, v_k)\}\,, \quad (3)$$

then Your **only possible logically-internally-consistent prior specification** is

$$p(\theta = v_j|\mathcal{B}) = \frac{1}{k} \text{ for all } j = 1, \ldots, k \text{ and 0 otherwise}\,. \quad (4)$$

One-sentence Proof 1 (by **contradiction**): **Suppose** that $p(\theta = v_j|\mathcal{B})$ is **not constant**; then this **violates** the **assumption** about the **totality** of **Your information**.

More elaborate Proof 2 (by **group invariance**, with a **nod** to **Einstein's relativity proofs**): *[short course web page: Jaynes (2003), pages 37–40]*

Optimal Diffuse Priors

Example 2. The **next simplest situation** is **when** the **unknown** θ **takes on values** in an **interval** on the **real line**, of the **form** (a, b) or $[a, b)$ or $(a, b]$ or $[a, b]$ for some **real numbers** a and b with $-\infty < a < b < \infty$ (**I'll** just **talk** about the **interval** (a, b) from **now on**).

If the **totality** of **Your information** about θ **external** to D is **captured** by the **single proposition**

$$B = \{\{\textbf{Possible values for } \theta\} = \{\text{the } \textbf{interval } (a, b)\}, a < b\}, \quad (5)$$

from which it would **again follow** that $B = \mathcal{B}$, then **intuition again suggests** that the **only possible prior specification** is

$$p(\theta|\mathcal{B}) = \textbf{Uniform}(a, b). \quad (6)$$

This intuition is **based** on the **following "proof" sketch**:

- the **interval** (a, b) can be approximated with $(k + 1)$ **equidistant discrete values** $\left(a, a + \frac{(b-a)}{k}, a + \frac{2(b-a)}{k}, \ldots, b\right)$;
- use the **Uniform Prior Distribution Theorem** above to **conclude** that **all** of **these values** must be **equally likely** (with the **given background information**); and

Optimal Diffuse Priors (continued)

- **pass** to the **limit** as $k \to \infty$.

However, You can **readily see this isn't completely satisfying**, through **simple examples** like the **following**:

Suppose that Your **background knowledge base** $\mathcal{B}$ **about** an **unknown** θ **says only** that θ can **take on all values from 1 to 10**.

This is **equivalent** to the **statement** that $\eta = \frac{1}{\theta}$ can **take on all values from** $\frac{1}{10} = 0.1$ **to 1**.

But a **Uniform prior** on θ **doesn't imply** a **Uniform prior** on η, and **vice versa**, as **You can see** either from the **Change-of-Variables formula** or by **simulation**:

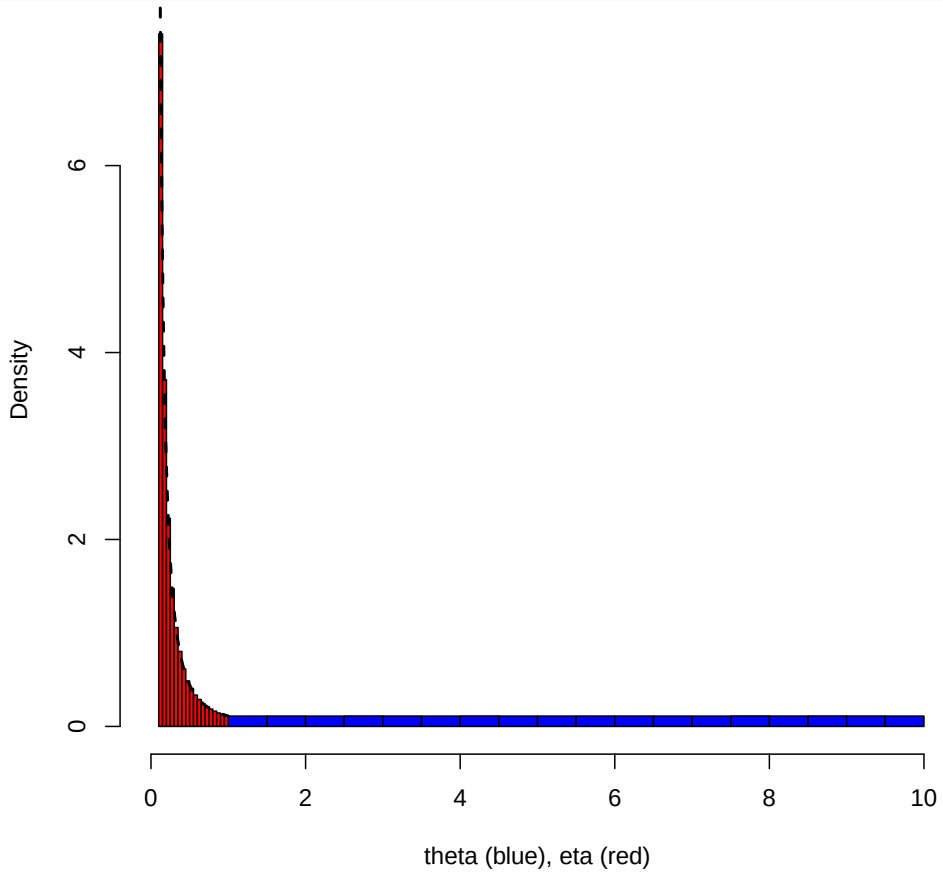
An **easy calculation shows** that $(\theta|\mathcal{B}) \sim$ **Uniform**$(1, 10)$ **iff** $p(\eta|\mathcal{B}) = \frac{1}{9\,\eta^2} \mathsf{I}(0.1, 1)$, and **simulation arrives** at the **same answer**.

The `R` **code** on the **next page stores 1,000,000 simulated draws** from the **Uniform**$(1, 10)$ **distribution** in `theta.star`, **computes** `eta.star` as the **reciprocal** of the `theta.star` **values, plots density-scale histograms** of `theta.star` and `eta.star` on the **same graph**, and **superimposes** the **density** $\frac{1}{9\,\eta^2} \mathsf{I}(0.1, 1)$ on the **histogram** for `eta.star`.

Optimal Diffuse Priors (continued)

```
n <- 1000000
theta.star <- runif( n, 1, 10 )
eta.star <- 1 / theta.star
theta.histogram <- hist( theta.star, plot = F )
eta.histogram <- hist( eta.star, plot = F )
xlim <- c( 0, 10 )
ylim <- range( 0, theta.histogram$intensities,
  eta.histogram$intensities)
# pdf( 'uniform-non-uniform.pdf' )
plot( theta.histogram, xlim = xlim, ylim = ylim,
  col = 'blue', xlab = 'theta (blue), eta (red)', freq = F,
  main = '')
par( new = F )
plot( eta.histogram, xlim = xlim, ylim = ylim,
  xaxt = 'n', yaxt = 'n', col = 'red', add = T,
  freq = F )
eta.grid <- seq( 0.1, 1, length = 500 )
lines( eta.grid, 1 / ( 9 * eta.grid^2 ), lty = 2, lwd = 2 )
# dev.off( )
```

Optimal Diffuse Priors (continued)



Thus, in **settings** in which **all You know** about θ is that it **lives continuously** in (a, b), there is **no unique prior** that **both captures this information** (and **no other information**) and **behaves reasonably** under **arbitrary monotonic transformation** (as **we'll see below**, this **problem bothered Fisher greatly**, and **turned him against** the **Bayesian paradigm**).

Ignorance Priors

Consider the **AMI mortality case study** from the **Day 1 Lecture Notes (Part 2)**, in which **exchangeability** (arising **directly** from **problem context**) led to the **model**

$$\begin{aligned} (\theta|\mathcal{B}) &\sim p(\theta|\mathcal{B}) \\ (y_i|\theta\,\mathcal{B}) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta) \end{aligned} \tag{7}$$

for $i = 1, \ldots, n$; **here** y_i is **1** if **AMI patient** i **died within 30 days of admission (0 otherwise)** and θ is the **underlying mortality rate** in the **population** of **patients**, **similar** to **those** in the **study**, to which **You're interested** in **generalizing**.

In this **case study, exchangeability uniquely identifies** the **sampling distribution (Bernoulli)**, but **what about** the **prior** $p(\theta|\mathcal{B})$?

In the **Day 1 Lecture Notes (Part 2)**, we **looked** for a **suitable member** of the **conjugate-prior Beta family**; will **some member** of **this family** be **suitable** as an **ignorance prior**, for **settings** in which **You have little information** about θ **external** to the **data set** y?

Ignorance Priors Via Group Invariance

One possibility — in fact, the **one favored independently** by **both Bayes** and **Laplace** — is of course the **Uniform = Beta**$(1, 1)$ **distribution**, but **we saw above** that **this** is **not invariant** to **monotonic transformation** of θ.

Three other possibilities have been **developed**, each **based** on its own **sensible-sounding principle**.

- **With** the $\mathcal{B}$ of **this problem**, **Jaynes** (2003) *[short course web page: pages 372–386]* uses a **group-invariance argument** to **show** that the **optimal ignorance prior** is

$$p(\theta|\mathcal{B}) \propto \frac{1}{\theta(1-\theta)} = \textbf{Beta}(0, 0)\,. \tag{8}$$

This prior is **improper**, but **leads** to a **proper posterior** for **any data set** in which **at least one 0** and **at least one 1** are **observed**.

However, to **complicate things**, a **different ignorance prior** was **developed** by **Jeffreys** (1939) — with a **different invariance calculation** — to **rebut** an **argument put forward** by **Fisher** (1922) **against** the **Bayes/Laplace prior**:

Fisher Weighs In

"(A) If You're completely ignorant about a success probability θ, aren't You also completely ignorant about any monotone function of θ, such as $\eta = \log \frac{\theta}{1-\theta}$?

"(B) You can't have a Uniform prior on θ and η simultaneously.

"(C) Therefore the entire Bayesian paradigm is rubbish."

Lindley (1954) **qualitative rebuttal**: **if**, on **Your information base**, θ could be **anywhere** in $(0, 1)$, with **no value favored** over **another**, then it's **absurd to say** that **You're equally ignorant** about a **monotone function** such as $\lambda = \theta^{100}$ — **You know for sure** that λ is **close to 0**.

- **Jeffreys** (1939) **quantitative rebuttal**: OK, **Fisher**, You **give me Your sampling distribution** $p(y|\theta\,\mathcal{B})$ (in **any problem** with a **univariate** θ), and **I'll give You** an **ignorance prior** that's **invariant** to **monotone increasing transformation**, namely

$$p(\theta|\mathcal{B}) \propto \sqrt{I(\theta)}\,, \quad \textbf{where} \quad I(\theta) = -E_{(y|\theta\,\mathcal{B})}\left[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta\,\mathcal{B})\right]. \quad (9)$$

Kullback-Leibler Divergence

In the **Bernoulli heart attack mortality example** above, the **Jeffreys idea gives** the **prior**

$$p(\theta|\mathcal{B}) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \textbf{Beta}(\tfrac{1}{2}, \tfrac{1}{2}) \,, \tag{10}$$

which **differs** from **both** the **Bayes/Laplace prior Beta**$(1, 1)$ and the **Jaynes prior Beta**$(0, 0)$.

- To **make things** even **worse**, **Bernardo** (1979) has **also articulated** a **reasonable-looking principle** that can be **used** to **derive ignorance priors**: **any prior** that's **far away** from the **posterior it leads to** (once the **data** have **arrived**) must have **relatively low information content**.

This **requires** a **notion** of **distance** between **two distributions**.

Definition. The **Kullback-Leibler (KL) divergence** of a **continuous density** $q(\theta|\mathcal{B})$ on $\Re^k$ from **another continuous density** $p(\theta|\mathcal{B})$ on $\Re^k$ — **also known as** the **relative entropy** of q to p — is

$$KL(q||p) = \int_{\Re^k} p(\theta|\mathcal{B}) \log \frac{p(\theta|\mathcal{B})}{q(\theta|\mathcal{B})} \, d\theta \,. \tag{11}$$

Bayes/Laplace, Jaynes, Jeffreys, Bernardo

(**This** is **not a distance metric**, because it's **asymmetric** in q and p, but $\frac{1}{2}\left[KL(q||p) + KL(p||q)\right]$ is a **proper distance metric**.)

Bernardo's idea is to **define** the **reference prior** $p(\theta|\mathcal{B})$ as **that distribution** which **maximizes** the **expectation** (over the **sampling distribution** $p(D|\theta\,\mathcal{B})$) of the **KL divergence** of $p(\theta|\mathcal{B})$ **from** the **posterior distribution** $p(\theta|D\,\mathcal{B})$.

Fact: when $k = 1$, **reference priors** and **Jeffreys priors coincide**, but this is **not necessarily true** for $k > 1$.

So: in the **AMI mortality example, principles** put forth by {**Bayes/Laplace, Jaynes, Jeffreys, Bernardo**} yield the **ignorance priors Beta**$(0, 0)$, **Beta**$(\frac{1}{2}, \frac{1}{2})$ and **Beta**$(1, 1)$.

Remembering that the **prior sample size** in the **Bernoulli sampling model** with the **conjugate** Beta(α, β) prior is $(\alpha + \beta)$, the **prior sample sizes** of the **three ignorance priors** above are **0**, **1** and **2** (**respectively**).

When Principles Collide

Another qualitative rebuttal to Fisher, **based on** those **prior sample sizes** (**many people** have **made this point**, including **Draper** (2009)):

Fisher's point would have **real force if nobody ever collected any data, because** in that case **(posterior = prior)** and **uncertainty** in **how** to **specify** a **diffuse prior can really matter**; but **with even** a **modest amount of data**, the **posteriors** with **prior sample sizes** of **0**, **1** and **2** **will** essentially **coincide**.

When Principles Collide: We **now have three different reasonable-looking principles** for **creating diffuse (ignorance, low-information) priors**: **Jeffreys's Transformation-Invariance**, **Jaynes's Group-Invariance** and **Bernardo's Reference-Prior**.

Q: **What** should You **do** when **reasonable diffuse-prior principles** lead to **priors** that are **different enough** to **matter**?

A: Here's **yet another Principle** that **(greatly) helps**.

The Calibration Principle

Calibration Principle: In **model specification**, it **helps** to **know something** about **how often** {the **methods** You're **using** to choose one **model** over **another**} **get** the **right answer**, and **this** can be **ascertained** by (a) **creating simulation environments** (structurally **similar** to the **setup** of the **scientific problem** You're **currently solving**) in which **You know what the right answer is** and (b) seeing **how often Your methods recover** known **truth**.

The **reasoning** behind the **Calibration Principle** is as follows:

(axiom) You want to **help positively advance** the **course of science**, and **repeatedly getting the wrong answer** runs **counter** to this desire.

(remark) There's **nothing** in the **Bayesian paradigm** to **prevent** You from making **one or both** of the following **mistakes** — (a) choosing $p(D|\theta\,\mathcal{B})$ **badly**; (b) inserting {**strong information** about θ **external** to D} into the **modeling process** that turns out **after the fact** to **have been** (badly) **out of step with reality** — and **repeatedly** doing this **violates the axiom** above.

Reasoning Behind the Calibration Principle

(remark) Paying attention to **calibration** is a **natural activity** from the **frequentist point of view**, but a **desire** to be **well-calibrated** can be **given** an **entirely Bayesian justification** via **decision theory**:

Taking a **broader perspective** over **Your career**, not just within any **single attempt** to solve an **inferential/predictive problem** in collaboration with **other investigators**, Your **desire** to **take part positively** in the **progress of science** can be **quantified** in a **utility function** that **incorporates** a **bonus** for being **well-calibrated**, and in this **context** (Draper, 2013) **calibration-monitoring** emerges as a **natural and inevitable Bayesian activity**.

This **seems to be** a **new idea**: **logical consistency** justifies **Bayesian uncertainty assessment** but **does not provide guidance** on **model specification**; if You accept the **Calibration Principle**, **some** of this **guidance** is **provided**, via **Bayesian decision theory**, through a **desire** on **Your part** to **pay attention to how often You get the right answer**, which is a **central scientific activity**.

A Calibration-Checking Case Study

I **bring up** the **Calibration Principle here** (and **it will come up again** several times **later**) because it **provides** an **answer** to the **question above** about **how** to **resolve conflicts** among **competing diffuse priors**: **You** can **create** a **simulation study**, **just like** the **problem** of **interest to You** but **in which You know** the **parameter values**, and **see which prior does** the **best job** of **correctly recovering known truth**.

Here's an **example** — **Case Study 1 (continued):** *In-home geriatric assessment (IHGA).*

This was the **actual distribution** of **number** of **hospitalizations** over a **two-year period**:

Group	Number of Hospitalizations 0	1	2	3	4	5	6	7	*n*	Mean	SD
Control	138	77	46	12	8	4	0	2	287	0.944	1.24
Treatment	147	83	37	13	3	1	1	0	285	0.768	1.01

(**We'll examine** this **data set** in a **number** of **other ways later**.)

Since the **outcome** in **each group** is a **count** of the **number of occurrences** of a **fairly rare phenomenon**, Your **initial impulse**

Extra-Poisson Variability/Unexplained Heterogeneity

would be to **fit** a **model** in which the **control observations** are **IID Poisson**(λ_C) and the **treatment values** are **independently IID Poisson**(λ_T) (**this** is **called** a **fixed-effects Poisson (FEP) model**).

However, the **Poisson**(λ) **distribution** has **mean** and **variance** both **equal** to λ; in **other words**, for **this distribution** the **variance-to-mean-ratio (VTMR)** is **1**.

Here the **control-** and **treatment-group VTMR values** are $\frac{1.24^2}{0.944} \doteq \mathbf{1.63}$ and $\frac{1.01^2}{0.768} \doteq \mathbf{1.33}$, **respectively**, so the **fixed-effects Poisson model** is **inadequate**.

Count data sets with **VTMR** > 1 are **said** to **exhibit extra-Poisson variability** or **unexplained heterogeneity**.

Consider just the **treatment values** for **now**, and **drop** the T **subscript**.

A **useful way** to **rewrite** the **IID Poisson**(λ_T) **model** for the **observations** $(y_1, \ldots, y_n)$, when **(as is the case here) little** is **known** about **hospitalization rates under IHGA external** to the **data set**, is

Making the Model More Realistic

$$\begin{aligned} (y_i|\lambda_i\,\mathcal{B}) &\overset{\text{indep}}{\sim} \textbf{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 \\ (\beta_0|\mathcal{B}) &\sim \textbf{diffuse} \end{aligned} \tag{12}$$

This **moves** toward **scientific realism**, in the **first line** of (12), by **allowing** each **elderly person** to have **her/his own** λ, but this would **create** an **unworkable model** with n **observations** and n **parameters**; the **second line** of the **model (unrealistically) reduces** the **number** of **parameters** from n to 1, by **pretending** that **everybody** in the **treatment group** has the **same underlying rate** λ of **hospitalization**.

In reality it's **far more reasonable** to **think** that **each person** has **his/her own underlying rate** of **hospitalization** that depends on **baseline health status**, **age**, and **various other things**.

If we **had** k **such covariates**, the **second** and **third lines** would be

$$\begin{aligned} \log(\lambda_i) &= \beta_0 + \sum_{j=1}^{k} \beta_j\, x_{ij} \\ (\beta|\mathcal{B}) &\sim \textbf{diffuse} \quad (\beta = (\beta_0, \ldots, \beta_k))\,. \end{aligned}$$

Random-Effects Poisson Model

Now **Hendriksen** (the **study's author**) **forgot to measure** (or **at least** to **report on**) any **covariates**, so the **best we can do** is to **lump** all of these **other latent (unobserved) predictor variables together** into a kind of **"error" term** e_i, as **follows**:

$$\begin{aligned} (y_i|\lambda_i\,\mathcal{B}) &\overset{\text{indep}}{\sim} \textbf{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + e_i \\ (e_i|\sigma\,\mathcal{B}) &\overset{\text{IID}}{\sim} N(0,\sigma^2) \\ (\beta_0\,\sigma|\mathcal{B}) &\sim \textbf{diffuse}\,. \end{aligned} \tag{13}$$

This is **referred to** as a **random-effects Poisson (REP) model**; the **latent variables** e_i are **also called random effects**.

The **Gaussian choice** for the **random-effects distribution** is **conventional**, **not dictated** by the **science** of the **problem** (although **if there were a lot** of **unobserved predictors hidden inside** the e_i, their **weighted sum** would be **close to Gaussian** by the **Central Limit Theorem**).

Model (13) is an **expansion** of the **earlier FEP model** (12), **because** You can **obtain model** (12) from (13) by **setting** $\sigma = 0$,

Unexplained Heterogeneity → Latent Variables

whereas with **(13)** we're **letting** σ **vary** and **learning about it from the data**.

The **addition** of the **random effects** e_i to the **model** is **one way** to **address** the **extra-Poisson variability**: **this model** could **also** be **called** a **lognormal mixture of Poisson distributions**, because **it's as if each person's** λ is **drawn** from a **lognormal distribution** and then **her/his number** of **hospitalizations** y is **drawn** from a **Poisson distribution** with **his/her** λ, and this **hierarchical mixing process** will **make** the **variance** of y **bigger than its mean**.

This is an **example** of a **valuable contemporary modeling approach** — when **unexplained heterogeneity** is **present**, **You** can **use latent variables** to **at least properly quantify it** (of course, **this** is **also** an **old idea**: **that's what** the **"error" term** in **linear regression** is **doing**).

However, a **new challenge** now arises: **how** to **make operational** the **statement** "$(\beta_0\, \sigma|\mathcal{B}) \sim$ **diffuse**".

In **this model** β_0 and σ are **clearly independent given** $\mathcal{B}$ — $p(\beta_0\, \sigma|\mathcal{B}) = p(\beta_0|\mathcal{B})\, p(\sigma|\mathcal{B})$ — so the **new questions** become:

Diffuse Priors on $(-\infty, \infty)$ and $(0, \infty)$

how to **specify** a **diffuse prior** in a **principled way** on (a) $(-\infty, \infty)$ (e.g., a **location parameter** such as β_0) and (b) $(0, \infty)$ (e.g., a **scale parameter** such as σ)?

Example 4. If the **totality** of **Your information** about an **unknown** θ is that **it can take on any value continuously** in $(-\infty, \infty)$, then an **arbitrary shift (left or right)** in the **location** of θ would **leave** the **problem invariant**.

Jaynes (2003) *[short course web page: pages 372–386]* **shows** that the **only prior satisfying** this **invariance property** is **Uniform**$(-\infty, \infty)$.

This is **improper**, but **can be approximated** to **arbitrary accuracy** with a **proper** $N(0, \sigma_\theta^2)$ **prior** with **huge variance** σ_θ^2, or **equivalently tiny precision** $\tau_\theta = \frac{1}{\sigma_\theta^2}$.

(**This** is **also** the **Jeffreys/reference prior** for **this problem**.)

Example 5. If the **totality** of **Your information** about an **unknown** θ is that **it can take on any value continuously** in $(0, \infty)$, then **several cases arise**, and **many of them lead** to **different diffuse priors**

Poisson and Location-Scale Problems

based on **different principles**, e.g.:

(1) **Jaynes** (2003) **shows** that if **You're observing** a **counting process** over **time** that **You're willing** to **model** as **Poisson**, then the **only prior** on the **Poisson rate** $\lambda \in (0, \infty)$ that's **invariant** to **arbitrary rescaling** of **time** by a **constant positive multiple** is $p(\lambda|\mathcal{B}) \propto \frac{1}{\lambda}$, but the **Jeffreys** and **reference priors** for the **IID Poisson model** are **both** $p(\lambda|\mathcal{B}) \propto \frac{1}{\sqrt{\lambda}}$.

(2) As **Jaynes** (2003) **points out**, if the **problem context** (with **real-valued observations** y_i) **leads You** to a **sampling model** of the **form** $(y_i|\mu\,\sigma\,\mathcal{B}) \overset{\text{IID}}{\sim} p(y_i|\mu\,\sigma\,\mathcal{B})$ for **some** p in which μ and σ are **location** and **scale parameters**, **respectively**, and in which **(therefore)** the **problem** should be **invariant** to **arbitrary left-right shifts** and **arbitrary positive rescaling**, the **only prior** that **expresses** this **information** (and **no other external information**) is of the **form** $p(\mu\,\sigma|\mathcal{B}) \propto \frac{1}{\sigma}$ (**this** is **also** the **Jeffreys prior** if $p(\mu\,\sigma|\mathcal{B})$ is **thought of** as $p(\sigma|\mathcal{B})\,p(\mu|\sigma\,\mathcal{B})$ and **Jeffreys's basic idea** is **applied separately** to **each term** in the **product**, and **it's also** the **reference prior** in **this setting**).

Hierarchical Models With Random Effects

However, if **instead** (but **still** with **real-valued observations** y_i) a **unique origin 0** has **special meaning**, but the **problem** is **invariant** to **arbitrary positive rescaling** (e.g., **temperature measured** in **degrees Kelvin**), then the **only prior** that **expresses this information** (and **no other external information**) is of the **form** $p(\mu\,\sigma|\mathcal{B}) \propto \frac{1}{\sigma^2}$.

(3) A **class** of **problems in which** there's **clear uncertainty** about **how** to **arrive** at a **good diffuse prior** is **hierarchical models** with **random effects**, such as the **REP model** above.

$$\begin{aligned} (y_i|\lambda_i\,\mathcal{B}) &\overset{\text{indep}}{\sim} \textbf{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + e_i \\ (e_i|\sigma\,\mathcal{B}) &\overset{\text{IID}}{\sim} N(0,\sigma^2) \\ (\beta_0\,\sigma|\mathcal{B}) &\sim \textbf{diffuse}\,. \end{aligned} \tag{14}$$

You can **factor** the **prior** as $p(\beta_0\,\sigma|\mathcal{B}) = p(\sigma|\mathcal{B})\,p(\beta_0|\sigma\,\mathcal{B})$ and **take** $p(\beta_0|\sigma\,\mathcal{B}) = p(\beta_0|\mathcal{B})$, **since** the **intent** is to **achieve diffuseness** with **both** β_0 and σ.

But what should **You take** for $p(\beta_0|\mathcal{B})$ and $p(\sigma|\mathcal{B})$, if **scientific context** (as it **does here**) implies **diffuseness**?

Diffuse Priors for Hierarchical Models

Simulation studies (Draper, 2013) with **this model** have **shown two things**:

- The **Normal**$(0, \sigma^2_{\beta_0})$ **prior** for β_0 with **small** $\tau_{\beta_0} = \frac{1}{\sigma^2_{\beta_0}}$ produces **good calibration** across a **wide range** of τ_{β_0} **values**, as **long** as τ_{β_0} is **small enough**.

For example, $(\beta_0|\mathcal{B}) \sim N(0, \tau_{\beta_0} = 10^{-6})$ **yields** a **Gaussian prior** with a **mean** of **0** and an **SD** of **1,000**; this **will be** a **flat prior** if the **likelihood** for β_0 is **concentrated** (say) between **−1** and **+1**, but **not** if it's **concentrated** (say) between **−5,000** and **+5,000**.

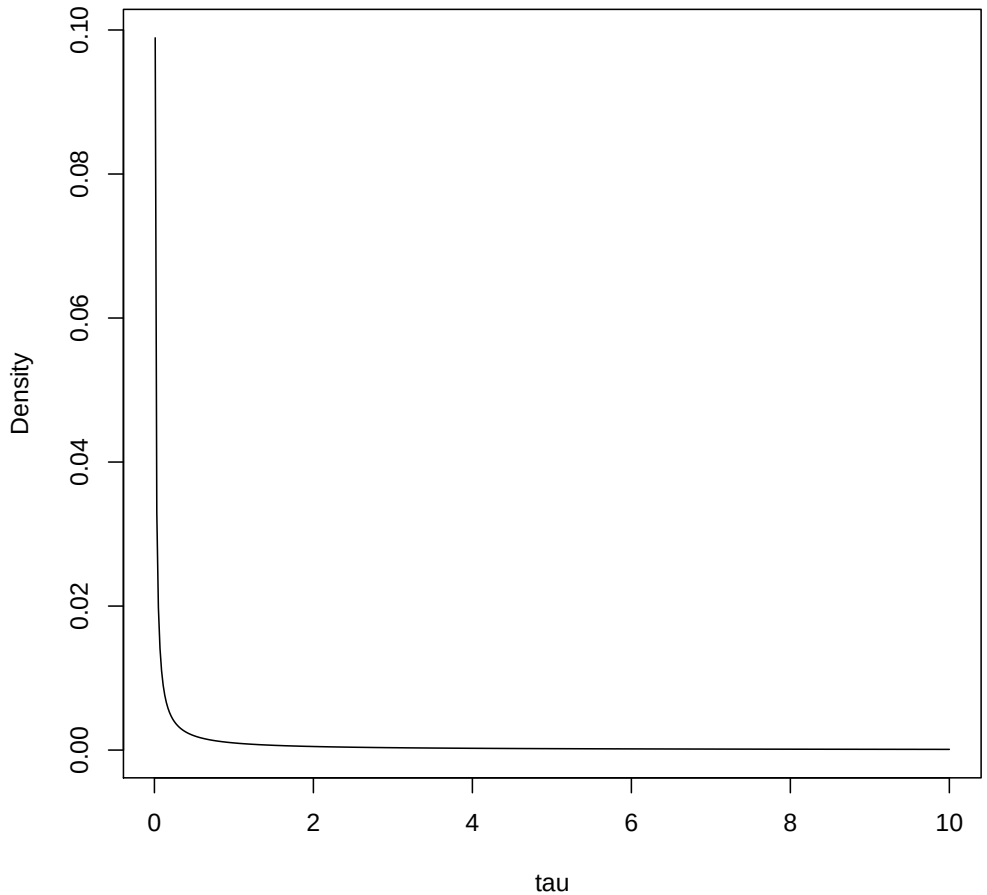
In the **IHGA data set**, **all** of the λ **values**, which are **approximately** e^{β_0}, are **not far from 1**, which **makes** β_0 **close to 0**, so the $N(0, \tau_{\beta_0} = 10^{-6})$ **prior** will be **diffuse**.

- **Much more care** is **required** in **specifying** a **diffuse prior** for σ, or **equivalently** for $\tau = \frac{1}{\sigma^2}$, to **achieve good calibration**.

The **choice** $(\tau|\mathcal{B}) \sim \Gamma(\epsilon, \epsilon)$, for a **small value** of ϵ **such as** 10^{-3}, has been **heavily popularized** by the `WinBUGS` people (**for instance**, it's **still frequently used** in the **examples volumes distributed** with `WinBUGS`).

$\Gamma(\epsilon, \epsilon)$ Prior on τ

This prior has a **mean** for τ of **1** and an **SD** of $\sqrt{10^3} \doteq$ **32**, and **does look flat** over a **broad range** of $\Re$, but (to **stay proper**) it **goes to** ∞ as $\tau \to 0$:

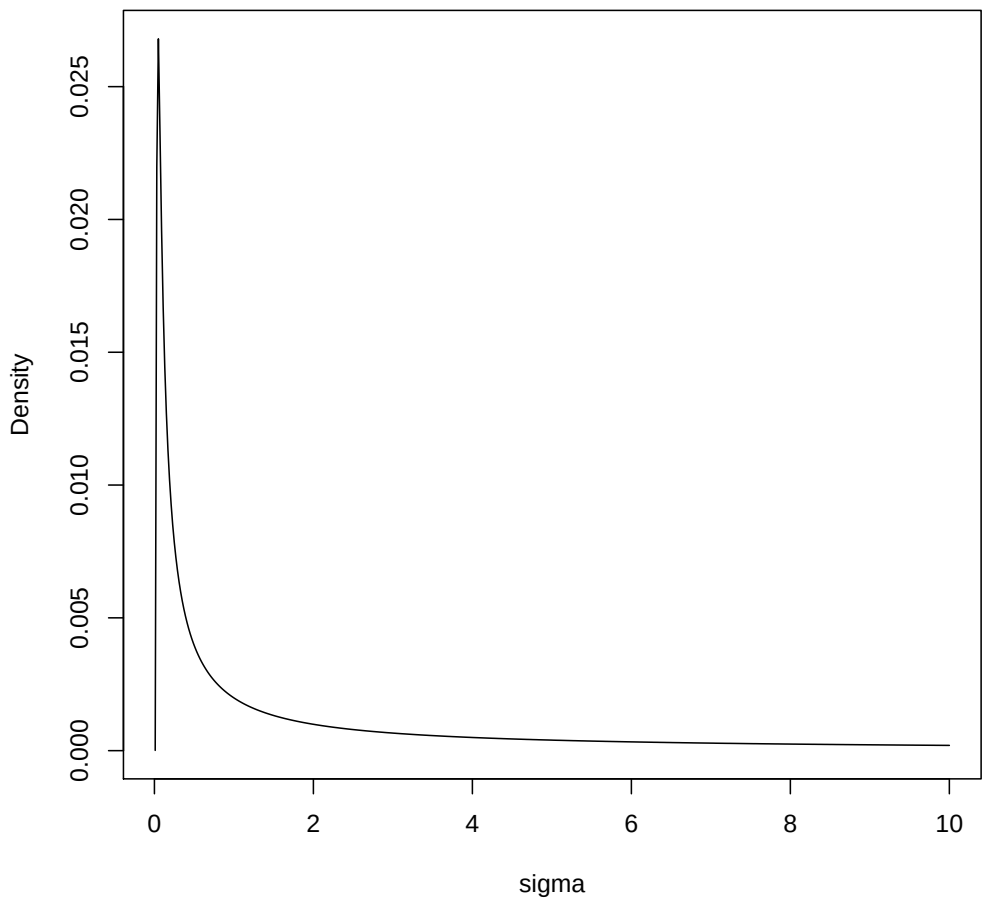


Now the **question** is: **what prior** does this **induce** on $\sigma = \frac{1}{\sqrt{\tau}}$?

The Induced Prior on σ

The **change-of-variables formula** **yields** the following **ugly-looking density expression** for σ:

$$p(\sigma|\mathcal{B}) = \frac{2\,\epsilon^{\epsilon}}{\Gamma(\epsilon)}\sigma^{-(2\epsilon+1)}\exp\left(-\frac{\epsilon}{\sigma^2}\right) . \tag{15}$$



Uniform Prior on σ

Surprisingly, the **tail** of the $\Gamma(\epsilon, \epsilon)$ **prior** on τ is **so heavy** that the **induced prior** on $\sigma = \frac{1}{\sqrt{\tau}}$ **also has** a **spike** near **0**.

From the **plot** on the **previous page**, You can **see what can go wrong** with **this prior**:

if the **likelihood** for σ is **concentrated on small positive values**, then the $\Gamma(\epsilon, \epsilon)$ **prior** on τ will be **unintentionally highly informative**, making the **posterior** on σ **even more concentrated** on **small positive values** than the **likelihood**.

- It **turns out** that a **diffuse prior** that has **good calibration properties** is $(\sigma|\mathcal{B}) \sim$ **Uniform**$(0, C)$, where $C > 0$ is a **constant chosen large enough** to **avoid artificially truncating** the **posterior distribution**.

Ironically, given the **bad results noted above** from the $\Gamma(\epsilon, \epsilon)$ **prior** on τ, it's **still useful** to **employ it** in a **two-step procedure**: You can **determine** a **good value** for C by

(a) **examining** the **preliminary posterior** on σ **produced** by the $\Gamma(\epsilon, \epsilon)$ **prior** on τ, and

Hierarchical Models With Random Effects

(b) **choosing** C — **based on** the **preliminary posterior** in (a) — for a **second** (and **definitive**) **posterior calculation** with the **Uniform**$(0, C)$ **prior** on σ.

[R code on short course web page, illustrating a calibration study comparing these two priors]

Another instance of **calibration sensitivity** to the **form** of the **diffuse prior** occurs with **hierarchical random-effects models** in **meta-analysis**.

Case Study: **Meta-analysis** of **effects** of **low-dose aspirin** on **heart attacks (Lecture Notes, Day 2 Part 1)**.

Recall that the **usual Gaussian random-effects meta-analysis model** in **this study** is

$$\begin{aligned} (\mu\,\sigma|\mathcal{B}) &\sim p(\mu\,\sigma|\mathcal{B}) && \text{(prior)} \\ (\theta_i|\mu\,\sigma\,\mathcal{B}) &\overset{\text{IID}}{\sim} N\left(\mu, \sigma^2\right) && \text{(underlying effects)} \\ (y_i|\theta_i\,\mathcal{B}) &\overset{\text{indep}}{\sim} N(\theta_i, V_i) && \text{(data)}\ . \end{aligned}$$

Well-Calibrated Diffuse Priors in Meta-Analysis

$$
\begin{array}{rcll}
(\mu\,\sigma|\mathcal{B}) & \sim & p(\mu\,\sigma|\mathcal{B}) & \text{(prior)} \\
(\theta_i|\mu\,\sigma\,\mathcal{B}) & \overset{\text{IID}}{\sim} & N\left(\mu,\sigma^2\right) & \text{(underlying effects)} \\
(y_i|\theta_i\,\mathcal{B}) & \overset{\text{indep}}{\sim} & N(\theta_i, V_i) & \text{(data) .}
\end{array}
\tag{16}
$$

Prior specification. The **top level** of **(16)** is **where** the **prior distribution** on the **regression parameters** from the **middle level** is **specified**; **what** should **You choose** for $p(\mu\,\sigma|\mathcal{B})$?

If — as was **true** in **this meta-analysis** — **scientific context** indicates **little information** about (μ, σ) **external** to the **data set, You need** a **good diffuse prior**, where (by the **Calibration Principle**) **"good"** means **well-calibrated**.

[Browne and Draper (2006) on course web page, including Gelman (2006)]

Simulation studies (Gelman 2006, Draper 2013) have **yielded results similar** to **those** with the **REP model above**:

Diffuse Priors in Meta-Analysis (continued)

- The **Normal**$(0, \sigma_\mu^2)$ **prior** for μ with **small** $\tau_\mu = \frac{1}{\sigma_\mu^2}$ produces **good calibration** across a **wide range** of τ_μ **values**, as **long** as τ_μ is **small enough**; here $(\mu|\mathcal{B}) \sim N(0, \tau_\mu = 10^{-6})$ again **works well**.

- **Once** again, **much more care** is **required** in **specifying** a **diffuse prior** for σ, or **equivalently** for $\tau = \frac{1}{\sigma^2}$, to **achieve good calibration**.

The **same Uniform**$(0, C)$ **prior** on σ (with C **chosen well** in the **same way**) **works** here **again**, and Gelman (2006) **has shown** that a **half-**t **prior** (see **his comment** for **details**) also **gives good results**.

[R code on short course web page, illustrating a calibration study comparing a variety of priors]