An Introduction to the Imprecise Dirichlet Model for Multinomial Data

Jean-Marc Bernard
CNRS ESA 8069 & Université Paris 5

Tutorial for the
Third International Symposium on
Imprecise Probabilities and Their Applications
(ISIPTA'03)

Lugano, Switzerland
14 July 2003

Outline

1. INTRODUCTION

The IDM in brief

☐ **Model for statistical inference**
Proposed by Walley (1996), generalizes the IBM (Walley, 1991).
Inference from data $\boldsymbol{x} = (x_1, \ldots, x_K)$, categorized in K categories C, with unknown chances $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$.

☐ **Prior ignorance** about $\boldsymbol{\theta}$, K and C

☐ **Imprecise probability model**, prior uncertainty about $\boldsymbol{\theta}$ expressed by a set of Dirichlet's.

☐ **Posterior uncertainty** about $\boldsymbol{\theta}|\boldsymbol{x}$ then described by a set of (updated) Dirichlet's.

☐ **Imprecise U&L probabilities**, interpreted as reasonable betting rates *for* or *against* an event.

☐ **Generalizes Bayesian inference**, prior/post. uncertainty described by a *single* Dirichlet.

☐ **Satisfies desirable principles** for inferences from prior ignorance, contrarily to alternative frequentist and objective Bayesian approaches.

Aims of this tutorial

☐ **Review objective Bayesian inference** based on Dirichlet distributions

☐ **Presentation of the IDM**

☐ **Review inferences produced by the IDM**
First simple cases.
Then more complex/recent applications.

☐ **Comparison of inferences** from the IDM, objective Bayesian models, and frequentist approach.
Review desirable principles for objective inference.

☐ **Arguments supporting specific values for s**, the single hyper-parameter of the IDM.

☐ **Mention some yet unsolved problems**

☐ **Scope/Interest of the IDM**

The "Bag of marbles" example

☐ **"Bag of marbles" problems** (Walley, 1996)

- "I have . . . a closed bag of coloured marbles. I intend to shake the bag, to reach into it and to draw out one marble. What is the probability that I will draw a red marble?"

- "Suppose that we draw a sequence of marbles whose colours are (in order):

  blue, *green*, *blue*, *blue*, *green*, *red*.

  What conclusions can you reach about the probability of drawing a red marble on a future trial?"

☐ **Caracteristics of this problem**

- Prediction problem: future observations?

- Prior ignorance about the chances $\boldsymbol{\theta}$ of the various colours (objective inference goal)

- Set C and number K of colours is partly arbitrary and may vary as data items are observed. There is prior ignorance about both C and K.

Desirable principles

□ **Symmetry principle (SP)**
Prior uncertainty should be invariant *w.r.t.* permutations of categories.

□ **Embedding principle (EP)**
Prior uncertainty should not depend on refinements or coarsenings of categories.

□ **Representation invariance principle (RIP)**
Inferences should not depend on refinements or coarsenings of categories.

□ **Stopping rule principle (SRP)**
Inferences should not depend on data that might have occurred, *i.e.* on why the data gathering stopped.

□ **Likelihood principle (LP)**
Inferences should depend on the data through the likelihood function only.

□ **Coherence** requirements, avoiding sure loss, when considering several inferences.

Inference from multinomial data

□ Multinomial data

- Infinite population, elements categorized in K categories from set $C = \{c_1, \ldots, c_K\}$.
- Unknown chances $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, $\sum_k \theta_k = 1$.
- Data are a random sample from the population, of size n, yielding counts $\boldsymbol{x} = (x_1, \ldots, x_K)$, with $\sum_k x_k = n$.

□ Multinomial likelihood

$$P(\boldsymbol{x}|\boldsymbol{\theta}) \;\propto\; \theta_1^{x_1} \ldots \theta_K^{x_K} \tag{1}$$

□ General problem: Make inferences about

- the unknown chances $\boldsymbol{\theta}$
- some derived parameter of interest $\lambda = g(\boldsymbol{\theta})$
- n' future observations

Usual approaches

☐ Two objective approaches

- Frequentist: significance tests, confidence limits and intervals (Fisher, Neyman & Pearson)
- objective Bayesian ("non-informative", *etc.*, priors) (*e.g.* Jeffreys, 1961)

☐ Difficulties of frequentist methods

- Do not obey LP
- Ad-hoc and/or asymptotic solutions to the problem of nuisance parameters

☐ Difficulties of Bayesian methods

Several priors proposed for prior ignorance, but none satisfies all desirable principles.

- Inferences often depend on C and/or K
- Some solutions violate LP (Jeffreys, 1946)
- Inferences about various derived parameters can be incoherent (Berger, Bernardo, 1992)

2. DIRICHLET DISTRIBUTIONS

Dirichlet distribution

□ **Dirichlet density**
Vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \sim \textit{Diri}(s\boldsymbol{t})$, $\boldsymbol{\theta} \in \mathcal{S}$ with $s > 0$ and $\boldsymbol{t} = (t_1, \ldots, t_K) \in \mathcal{S}^\star$,

$$h(\boldsymbol{\theta}) \;\propto\; \theta_1^{st_1} \ldots \theta_K^{st_K - 1} \tag{2}$$

($\mathcal{S}$ and $\mathcal{S}^\star$ are the closed/open simplices.)

□ **Parameterization** (usual one) in terms of the *strengths* $\boldsymbol{\alpha} = s\boldsymbol{t} = (\alpha_1, \ldots, \alpha_K)$

□ **Generalization** of Beta distribution ($K = 2$)

$$(\theta_1, \theta_2) \sim \textit{Diri}(\alpha_1, \alpha_2) = \textit{Beta}(\alpha_1, \alpha_2)$$

□ **Basic properties**

- Expectations given by the *relative strengths*:

$$E(\theta_k) \;=\; t_k \tag{3}$$

- Hyper-parameter s determines the dispersion of the distribution.

Examples of Dirichlet's

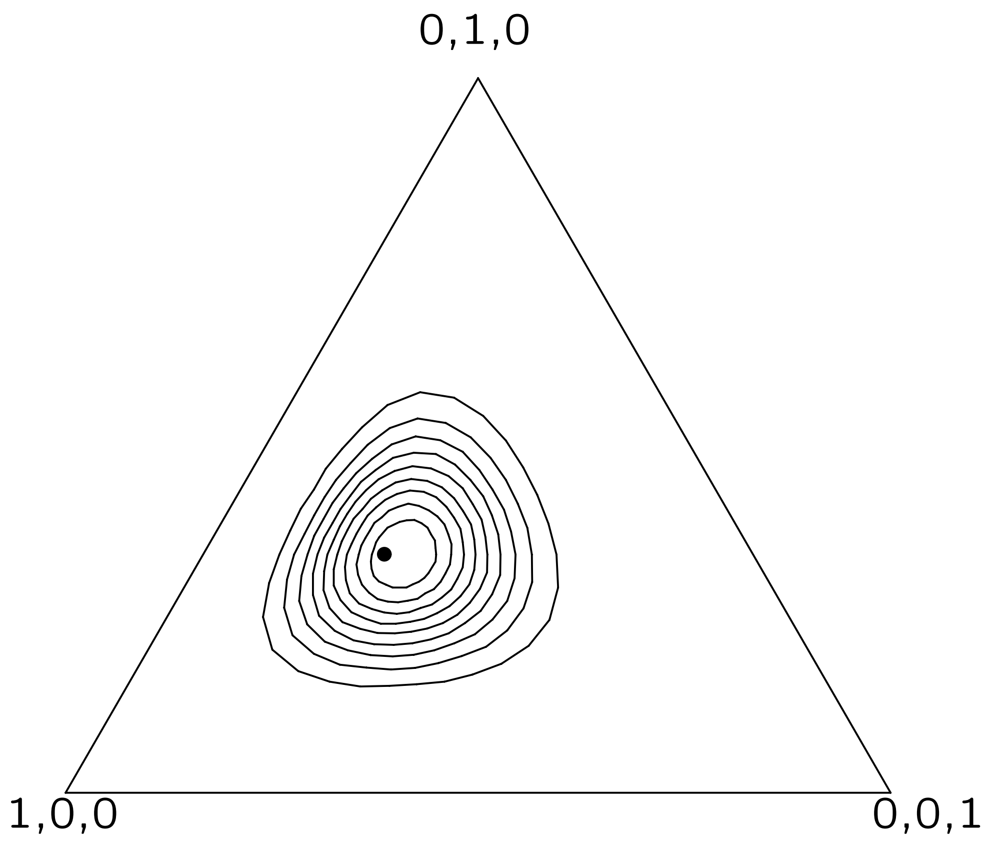
□ Example 1

$$Diri(1, 1, \ldots, 1) \text{ is uniform on } \mathcal{S}$$

□ Example 2

$$(\theta_1, \theta_2, \theta_3) \sim Diri(10, 8, 6)$$

□ Highest density contours $[100\%, 90\%, \ldots, 10\%]$



Properties of the Dirichlet

General properties given on an example.
Assume $(\theta_1, \ldots, \theta_5) \sim \mathit{Diri}(\alpha_1, \ldots, \alpha_5)$. Then,

□ **Pooling property**

$$(\theta_1, \theta_{234}, \theta_5) \sim \mathit{Diri}(\alpha_1, \alpha_{234}, \alpha_5),$$

where pooling categories amounts to add corresponding chances and strengths.

□ **Tree T underlying C**
Consider any tree T underlying the set of categories C. Then, the pooling property implies that

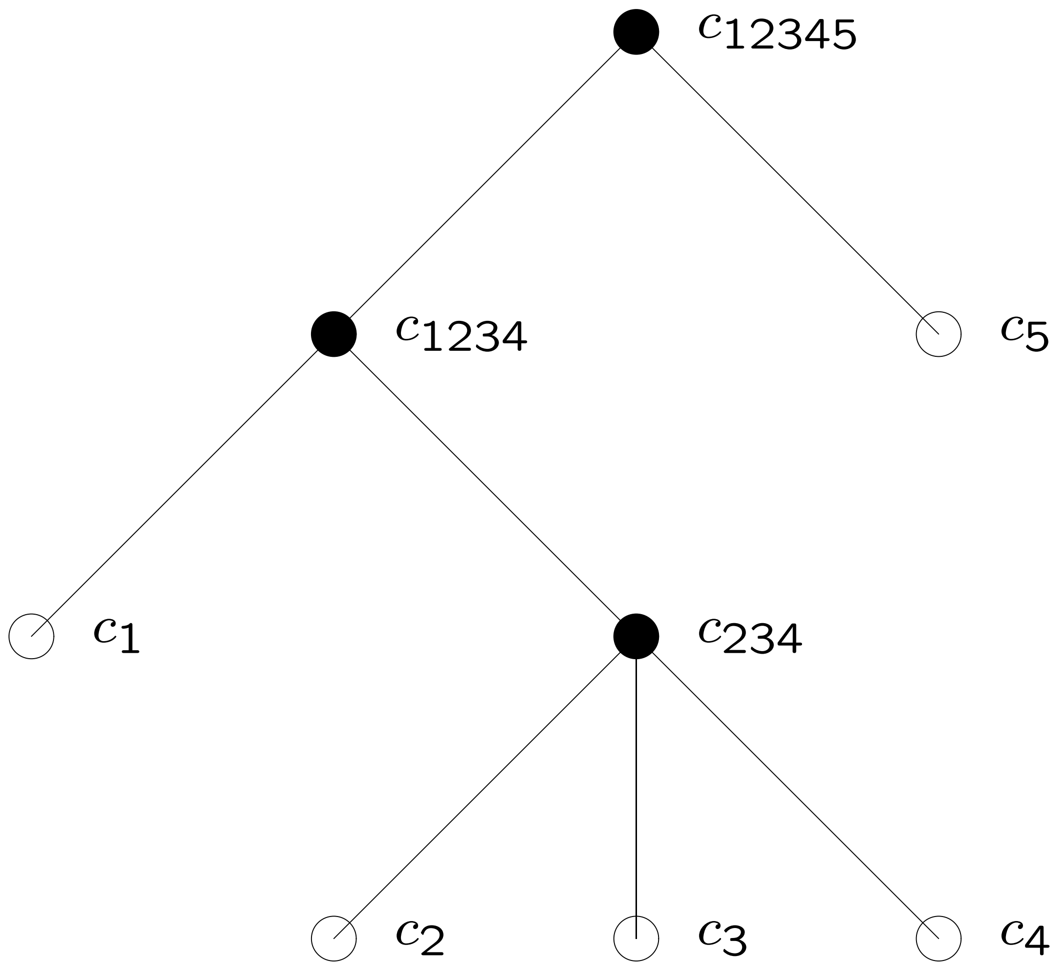
$$\boldsymbol{\theta}_T \sim \mathit{Diri}(\boldsymbol{\alpha}_T)$$

□ **Restriction property**

$$(\theta_2^{234}, \theta_3^{234}, \theta_4^{234}) \sim \mathit{Diri}(\alpha_2, \alpha_3, \alpha_4),$$

where $\theta_2^{234} = \theta_2/\theta_{234}$, *etc.*, are conditional chances.

Tree representation of categories



“Node-cutting” a Dirichlet

☐ **Cutting a tree T** at node c amounts to spliting T into two sub-trees

- $\overline{T}$, where c is a terminal-leaf
- $\underline{T}$, where c is the root

☐ **Corresponding chances and strengths**

- Chances θ_k are normalized
- Strengths α_k remain unchanged

☐ **Theorem** (Bernard, 1997)
Consider any tree T, cut at any node c, giving two sub-trees $\overline{T}$ and $\underline{T}$, then

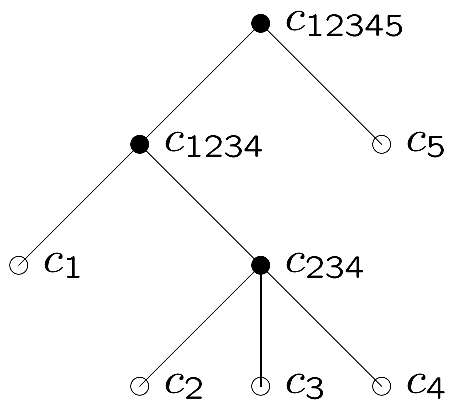
$$\begin{array}{rcl} \boldsymbol{\theta}_{\overline{T}} & \sim & Diri(\boldsymbol{\alpha}_{\overline{T}}) \\ \boldsymbol{\theta}_{\underline{T}} & \sim & Diri(\boldsymbol{\alpha}_{\underline{T}}) \\ \boldsymbol{\theta}_{\overline{T}} & \perp\!\!\!\perp & \boldsymbol{\theta}_{\underline{T}} \end{array}$$

See also Connor, Mosimann, 1969; Darroch, Ratcliff, 1971; Fang, Kotz, Ng, 1990.

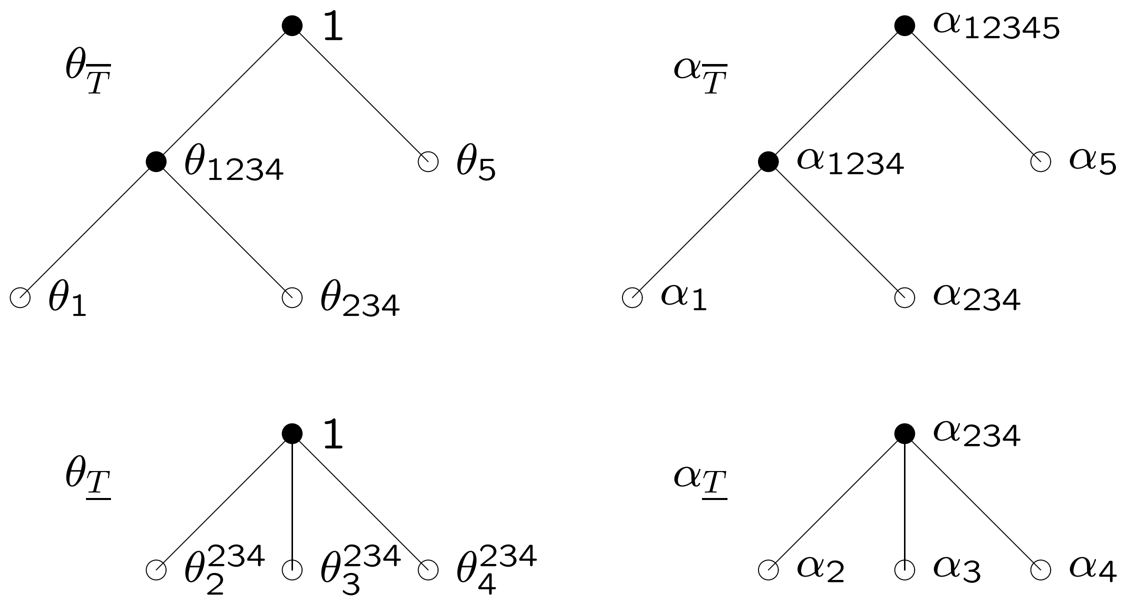
☐ **Key** to computations of the Dirichlet.

“Node-cutting” a Dirichlet (contd)

□ Set C and underlying tree T



□ Cut at node c_{234}



3. THE BAYESIAN APPROACH

Conjugate Bayesian inference

□ Dirichlet prior

Prior uncertainty about $\boldsymbol{\theta}$ is expressed by

$$\boldsymbol{\theta} \;\sim\; \mathit{Diri}(s\boldsymbol{t})$$

with hyper-parameters, s, the *total prior strength*, and $\boldsymbol{t} = (t_1, \ldots, t_K)$, with $t_k > 0$, $\sum_k t_k = 1$ ($\boldsymbol{t}$ belongs to the K-dimensional unit simplex $\mathcal{S}^\star(1, K)$). We call $\alpha_k = st_k$ the *prior strength* of c_k.

Prior expectations

$$E(\theta_k) \;=\; t_k,$$

□ Dirichlet posterior

Posterior uncertainty about $\boldsymbol{\theta}|\boldsymbol{x}$ is expressed by

$$\boldsymbol{\theta}|\boldsymbol{x} \;\sim\; \mathit{Diri}(\boldsymbol{x} + s\boldsymbol{t})$$

Posterior expectations

$$E(\theta_k|\boldsymbol{x}) \;=\; \frac{x_k + s_k}{n + s} = \frac{nf_k + st_k}{n + s}$$

The objective Bayesian approach

□ Priors proposed for objective inference

Idea: $\boldsymbol{\alpha}$ expressing prior ignorance about $\boldsymbol{\theta}$ (Kass & Wasserman, 1996)

Almost all proposed solutions for fixed n are *symmetric* Dirichlet priors, *i.e.* $t_k = 1/K$:

- Haldane (1948): $\alpha_k = 0$ $(s = 0)$
- Perks (1947): $\alpha_k = \frac{1}{K}$ $(s = 1)$
- Jeffreys (1946, 1961): $\alpha_k = \frac{1}{2}$ $(s = K/2)$
- Bayes-Laplace: $\alpha_k = 1$ $(s = K)$
- Berger-Bernardo *reference priors*

□ Difficulties of objective Bayesian approach

None of these solutions simultaneously satisfies all desirable principles for prior ignorance:

- no SP: all except Haldane
- no RIP & EP: all except Haldane
- no LP & SRP: Jeffreys, Berger-Bernardo

4. IMPRECISE DIRICHLET MODEL

Prior and posterior IDM

□ Prior IDM

The prior IDM(s) is defined as the set $\mathcal{M}_0$ of all Dirichlet distributions on $\boldsymbol{\theta}$ with a fixed total prior strength $s > 0$:

$$\mathcal{M}_0 = \{\mathit{Diri}(s\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{S}^\star\} \tag{4}$$

□ Updating

Each Dirichlet distribution on $\boldsymbol{\theta}$ in the set $\mathcal{M}_0$ is updated into another Dirichlet on $\boldsymbol{\theta}|\boldsymbol{x}$, using Bayes' theorem.

This procedure guarantees the *coherence* of inferences (Walley, 1991, Thm 7.8.1).

□ Posterior IDM

Posterior uncertainty about $\boldsymbol{\theta}$ is expressed by the set

$$\mathcal{M}_n = \{\mathit{Diri}(\boldsymbol{x} + s\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{S}^\star\}. \tag{5}$$

Upper and lower probabilities

□ Prior U&L probabilities

Consider event B relative to $\boldsymbol{\theta}$, and $P_{s\boldsymbol{t}}(B)$ the prior probability obtained from the distribution *Diri*$(s\boldsymbol{t})$ in $\mathcal{M}_0$.

Prior uncertainty about B is expressed by

$$\underline{P}(B) \quad \text{and} \quad \overline{P}(B),$$

obtained by min-/maximization of $P_{s\boldsymbol{t}}(B)$ *w.r.t.* $\boldsymbol{t} \in \mathcal{S}^{\star}(1, K)$.

□ Posterior U&L probabilities

Denote $P_{s\boldsymbol{t}}(B|\boldsymbol{x})$ the posterior probability of B obtained from the prior *Diri*$(s\boldsymbol{t})$ in $\mathcal{M}_0$, *i.e.* the posterior *Diri*$(\boldsymbol{x} + s\boldsymbol{t})$ in $\mathcal{M}_n$.

Posterior uncertainty about B is expressed by

$$\underline{P}(B|\boldsymbol{x}) \quad \text{and} \quad \overline{P}(B|\boldsymbol{x}),$$

obtained by min-/maximization of $P_{s\boldsymbol{t}}(B|\boldsymbol{x})$ *w.r.t.* $\boldsymbol{t} \in \mathcal{S}^{\star}(1, K)$.

Posterior inferences about $\lambda = g(\boldsymbol{\theta})$

□ **Derived parameter of interest**

$$\lambda = g(\boldsymbol{\theta}) \;\; = \;\; \begin{cases} \theta_k \\ \sum_k y_k \theta_k \\ \theta_i/\theta_j \\ \textit{etc.} \end{cases}$$

Posterior inferences about λ can be summarized by

□ **U&L expectations**

$$\underline{E}(\lambda|\boldsymbol{x}) \quad \text{and} \quad \overline{E}(\lambda|\boldsymbol{x}),$$

obtained by min-/maximization of $E_{s\boldsymbol{t}}(\lambda|\boldsymbol{x})$ *w.r.t.* $\boldsymbol{t} \in \mathcal{S}^\star(1, K)$,

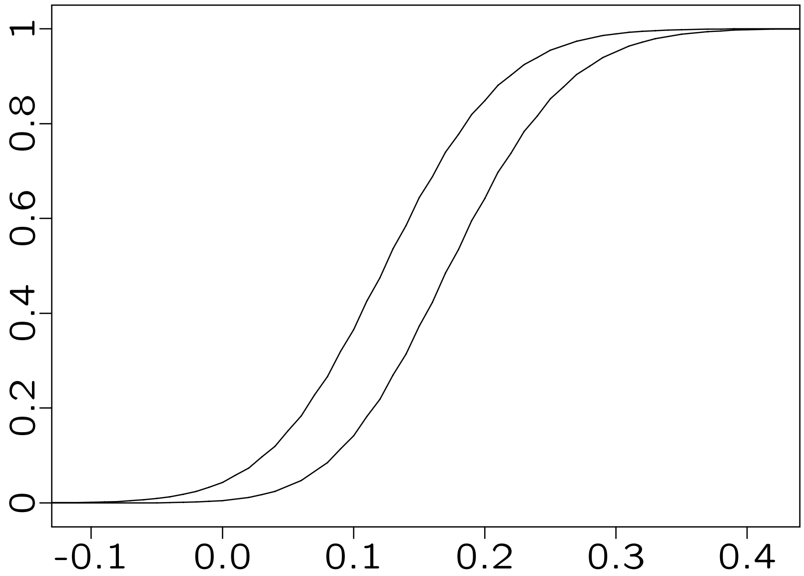
□ **U&L cumulative distribution fonctions (cdf)**

$$\underline{F}(u|\boldsymbol{x}) = \underline{P}(\lambda \le u|\boldsymbol{x}) \quad \text{and} \quad \overline{F}(u|\boldsymbol{x}) = \overline{P}(\lambda \le u|\boldsymbol{x}).$$

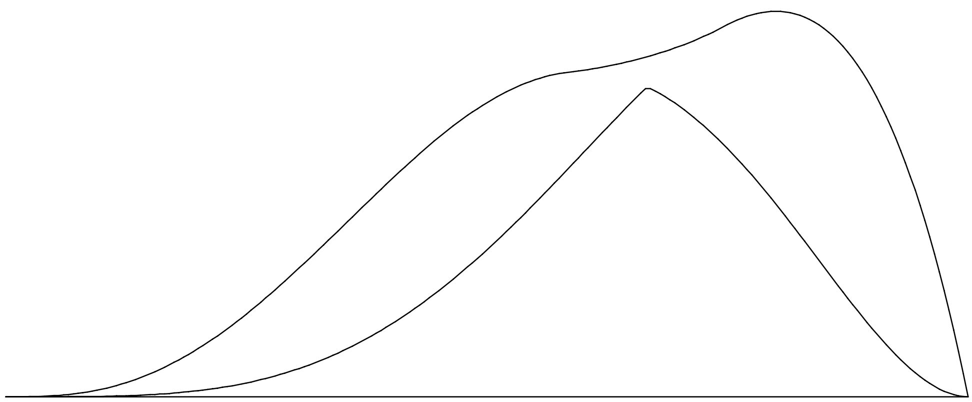
□ **Conjecture:** The two min-/maximization problems above have the same solution, in general, or for some class of functions $g(.)$ to be found?

Examples of U&L df's and cdf's

□ **U&L cdf's**, $\lambda = \sum_k y_k \theta_k$



□ **U&L df's**, $\lambda = \theta_k$



Inferences about θ_k from the IDM

□ **Prior U&L expectations and cdf's**

Expectations

$$\underline{E}(\theta_k) = 0 \quad \text{and} \quad \overline{E}(\theta_k) = 1$$

Cdf's

$$\begin{aligned}
\underline{P}(\theta_k \leq u) &= P(Beta(s, 0) \leq u) \\
\overline{P}(\theta_k \leq u) &= P(Beta(0, s) \leq u)
\end{aligned}$$

□ **Posterior U&L expectations and cdf's**

Expectations

$$\underline{E}(\theta_k|\boldsymbol{x}) = \frac{x_k}{n+s} \quad \text{and} \quad \overline{E}(\theta_k|\boldsymbol{x}) = \frac{x_k + s}{n+s}$$

Cdf's

$$\begin{aligned}
\underline{P}(\theta_k \leq u|\boldsymbol{x}) &= P(Beta(x_k + s, n - x_k) \leq u) \\
\overline{P}(\theta_k \leq u|\boldsymbol{x}) &= P(Beta(x_k, n - x_k + s) \leq u)
\end{aligned}$$

□ **Optimization** attained for $t_k \to 0$ or $t_k \to 1$.
Equivalent to:

Haldane + s extreme observations.

Hyper-parameter s

□ Interpretations of s

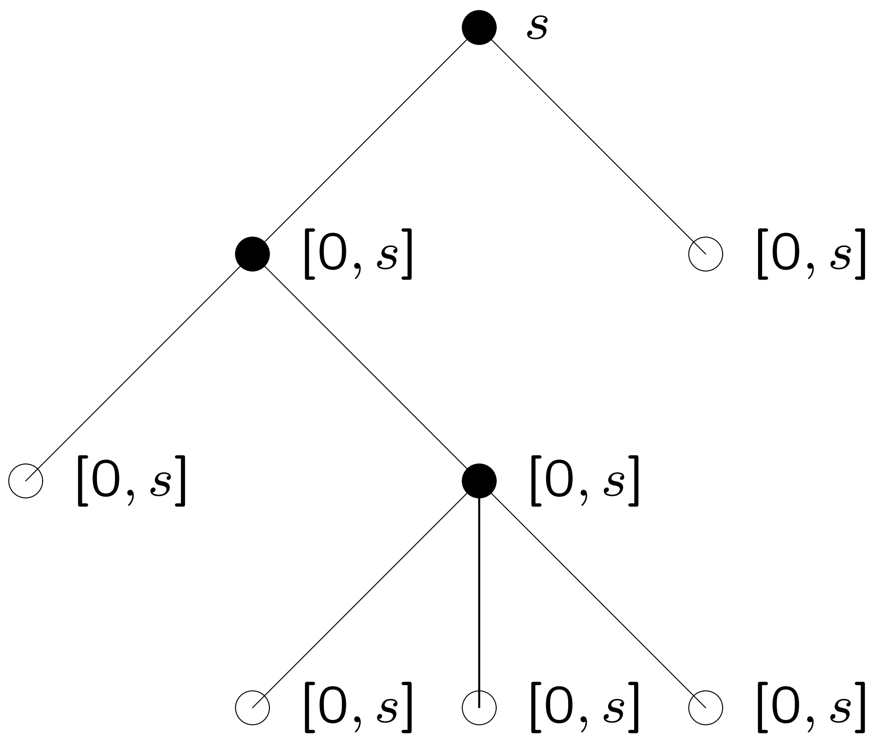
- Determines the degree of imprecision in *posterior* inferences; the larger s, the more cautious inferences are
- s as a number of additional *unknown* observations

□ Criteria for choosing s

- Encompass objective Bayesian inferences:
  Haldane: $s > 0$
  Perks: $s \geq 1$
  Other solutions? Problem: $s \geq K/2$ or $\geq K$
- Encompass frequentist inferences
- If too high, inferences are too weak

□ **Suggested values**: $s = 1$ or $s = 2$ (Walley, 1996)

Why does the IDM satisfy the RIP?



- Dirichlet distributions compatible with any tree. But, under a Dirichlet model, total prior strength s scatters when moving down the tree.

- In the IDM, all allocations of s to the nodes are possible (due to imprecision).

- Each sub-tree inheritates the same IDM(s) caracteristic.

5. EXAMPLES OF INFERENCES FROM THE IDM

5.1. PREDICTIVE INFERENCE & THE RULE OF SUCCESSION

Predictive inference, the IDMM

□ Predictive inference

Imprecise Dirichlet-multinomial model (IDMM) proposed by Walley & Bernard (1999).

Model for statistical inference about future observations $\boldsymbol{x}' = (x'_1, \ldots, x'_K)$ of size $n' = \sum_k x'_k$, sampled without replacement (multi-hypergeometric).

Prior uncertainty about $\boldsymbol{x}^* = \boldsymbol{x} + \boldsymbol{x}'$ is described by a set of *Dirichlet-multinomial* (*DiMn*) distributions.

$$P(\boldsymbol{x}^*) \propto \prod_k \binom{x^*_k + st_k - 1}{x^*_k} \tag{6}$$

□ Prior prediction about $\boldsymbol{x}^*$

$$\mathcal{M}_0 = \{DiMn(s\boldsymbol{t}, n^*) : \boldsymbol{t} \in \mathcal{S}^\star\} \tag{7}$$

□ Posterior prediction about $\boldsymbol{x}'|\boldsymbol{x}$

$$\mathcal{M}_n = \{DiMn(\boldsymbol{x} + s\boldsymbol{t}, n') : \boldsymbol{t} \in \mathcal{S}^\star\} \tag{8}$$

Links between IDM and IDMM

□ Relationship with inferences about θ

In general, in both Bayesian inference and in the IDM,

- $\boldsymbol{\theta}$ leads to $\boldsymbol{x}'$ (side-product of Bayes' theorem)
- $\boldsymbol{x}'$ gives $\boldsymbol{\theta}$ as $n' \to \infty$

The IDM and the IDMM are equivalent, if we assume that n' can tend to infinity.

□ Predictive model more fundamental

(see, Geisser, 1993)

- Finite population & data
- Models observables only, not hypothetical parameters
- Relies on exchangeability assumptions only.
- Gives the IDM as a limiting case as $n' \to \infty$

Rule of succession under the IDM

☐ **Prediction about the next observation**

Let B_j be the event that the next observation is of type c_j, where c_j is a subset of C with $1 \leq J \leq K$ elements and $x_j = \sum_{k \in j} x_k$.

☐ **Prior rule of succession**

The U&L prior probabilities of B_j are vacuous:

$$\underline{P}(B_j) = 0 \quad \text{and} \quad \overline{P}(B_j|\boldsymbol{x}) = 1,$$

obtained as $t_j \to 0$ and $t_j \to 1$ resp..

☐ **Posterior rule of succession**

After data $\boldsymbol{x}$ have been observed, the posterior U&L probabilities of event B_j are

$$\underline{P}(B_j|\boldsymbol{x}) = \frac{x_j}{n+s} \quad \text{and} \quad \overline{P}(B_j|\boldsymbol{x}) = \frac{x_j + s}{n+s},$$

obtained as $t_j \to 0$ and $t_j \to 1$ resp..

The interval contains $f_j = x_j/n$.

☐ **Rule independent** from C, K and J

Rule of succession and imprecision

□ Degree of imprecision about B_j

- Prior state: imprecision is maximal

$$\Delta(B_j) = \overline{P}(B_j) - \underline{P}(B_j) = 1$$

- Posterior state:

$$\Delta(B_j|\boldsymbol{x}) = \overline{P}(B_j|\boldsymbol{x}) - \underline{P}(B_j|\boldsymbol{x}) = \frac{s}{n+s}$$

□ Prior ignorance

Caracterized by a maximal imprecision, *i.e.* vacuous probabilities.

□ Interpretation of s

Hyper-parameter s controls how fast imprecision diminishes with n: s is the number of observations necessary to halve imprecision about B_j.

Bayesian rule of succession

□ Bayesian rule of succession

The rule of succession obtained from a single symmetric Dirichlet distribution, *Diri*$(\boldsymbol{\alpha})$ with $\alpha_k = s/K$, is

$$P(B_j) = \frac{x_j + \alpha_j}{n + s} = \frac{nf_j + sJ/K}{n + s} \qquad (9)$$

□ Objective Bayesian rules

Bayes	$P(B_j) = (x_j + J)/(n + K)$
Jeffreys	$P(B_j) = (x_j + J/2)/(n + K/2)$
Perks	$P(B_j) = (x_j + J/K)/(n + 1)$
Haldane	$P(B_j) = x_j/n$

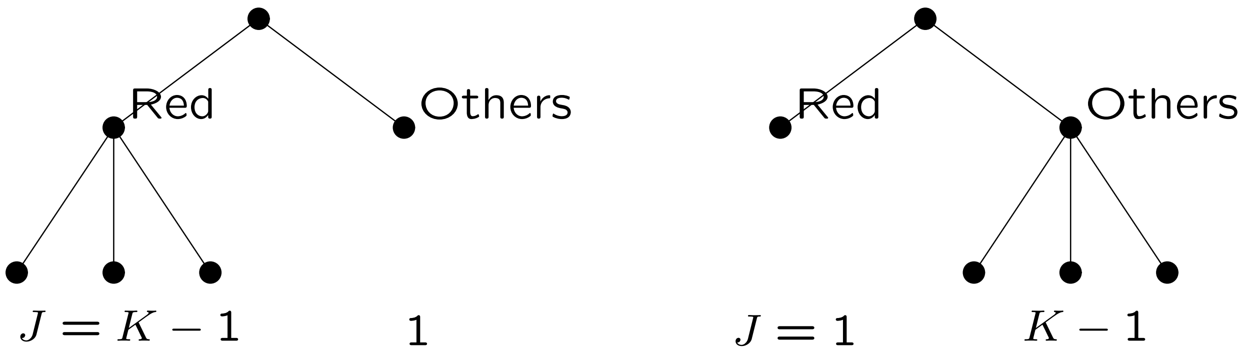
□ Dependence on K and J except Haldane

□ Particular case $J = 1$, $K = 2$

If $x_1 = n/2$, *i.e.* $f = 1/2$, each Bayesian rule leads to $P(B) = 1/2$, whether $n = 0$, or $n = 10$, 100 or 1000.

Categorization arbitrariness

□ **Arbitrariness of C, *i.e.* J and K**



Most extremes cases obtained as $K \to \infty$

□ **Bayesian rules** lead to intervals when arbitrariness is introduced

Bayes-Laplace	$[0; 1]$,	IDM$(s = \infty)$
Jeffreys	$[0; 1]$,	IDM$(s = \infty)$
Perks	$[\frac{x_k}{n+1}; \frac{x_k+1}{n+1}]$,	IDM$(s = 1)$
Haldane	$[x_k/n; x_k/n]$,	IDM$(s \to 0)$

Frequentist prediction

□ **"Bayesian and confidence limits for prediction"** (Thatcher, 1964)

- Considers binomial or hypergeometric data $(K = 2)$, $\boldsymbol{x} = (x_1, n - x_1)$.
- Studies the prediction about n' future observations $\boldsymbol{x'} = (x'_1, n' - x'_1)$.
- Derives lower and upper *confidence* (frequentist) limits for x'_1.
- Compares these confidence limits to *credibility* (Bayesian) limits from a Beta prior.

□ **Main result**

- Upper confidence and credibility limits for x'_1 coincide *iff* the prior is *Beta*$(\alpha_1 = 1, \alpha_2 = 0)$.
- Lower confidence and credibility limits for x'_1 coincide *iff* the prior is *Beta*$(\alpha_1 = 0, \alpha_2 = 1)$.

Frequentist rule of succession

□ Frequentist "rule of succession"

For $n' = 1$, the lower and upper confidence limits resp. correspond to the following Bayesian rules:

$$P(B_j|\boldsymbol{x}) = \frac{x_j}{n+1} \quad \text{and} \quad P(B_j|\boldsymbol{x}) = \frac{x_j+1}{n+1}$$

i.e. to the IDM interval for $s = 1$.

□ A "difficulty"

"... is there a prior distribution such that both the upper and lower Bayesian limits always coincide with confidence limits? ... In fact there are not such distributions." (Thatcher, 1964, p. 184)

□ Reconciling frequentist and Bayesian

"... we shall consider whether these difficulties can be overcome by a more general approach to the prediction problem: in fact, by ceasing to restrict ourselves to a single set of confidence limits or a single prior distribution." (Thatcher, 1964, p. 187)

5.2. IMPRECISE BETA MODEL (IBM)

Bernoulli process, frequentist *vs.* Bayesian (Bernard, 1996)

☐ **Data from a Bernoulli process**

Sequential binary data (success/failure), *e.g.* sequence

S, F, S, S, S, S, S, F, S, S,

so that $a = x_S = 8$, $b = x_F = 2$, $n = 10$.

☐ **Problem of testing a one-sided hypothesis**

$$H_0 : \theta_S \leq \theta_0 \quad \textit{vs.} \quad H_1 : \theta_S > \theta_0$$

☐ **Example:** $f_S = 8/10$, $\theta_S > \theta_0 = 1/2$?

☐ **Comparison** of frequentist solutions and objective Bayesian solutions to this problem.

Frequentist approach

☐ Principle

Consider all *possible* data sets, that are *more extreme* than the observed data under H_0, *i.e.* such that F_S greater than $f_S = \frac{8}{10}$, and add up their probabilities under H_0 (yielding "the" p-value).

☐ "Possible": depends on stopping rule; either stop after

- n observations: n-rule
- a successes: a-rule (neg. sampling)
- b failures: b-rule (neg. sampling)

☐ "More extreme": three conventions for computing the p-value

- Inclusive: $p_{inc} = P(F_S \geq f_S | H_0)$
- Exclusive: $p_{exc} = P(F_S > f_S | H_0)$
- Mid-P convention: $p_{mid} = (p_{exc} + p_{inc})/2$

Objective Bayesian approach

□ Principle

Consider an *objective Beta*(α, β) prior on θ_S, derive an (updated) posterior on $\theta_S|\boldsymbol{x}$, then compute

$$PB_{\alpha,\beta} = P_{\alpha,\beta}(H_0|\boldsymbol{x}).$$

□ Objective Beta priors

$\alpha = 0,\ \beta = 0$: Haldane
$\alpha = \frac{1}{2},\ \beta = \frac{1}{2}$: Jeffreys-(n), Perks
$\alpha = 1,\ \beta = 1$: Bayes-Laplace
$\alpha = 0,\ \beta = \frac{1}{2}$: Jeffreys-(a)
$\alpha = \frac{1}{2},\ \beta = 0$: Jeffreys-(b)
$\alpha = 0,\ \beta = 1$: Hartigan-(b) ALI prior
$\alpha = 1,\ \beta = 0$: Hartigan-(a) ALI prior

Main results

□ **Comparison frequentist** *vs.* **Bayesian**
(Bernard, 1996)

$$\begin{aligned} PB_{1,0} &= P_{n,I} = P_{a,I} = 11/1024 \\ PB_{0,1} &= P_{n,E} = P_{b,E} = 56/1024 \end{aligned}$$

$$PB_{1,0} \;\leq \text{all } P\text{'s and } PB\text{'s } \;\leq PB_{0,1}$$

□ **Ignorance zone**

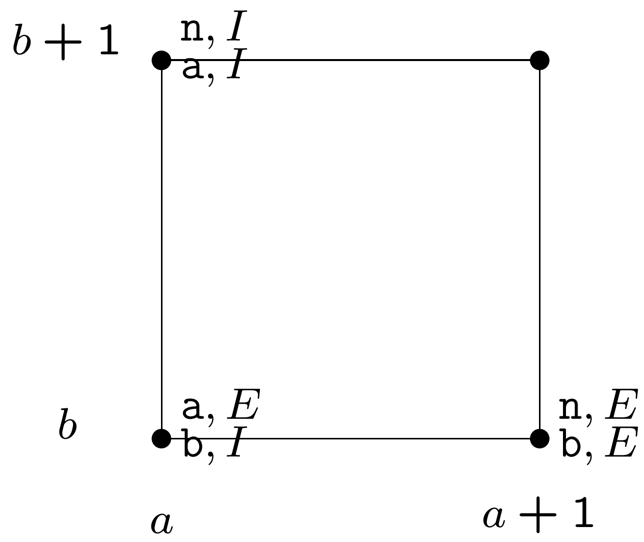
The bounds of this ignorance zone correspond to the *Imprecise Beta Model* (IBM) with $s = 1$.

□ **Reconcile frequentist principles & LP**
(Walley, 2002)

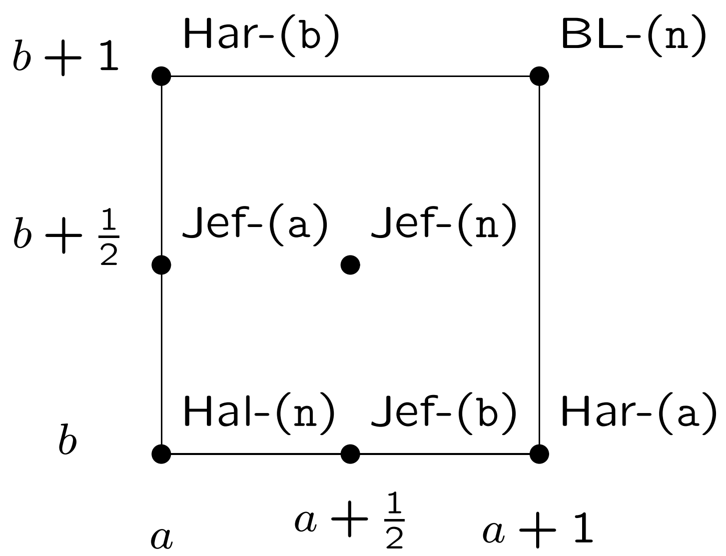
The IBM with $s = 1$ produces statements about one-sided or equi-tailed two-sided hypotheses relative to θ_S, which satisfies weak frequentist principles (validity under any monotone stopping-rule), LP and coherence.

Frequentist and Bayesian levels maps

□ Frequentist significance levels



□ Bayesian significance levels



5.3. TWO BY TWO CONTINGENCY TABLES

Independence in a 2×2 contingency table

□ Data

	$b1$	$b2$
$a1$	x_{11}	x_{12}
$a2$	x_{21}	x_{22}

	$b1$	$b2$
$a1$	8	4
$a2$	2	5

□ Problem

Positive association between A and B?

Derived parameter: contingency coefficient

$$\rho = \frac{\theta_{11}}{\theta_{1.}\theta_{.1}} \qquad r_{obs} = 0.467$$

Hypothesis to be tested:

$$H_0 : \rho \leq 0 \quad \text{vs.} \quad H_1 : \rho > 0$$

□ **Comparison** of frequentist, Bayesian & IDM inferences (Altham, 1969; Walley, 1996; Walley et al., 1996; Bernard, 2003)

Frequentist inference

□ Fisher's exact test for a 2×2 table

Amounts to considering all 2×2 tables $\boldsymbol{x}$ with the same margins than those observed.

Frequentist probability of any $\boldsymbol{x}$ under H_0 is

$$P(\boldsymbol{x}|H_0) = \frac{x_{1.}!x_{2.}!x_{.1}!x_{.2}!}{n!x_{11}!x_{12}!x_{21}!x_{22}!}$$

The p-value of the test is defined as,

$$p_{obs} = P(\textit{more extreme data}|H_0)$$

where "more extreme data" means all $\boldsymbol{x}$ with R larger than r_{obs}.

□ Frequentist solutions

- $p_{obs} = p_{inc}$, more or as extreme
- $p_{obs} = p_{exc}$, strictly more extreme

Inclusive convention is the usual one; but roles of "inclusive" and "exclusive" are permuted when considering the test of $H_O : \phi \geq 0$ vs. $H_1 : \phi < 0$.

Bayesian & Imprecise models

□ **Objective Bayesian models**, for fixed n:

Haldane, Perks, Jeffreys, Bayes-Laplace

□ **IBM**

Suggested by Walley (1996) and Walley et al. (1996) for the ECMO data: A are groups of patients and B outcomes of treatment.

Suggest using two independent IBM's with $s = 1$ each for each group.

□ **IDM**, with $s = 1$ or $s = 2$

□ **Relationships** between models

$$\underline{P}[\text{IDM}_2] \quad \leq \quad \underline{P}[\text{IBM}] = p_{exc} \quad \leq \quad \underline{P}[\text{IDM}_1]$$

$$\leq \quad PB[\text{Hal}],\ PB[\text{Per}],\ PB[\text{Jef}],\ PB[\text{BL}] \quad \leq$$

$$\overline{P}[\text{IDM}_1] \quad \leq \quad \overline{P}[\text{IBM}] = p_{inc} \quad \leq \quad \overline{P}[\text{IDM}_2]$$

Comparison with objective models

Haldane +	$\begin{array}{|c|c|}\hline 8 & 2 \\ \hline 4 & 5 \\ \hline\end{array}$	Freq.	Bayesian	Imprecise
$\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 2 \\ \hline\end{array}$	.015			$\underline{P}$ IDM($s = 2$)
$\begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 1 \\ \hline\end{array}$	.017	p_{exc}		$\underline{P}$ IBM($2 \times s = 1$)
$\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 1 \\ \hline\end{array}$	.025			$\underline{P}$ IDM($s = 1$)
$\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 0 \\ \hline\end{array}$	.043		Haldane	
$\begin{array}{|c|c|}\hline \frac{1}{4} & \frac{1}{4} \\ \hline \frac{1}{4} & \frac{1}{4} \\ \hline\end{array}$	.047		Perks	
$\begin{array}{|c|c|}\hline \frac{1}{2} & \frac{1}{2} \\ \hline \frac{1}{2} & \frac{1}{2} \\ \hline\end{array}$	.053		Jeffreys	
$\begin{array}{|c|c|}\hline 1 & 1 \\ \hline 1 & 1 \\ \hline\end{array}$	.063		Bay.-Lap.	
$\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 1 & 0 \\ \hline\end{array}$	.088			$\overline{P}$ IDM($s = 1$)
$\begin{array}{|c|c|}\hline 0 & 1 \\ \hline 1 & 0 \\ \hline\end{array}$	.130	p_{inc}		$\overline{P}$ IBM($2 \times s = 1$)
$\begin{array}{|c|c|}\hline 0 & 0 \\ \hline 2 & 0 \\ \hline\end{array}$	.144			$\overline{P}$ IDM($s = 2$)

5.4. LARGE n AND POSTERIOR IMPRECISION

Large n, Bayesian models and IDM

□ Claim by Bayesians or IP papers

When n is large, all objective Bayesian priors lead to similar inferences.

This claim is also (implicitly) present in many IP writings.

□ This claim is FALSE!

□ Counter-examples

- Inference about a chance θ in binary data
- Inference about association in 2×2 table
- Inference about a universal law (Walley, Bernard, 1999)
- Inference about quasi-implications in multi-variate binary data (Bernard, 2001)

Inference about a single chance θ

☐ Problem

- Observed counts $\boldsymbol{x} = (x_1, x_2)$, $n = x_1 + x_2$
- Test $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$

☐ U&L probs. of H_0 under the IDM($s = 1$)

$$\begin{aligned}\underline{P}(\theta \leq \theta_0|\boldsymbol{x}) &= P(X_1 > x_1|H_0, n) \\ \overline{P}(\theta \leq \theta_0|\boldsymbol{x}) &= P(X_1 \geq x_1|H_0, n)\end{aligned}$$

$$\begin{aligned}\Delta(\theta \leq \theta_0|\boldsymbol{x}) &= P_n(X_1 = x_1|H_0, n) \\ &= \binom{n}{x_1}\theta_0^{x_1}(1-\theta_0)^{x_2}\end{aligned}$$

☐ Example: $x_1 = 0$, $x_2 = 100$, $\theta_0 = 0.001$

$$\begin{aligned}\underline{P}(\theta \leq \theta_0|\boldsymbol{x}) &= 0 \\ \overline{P}(\theta \leq \theta_0|\boldsymbol{x}) &= 0.905 \\ \Delta(\theta \leq \theta_0|\boldsymbol{x}) &= 0.905\end{aligned}$$

☐ Why? $P(\textit{observed data}|H_0)$ is high

Association in 2×2 tables

☐ **Example** $n = 115$

	$b1$	$b2$
$a1$	0	4
$a2$	4	107

☐ **Fisher's test:** $H_0 : \Phi \geq 0$ *vs.* $H_1 : \Phi < 0$

Exclusive: $p_{exc} = 0$
Inclusive: $p_{inc} = 0.866$

☐ **Bayesian answers** (taking $K = 4$)

Haldane: $P(H_1) = 0$
Perks: $P(H_1) = 0.350$
Jeffreys: $P(H_1) = 0.571$
Bayes: $P(H_1) = 0.802$

☐ **IDM answers**

$s = 1$: $\underline{P}(H_1) = 0$, $\overline{P}(H_1) = 0.866$
$s = 2$: $\underline{P}(H_1) = 0$, $\overline{P}(H_1) = 0.986$

☐ **Why?** Indepence is compatible with data (despite $x_{11} = 0$), because f_a and f_b are small.

Comments

□ **What happens?** There are situations in which

- n is large
- objective Bayesian inferences do not agree
- inferences from the IDM are highly imprecise

□ **Tentative explanation**

From the frequentist viewpoint, in the two examples, the two hypotheses H_0 and H_1 are both extremely compatible with the data.

This occurs because, in both cases, the frequentist probability $P(x|H_0)$ is high.

□ **Consequences for the IDM**

Within a unique dataset, imprecision in the inferences from the IDM can vary considerably (Bernard, 2001, 2003)

5.5. NON-PARAMETRIC ESTIMATION OF A MEAN

Non-parametric estimation of a mean

□ **Problem**

Numerical data, bounded with finite precision. Possible values amongst the set $\{y_1, y_2, \ldots, y_K\}$ such that $y_1 < y_2 < \cdots < y_K$.

A sample yields the counts $\boldsymbol{x} = (x_1, \ldots, x_K)$.

More realistic than assumption of normality, *etc.*.

□ **Parameter of interest**, the unknown mean

$$\mu = \sum_k y_k \theta_k$$

□ **Bayesian inference**, from a *Diri*$(\boldsymbol{\alpha})$ prior,

$$\begin{aligned} \mu &\sim \textit{L-Diri}(\boldsymbol{y}, \boldsymbol{\alpha}) \\ \mu|\boldsymbol{x} &\sim \textit{L-Diri}(\boldsymbol{y}, \boldsymbol{x} + \boldsymbol{\alpha}) \end{aligned}$$

Inferences from the IDM

□ Prior expectations

$$\underline{E}(\mu) = y_1 \quad \text{and} \quad \overline{E}(\mu) = y_K$$

□ Posterior expectations

$$\begin{aligned}
\underline{E}(\mu|\boldsymbol{x}) &= \frac{n\mathit{Mean}(\boldsymbol{y},\boldsymbol{x}) + sy_1}{n+s} \\
\overline{E}(\mu|\boldsymbol{x}) &= \frac{n\mathit{Mean}(\boldsymbol{y},\boldsymbol{x}) + sy_K}{n+s}
\end{aligned}$$

obtained as $t_1 \to 1$ or $t_K \to 1$ resp..

□ U&L cdf's

The same limits lead to the U&L prior and posterior cdf's of μ.

All inferences from the IDM can be carried out using the two extreme distributions

$$\mathit{L\text{-}Diri}(\boldsymbol{y}, \boldsymbol{x} + \boldsymbol{\alpha} = (x_1 + n, x_2, \ldots, x_K))$$

$$\mathit{L\text{-}Diri}(\boldsymbol{y}, \boldsymbol{x} + \boldsymbol{\alpha} = (x_1, \ldots, x_{K-1}, \ldots, x_K + n))$$

Implications for the choice of s

☐ **Theorem** (Bernard, 2001)

$$\textit{L-Diri}(\boldsymbol{y}, \boldsymbol{\alpha}) \to \textit{Uni}(y_1, y_K)$$

for $\alpha_1 = \alpha_K = 1$ and $\alpha_k \to 0$, $k \neq 1, K$

☐ **Objective Bayesian inference & IDM**

Three reasonable priors encompassed by the IDM

Haldane	if	$s > 0$
Perks	if	$s \geq 1$
Uniform	if	$s \geq 2$ (from theorem above)

Jeffreys' and Bayes-Laplace's priors on set Y lead to highly informative priors about μ.

☐ **Conclusion:** Case with large K, where $s = 2$ encompasses all reasonable Bayesian alternatives.

5.6 SOME APPLICATIONS OF THE IDM

Some applications of the IDM

- Reliability analysis: Analysis of failure data including right-censored observations (Coolen, 1997; Yan, 2002).
- Predictive inferences from multinomial data (Walley, Bernard, 1999; Coolen, Augustin, in prep.).
- Non-parametric inference about a mean (Bernard, 2001).
- Classification, networks, tree-dependencies structures, estimation of entropy or mutual information (Cozman, Chrisman, 1997; Zaffalon, 2001a, 2001b; Hutter, 2003).
- Treatment of missing data (Zaffalon, 2002).
- Implicative analysis for multivariate binary data (large $K = 2^q$) (Bernard, 2002).
- Analysis of local associations in contingency tables (Bernard, 2003).
- Game-theoretic learning (Quaeghebeur, de Cooman, 2003)

6. CHOICE OF s

Interpretations of s

☐ Caution parameter

- Prior uncertainty: In many cases, any $s > 0$ produces vacuous prior probabilities.
- Posterior uncertainty: s determines the degree of imprecision in *posterior* inferences; the larger s, the more cautious inferences are.

☐ IDM's nested according to s

The probability intervals produced by two IDM's such that $s_1 < s_2$ are nested:

$$Int[s_2] \subset Int[s_2]$$

☐ Number of additional observations

In several examples, using the IDM amounts to making Bayesian inferences

- from Haldane's prior
- taking the observed data $\boldsymbol{x}$ into account
- adding s observations to the more extreme categories

Note: cf. some ad-hoc frequentist methods

Choice of hyper-parameter s

☐ **Two contradictory aims**

- Large enough to encompass alternative objective models
- Not too large, because inferences are too weak

☐ **Encompassing alternative models**

- Haldane: $s > 0$
- Perks: $s \geq 1$
- Jeffreys or Bayes-Laplace: would require $s \geq K/2$ or $\geq K$, but produce unreasonable inferences when K large (cf. categ. arbitrariness, infer. on a mean).
- Berger-Bernardo: open question.
- Encompass frequentist inferences: some arguments for $s = 1$ for $K = 2$ or $K = 4$.

☐ **Additional new principle?** (Walley, 1996)

Which value for s

☐ Suggested value(s) for s?

- First results suggested $1 \leq s \leq 2$, but mostly based on cases with $K = 2$ or small K (Walley, 1996).
- Some new arguments, in the case of large K, for $s = 2$ (Bernard, 2001, 2003).

☐ Problem not settled yet

- Need to study more situations with K large.
- Need to compare the IDM with alternative objective models in such cases.

7. COMPUTATIONAL ASPECTS

Computational aspects

□ General problem

Min-/maximization of $E_{\boldsymbol{st}}(\lambda)$ and $P_{\boldsymbol{st}}(\lambda \leq u)$ for general $\lambda = g(\boldsymbol{\theta})$.

- Simple (and identical) solution to both problems when $g(.)$ is linear: $t_k \to 1$ for extreme k's (*w.r.t.* to $g(.)$) (Walley, Bernard, 1999; Bernard, 2001).

- Some exact & approximate solutions for specific cases (Bernard, 2003; Hutter, 2003).

□ Remaining issues

- Find class of functions $g(.)$ for which $t_k \to 1$ for some k provides the solution.

- Is saying $t_k \to 1$ enough to specify the min-/maximization solution? NO: in some case, necessity to say how the other t_k's tend to 0.

- Find exact or conservative approximate solutions for general $g(.)$.

- Find non-conservative approximate solutions (useful in practical applications).

- Can the predictive approach help?

8. CONCLUSIONS

Why using a set of Dirichlet's
Walley (1996, p. 7)

(a) Dirichlet prior distributions are mathematically tractable because . . . they generate Dirichlet posterior distributions;

(b) when categories are combined, Dirichlet distributions transform to other Dirichlet distributions (this is the crucial property which ensures that the RIP is satisfied);

(c) sets of Dirichlet distributions are very rich, because they produce the same inferences as their convex hull and any prior distribution can be approximated by a finite mixture of Dirichlet distributions;

(d) the most common Bayesian models for prior ignorance about $\boldsymbol{\theta}$ are Dirichlet distributions.

Fundamental properties of the IDM

□ Principles

Satisfies several desirable principles for prior ignorance: SP, EP, RIP, LP, SRP, coherence.

□ IDM *vs.* Bayesian and frequentist

- Answers several difficulties of alternative approaches
- Provides means to reconcile frequentist and objective Bayesian approaches (Walley, 2002)

□ Generality

More general than for multinomial data. Valid under a general hypothesis of exchangeability between observed and future data. (Walley, Bernard, 1999).

□ Degree of imprecision and n

Degree of imprecision in posterior inferences enables one to distinguish between: *(a)* prior uncertainty still dominates, *(b)* there is substantial information in the data.
The two cases can occur within the same data set.

Future research, open questions

- Find a new principle suggesting an upper bound for s.
- Major argument for Jeffreys' prior is that it is reparameterization invariant. Does this concept have a meaning within the IDM?
- Compare the IDM with Berger-Bernardo reference priors.
- Study the properties of the IDM in situations with possibly large K, compare it with alternative models.
- Further applications of the IDM for non-parametric inference from numerical data.
- Applications to classification, networks, tree-dependencies structures.
- Elaborate theory & algorithms for computing inferences from the IDM in general cases.