The Imprecise Dirichlet Model

Jean-Marc Bernard
CNRS UMR 8069 & University Paris 5

Second Summer School of the
Society for Imprecise Probabilities,
Theory and Applications

Madrid, Spain
25 July 2006

INTRODUCTION

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?"

Two problems of predictive inference

- Prior prediction, before observing any item
- Posterior prediction, after observing n items

Inference from a state of prior ignorance

about the proportions of the various colours

Categorical data (1)

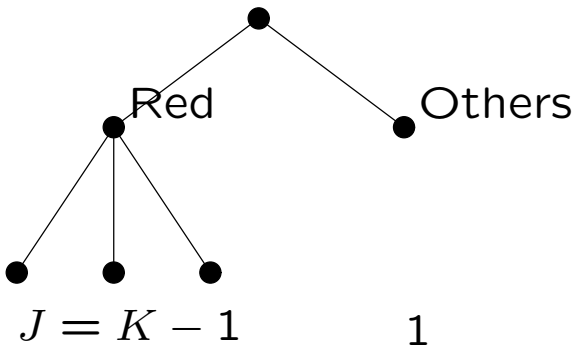
Categories

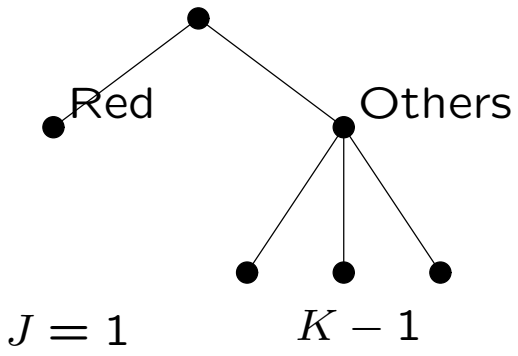
- Set of K of categories or types

$$C = \{c_1, \ldots, c_K\}$$

- Categories c_k are exclusive and exhaustive
- Possible to add an extra category: "other colours", "other types"

Categorisation is partly arbitrary





Categorical data (2)

□ Data

- Set, or sequence, I of n observations, items, individuals, *etc.*
- For each individual $i \in I$, we observe the corresponding category

$$\begin{array}{ccc} I & \to & C \\ i & \mapsto & c_k \end{array}$$

- Observed composition, in counts:

$$\boldsymbol{a} = (a_1, \ldots, a_K)$$

with $\sum_k a_k = n$

- Observed composition, in frequencies:

$$\boldsymbol{f} = (f_1, \ldots, f_K) = \frac{\boldsymbol{a}}{n}$$

with $\sum_k f_k = 1$

□ **Compositions:** order considered as not important

Statistical inference problems

☐ Inference about what?

- Future counts or frequencies in n' future observations

$$\begin{aligned} \boldsymbol{a'} &= (a'_1, \ldots, a'_K) \\ \boldsymbol{f'} &= (f'_1, \ldots, f'_K) \end{aligned}$$

$n' = 1$ Immediate prediction
$n' \geq 2$ Predictive inference

- True/parent frequencies (parameters) in whole population of size N

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$$

$N < \infty$ Parametric inference
$N = \infty$ Parametric inference

☐ Prior *vs.* posterior inferences

- Prior inference: $n = 0$
- Posterior inference: $n \geq 1$

Model relating past and future observations

Random sampling

- Population with a fixed, but unknown, true composition in frequencies

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$$

- Data (observed & future): random samples from the same population
- Ensures that the data are representative of the population *w.r.t.* C

Stopping rule

- Fixed n
- Fixed a_k, "negative" sampling
- More complex stopping rules

Finite/infinite population

- Multi-hypergeometric (N finite)
- Multinomial ($N = \infty$)

These elements define a sampling model

Alternative model: Exchangeabiblity

□ Exchangeability

- Consider any sequence S of $n^* = n + n'$ observations,

$$S = (c_1, \ldots, c_n, c_{n+1}, \ldots, c_{n^*})$$

having composition

$$\boldsymbol{a}^* = (a_1^*, \ldots, a_K^*)$$

- Assumption of order-invariance, or permutation-invariance

$$\forall\, S, \quad P(S \mid \boldsymbol{a}^*) = constant$$

□ Relates past and future observations directly

□ No need to invoke unknown parameters θ

□ see Gert's lecture

A statistical challenge

□ **Challenge:**

- Model prior ignorance about $\boldsymbol{\theta}$
- Arbitrariness of C and K, both may vary as data items are observed
- Model prior ignorance about both C and K

□ **Make reasonable inferences**
from such a state of prior ignorance

- Idea of "objective" methods: "let the data speak for themselves"
- Frequentist methods
- objective Bayesian methods

□ **"Reasonable":** Several desirable principles

Desirable principles / properties (1)

□ Prior ignorance

- Symmetry (SP): Prior uncertainty should be invariant *w.r.t.* permutations of categories
- Embedding pcple (EP): Prior uncertainty should not depend on refinements or coarsenings of categories

□ Independence from irrelevant information

of posterior inferences

- Stopping rule pcple (SRP): Inferences should not depend on the stopping rule, *i.e.* on data that might have occurred but have actually not
- Likelihood pcple (LP): Inferences should depend on the data through the likelihood function only
- Representation invariance (RIP): Inferences should not depend on refinements or coarsenings of categories
- Specificity pcple:?

Desirable principles / properties (2)

□ **Reasonable account of uncertainty** in prior and posterior inferences

□ **Consistency requirements** when considering several inferences

- Avoiding sure loss (ASL): Probabilistic assessments, when interpreted as betting dispositions, should not jointly lead to a sure loss
- Coherence (CP): Stronger property of consistency of all probabilistic assessments

□ **Frequentist interpretation(s)**

- Repeated sampling pcple (RSP): Probabilities should have an interpretation as relative frequencies in the long run

□ **See** Walley, 1996; 2002

Methods for statistical inference: Frequentist approach

Frequentists methods

- Based upon sampling model only
- Probabilities can be assimilated to long-run frequencies
- Significance tests, confidence limits and intervals (Fisher, Neyman & Pearson)

Difficulties of frequentist methods

- Depend on the stopping rule. Hence do not obey SRP, nor LP
- Not conditional on observed data; May have relevant subsets
- For multidimensional parameters' space: ad-hoc and/or asymptotic solutions to the problem of nuisance parameters

Methods for statistical inference: Objective Bayesian approach (1)

Bayesian methods

- Two ingredients: sampling model + prior
- Prior distribution for multinomial data: conjugate Dirichlet family
- Depend on the sampling model through the likelihood function only, when prior chosen independently

Objective Bayesian methods

- Data analysis goal: let the data say what they have to say about unknown parameters
- Priors formalizing "prior ignorance"
- objective Bayesian: "non-informative" priors, *etc.* (*e.g.* Kass, Wasserman, 1996)
- Exact or approximate frequentist reinterpretations: "matching priors" (*e.g.* Datta, Ghosh, 1995)

Methods for statistical inference: objective Bayesian approach (2)

□ **Difficulties of Bayesian methods** for categorical data

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)
- Some solutions can generate incoherent inferences (Berger, Bernardo, 1992)
- If $K = 2$, uncertainty about next observation (case $n' = 1$) is the same whether $a_1 = a_2 = 0$ (prior) or $a_1 = a_2 = 100$ (posterior)

$$P(\boldsymbol{a}' = (1,0)) \;\; = \;\; P(\boldsymbol{a}' = (1,0) \mid \boldsymbol{a})$$

□ **Only approximate agreement** between frequentist methods and objective Bayesian methods, for categorical data

The IDM in brief

□ **Model for statistical inference** for categorical data
Proposed by Walley (1996), generalizes the IBM (Walley, 1991).
Inference from data $\boldsymbol{a} = (a_1, \ldots, a_K)$, categorized in K categories C, with unknown chances $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$.

□ **Imprecise probability model**
Prior uncertainty about $\boldsymbol{\theta}$ expressed by a set of Dirichlet's.
Posterior uncertainty about $\boldsymbol{\theta}|\boldsymbol{a}$ then described by a set of updated Dirichlet's.
Generalizes Bayesian inference, where prior/ posterior uncertainty is described by a *single* Dirichlet.

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

□ **Models prior ignorance** about $\boldsymbol{\theta}$, K and C

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

Outline

1. Introduction
2. Bayesian inference
3. From Bayesian to imprecise models
4. Definition of the IDM
5. Important distributions
6. Properties of the IDM
7. Inferences from the IDM
   - Predictive inference
   - The rule of succession
   - Imprecise Beta model
   - Contingency tables
   - Large n and the IDM
   - Non-parametric inference on a mean
8. Choice of hyper-parameter s
9. Some applications
10. Computational aspects
11. Conclusions

THE BAYESIAN APPROACH

Presentation

Hypotheses

- We assume $N = \infty$
- Set C, and number of categories, K, are considered as known and fixed
- Data have a multinomial likelihood

Focus on the Bayesian approach since

- Bayesian: a single Dirichlet prior yields a single Dirichlet posterior, precise model
- IDM: a prior set of Dirichlet's yields a posterior set of Dirichlet's, imprecise model

Goal

- Sketch Bayesian approach to inference
- Specifically: objective Bayesian models
- Indicate shortcomings of these models

Inference from multinomial data

Multinomial data

- Elements of population are categorized in K categories from set $C = \{c_1, \ldots, c_K\}$.
- Unknown true chances $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, with $\theta_k \geq 0$ and $\sum_k \theta_k = 1$, *i.e.* $\boldsymbol{\theta} \in \Theta = \mathcal{S}(1, K)$.
- Data are a random sample of size n from the population, yielding counts $\boldsymbol{a} = (a_1, \ldots, a_K)$, with $\sum_k a_k = n$.

Multinomial sampling distribution

$$P(\boldsymbol{a}|\boldsymbol{\theta}) \;=\; \binom{n}{\boldsymbol{a}} \theta_1^{a_1} \ldots \theta_K^{a_K}$$

When seen as a function of $\boldsymbol{\theta}$, leads to the likelihood function

$$L(\boldsymbol{\theta}|\boldsymbol{a}) \;\propto\; \theta_1^{a_1} \ldots \theta_K^{a_K}$$

Same likelihood is obtained from observing $\boldsymbol{a}$, for a variety of stopping rules: n fixed, a_k fixed, *etc.*

Bayesian inference: a learning model

General scheme

$$\left\{\begin{array}{c}\text{Prior } P(\boldsymbol{\theta}) \\ + \\ \text{Sampling } P(\boldsymbol{a}|\boldsymbol{\theta})\end{array}\right. \longrightarrow \left\{\begin{array}{c}\text{Posterior } P(\boldsymbol{\theta}|\boldsymbol{a}) \\ + \\ \text{Prior predictive } P(\boldsymbol{a})\end{array}\right.$$

Iterative process

$$\left\{\begin{array}{c}\text{Prior' } P(\boldsymbol{\theta}|\boldsymbol{a}) \\ + \\ \text{Sampl.' } P(\boldsymbol{a'}|\boldsymbol{\theta},\boldsymbol{a})\end{array}\right. \longrightarrow \left\{\begin{array}{c}\text{Posterior' } P(\boldsymbol{\theta}|\boldsymbol{a'},\boldsymbol{a}) \\ + \\ \text{Post. pred. } P(\boldsymbol{a'}|\boldsymbol{a})\end{array}\right.$$

Learning model about

- unknown chances: $P(\boldsymbol{\theta})$ updated to $P(\boldsymbol{\theta}|\boldsymbol{a})$
- future data: $P(\boldsymbol{a})$ updated to $P(\boldsymbol{a'}|\boldsymbol{a})$

Warning: probabilities or probabilities?

□ **Several quantities behave like "probabilities"**

$$f \quad f'$$

$$\theta \quad P(a|\theta)$$

and

$$P(\theta) \;\; P(\theta|a), \quad P(a) \;\; P(a'|a)$$

□ **Major distinction**

- Frequentist probabilities, and frequencies: can be given a (relative) frequency interpretation
- Epistemic probabilities: describe degrees of belief, express (personal) states of uncertainty

□ **Problem:** Our subject deals about "Probabilities about probabilities"!

Frequencies & frequentist probabilities

☐ Proportions, (relative) frequencies

$$f \quad f'$$

- f: known observed frequencies
- f': unknown future frequencies

☐ Interpretable as (relative) frequencies

$$\theta \quad P(a|\theta)$$

- θ: unknown chances
- $P(a|\theta)$: sampling probabilities, are functions of the chances θ
- "Frequentist probabilities": can be viewed as frequencies in the long run
- Also called aleatory probabilities

Epistemic probabilities

Four kinds of epistemic probabilities

$$P(\theta) \quad P(\theta|a) \quad P(a) \quad P(a'|a)$$

About what?

- unknown chances, θ
- future data, a or a'

Conditional on what?

- unconditional, describe a prior state of uncertainty
- conditional on observed data a, describe a posterior state of knowledge / uncertainty

Behavioural interpretation

Epistemic probabilities describe dispositions to accept bets (de Finetti, 1974-75; Walley, 1991)

What about chances?

Principle of direct inference (Walley, 1991)

Bayesian inference

Continuous parameters space

Since the parameters space, Θ, is continuous, probabilities on $\boldsymbol{\theta}$, $P(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\boldsymbol{a})$, are defined via densities, denoted $h(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta}|\boldsymbol{a})$

Bayes' theorem (or rule)

$$h(\boldsymbol{\theta}|\boldsymbol{a}) = \frac{h(\boldsymbol{\theta})\ P(\boldsymbol{a}|\boldsymbol{\theta})}{\int_{\Theta} h(\boldsymbol{\theta})\ P(\boldsymbol{a}|\boldsymbol{\theta})\ d\boldsymbol{\theta}}$$

$$= \frac{h(\boldsymbol{\theta})\ L(\boldsymbol{\theta}|\boldsymbol{a})}{\int_{\Theta} h(\boldsymbol{\theta})\ L(\boldsymbol{\theta}|\boldsymbol{a})\ d\boldsymbol{\theta}}$$

Likelihood principle satisfied if prior $h(\boldsymbol{\theta})$ is chosen independently of $P(\boldsymbol{a}|\boldsymbol{\theta})$

Conjugate inference

- Prior $h(\boldsymbol{\theta})$ and posterior $h(\boldsymbol{\theta}|\boldsymbol{a})$ are from the same family
- For multinomial likelihood: Dirichlet family

Dirichlet prior for θ

Dirichlet prior

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

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

with prior strengths

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

such that $\alpha_k > 0, \quad \sum_k \alpha_k = s$

Dirichlet distribution

Density defined for any $\boldsymbol{\theta} \in \Theta$, with $\Theta = \mathcal{S}(1, K)$

$$h(\boldsymbol{\theta}) = \frac{\Gamma(s)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1}$$

Generalisation of the Beta distribution

$$(\theta_1, 1 - \theta_1) \sim \mathit{Diri}(\alpha_1, \alpha_2) \iff \theta_1 \sim \mathit{Beta}(\alpha_1, \alpha_2)$$

Alternative parameterization

☐ **Dirichlet prior on $\boldsymbol{\theta}$**

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

☐ **Alternative parameterization** in terms of s, the total prior strength, and the relative prior strengths

$$\boldsymbol{t} = (t_1, \ldots, t_K) = \boldsymbol{\alpha}/s$$

with $t_k > 0$, $\sum_k t_k = 1$, *i.e.* $\boldsymbol{t} \in \mathcal{S}^\star(1, K)$

Hence,

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

☐ **Prior expectation** of θ_k

$$E(\theta_k) = t_k$$

☐ **Interpretation**

- $\boldsymbol{t}$ determines the center of the distribution
- s determines its dispersion / concentration

Dirichlet posterior for θ

Dirichlet posterior

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

$$\begin{aligned} \boldsymbol{\theta}|\boldsymbol{a} &\sim Diri(\boldsymbol{a}+\boldsymbol{\alpha}) \\ &\sim Diri(\boldsymbol{a}+s\boldsymbol{t}) \end{aligned}$$

Parameters/strengths of the Dirichlet play a role of counters: the prior strength α_k is incremented by the observed count a_k to give the posterior strength $a_k + \alpha_k$

Posterior expectation of θ_k

$$\begin{aligned} E(\theta_k|\boldsymbol{a}) &= \frac{a_k+\alpha_k}{n+s} \\ &= \frac{nf_k+st_k}{n+s} \end{aligned}$$

i.e. a weighted average of prior expectation, t_k, and observed frequency, f_k, with weights s and n

Objective Bayesian models

☐ Priors proposed for objective inference

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

☐ For multinomial sampling

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): $\alpha_k = \frac{1}{2}$ $(s = K/2)$
- Bayes-Laplace: $\alpha_k = 1$ $(s = K)$
- Berger-Bernardo reference priors

☐ For negative-multinomial sampling

Some proposed solutions for fixed a_k are *non-symmetric* Dirichlet priors

Which principles are satisfied? (1)

Prior ignorance

- Symmetry (SP). Yes: for all usual symmetric priors with $t_k = 1/K$. No: for some priors proposed for negative-sampling.
- Embedding Pcple (EP). Yes: for Haldane's prior. No: for all other priors

Internal consistency

- Coherence (CP), including ASL. Yes: if prior is proper. No: for Haldane's improper prior.

Frequentist interpretation

- Repeated sampling pcple (RSP). No in general. Yes asymptotically. Exact or conservative agreement for some procedures.

Which principles are satisfied? (2)

□ Invariance, Independence from irrelevant information

- Likelihood pcple (LP), including SRP. ??Yes, if prior $P(\boldsymbol{\theta})$ chosen independently of $P(\boldsymbol{a}|\boldsymbol{\theta})$. No, for Jeffreys' or Berger-Bernardo's priors
- Representation invariance (RIP). Yes: Haldane. No: all other priors
- Invariance by reparameterisation. Yes, for Jeffreys' or Berger-Bernardo's priors
- Specificity pcple? Yes: Haldane. No: all other priors

□ Difficulties of objective Bayesian approach

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

Focus on Haldane's prior

☐ Satisfies most principles

- Satisfies most of the principles: symmetry, LP, EP and RIP, specificity
- Incoherent because of improperness, but can be extended to a coherent model (Walley, 1991)

☐ But

- Improper prior
- Improper posterior if some $a_k = 0$
- Too data-glued:
  If $n = a_k = 1$, essentially says that $\theta_k = 1$ with probability 1.
  If $a_k = 0$, essentially says that $\theta_k = 0$ with probability 1.
- Doesn't give a reasonable account of uncertainty.

☐ Limit case of the IDM

Prior predictive distribution

From Bayes theorem

$$h(\boldsymbol{\theta}|\boldsymbol{a}) \;=\; \frac{h(\boldsymbol{\theta})\; P(\boldsymbol{a}|\boldsymbol{\theta})}{\int_{\Theta} h(\boldsymbol{\theta})\; P(\boldsymbol{a}|\boldsymbol{\theta})\; d\boldsymbol{\theta}}$$

Prior predictive distribution on $\boldsymbol{a}$

$$P(\boldsymbol{a}) \;=\; \int_{\Theta} h(\boldsymbol{\theta})\; P(\boldsymbol{a}|\boldsymbol{\theta})\; d\boldsymbol{\theta}$$

$$= \;\frac{h(\boldsymbol{\theta})\; P(\boldsymbol{a}|\boldsymbol{\theta})}{h(\boldsymbol{\theta}|\boldsymbol{a})}$$

which yields

$$P(\boldsymbol{a}) \;=\; \frac{\prod_k \binom{a_k+\alpha_k-1}{a_k}}{\binom{n+s-1}{n}}$$

with $\binom{m+x-1}{m} = \frac{\Gamma(m+x)}{m!\Gamma(x)}$, for any positive integer $m \geq 0$, and any real $x > 0$

Dirichlet-multinomial distribution

$$\boldsymbol{a} \;\sim\; DiMn(n;\boldsymbol{\alpha})$$

Posterior predictive distribution

- **Similarly, from Bayes theorem**

$$P(\boldsymbol{a}'|\boldsymbol{a}) = \frac{h(\boldsymbol{\theta}|\boldsymbol{a})\ P(\boldsymbol{a}'|\boldsymbol{\theta},\boldsymbol{a})}{h(\boldsymbol{\theta}|\boldsymbol{a}',\boldsymbol{a})}$$

$$= \frac{h(\boldsymbol{\theta}|\boldsymbol{a})\ P(\boldsymbol{a}'|\boldsymbol{\theta})}{h(\boldsymbol{\theta}|\boldsymbol{a}'+\boldsymbol{a})}$$

which yields

$$P(\boldsymbol{a}'|\boldsymbol{a}) = \frac{\prod_k \binom{a'_k+a_k+\alpha_k-1}{a'_k}}{\binom{n'+n+s-1}{n'}}$$

- **Dirichlet-multinomial posterior**

$$\boldsymbol{a}'|\boldsymbol{a} \sim DiMn(n';\boldsymbol{a}+\boldsymbol{\alpha})$$

- **Interpretation in terms of "counters"**

Here too, prior strengths $\boldsymbol{\alpha}$ are updated into posterior strengths $\boldsymbol{a}+\boldsymbol{\alpha}$

Bayesian answers to inference problems

- **Prior uncertainty**: $P(\boldsymbol{\theta})$ or $P(\boldsymbol{a})$
- **Posterior uncertainty**: $P(\boldsymbol{\theta}|\boldsymbol{a})$ or $P(\boldsymbol{a'}|\boldsymbol{a})$

For drawing all inferences, from observed data to unknowns (parameters or future data)

- **Inferences** about $\boldsymbol{\theta}$
  - Expectations, $E(\theta_k|\boldsymbol{a})$; Variances, $Var(\theta_k|\boldsymbol{a})$; *etc.*
  - Any event about $\boldsymbol{\theta}$: $P(\boldsymbol{\theta} \in \Theta^* \mid \boldsymbol{a})$
- **Inferences** about real-valued $\lambda = g(\boldsymbol{\theta})$
  - Marginal distribution function: $h(\lambda|\boldsymbol{a})$
  - $E(\lambda|\boldsymbol{a})$, $Var(\lambda|\boldsymbol{a})$
  - Cdf: $F_\lambda(u) = P(\lambda < u|\boldsymbol{a}) = \int_{-\infty}^{u} h(\lambda|\boldsymbol{a})\, d\lambda$
  - Credibility intervals: $P(\lambda \in [u_1; u_2] \mid \boldsymbol{a})$
  - Any event about λ

FROM PRECISE BAYESIAN MODELS TO AN IMPRECISE PROBABILITY MODEL

Precise Bayesian Dirichlet model

☐ Elements of a (precise) standard Bayesian model

- Prior distribution: $P(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$
- Sampling distribution: $P(\boldsymbol{a}|\boldsymbol{\theta})$, $\boldsymbol{a} \in \mathcal{A}$, $\boldsymbol{\theta} \in \Theta$
- Posterior distribution: $P(\boldsymbol{\theta}|\boldsymbol{a})$, $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{a} \in \mathcal{A}$, obtained by Bayes' theorem

☐ Elements of a precise Dirichlet model

- Dirichlet $P(\boldsymbol{\theta})$
- Multinomial $P(\boldsymbol{a}|\boldsymbol{\theta})$
- Dirichlet $P(\boldsymbol{\theta}|\boldsymbol{a})$

☐ Restriction: proper prior and posterior

Probability *vs.* Prevision (1)

☐ Three distributions

$$P(\theta) \quad P(a|\theta) \quad P(\theta|a)$$

These are probability distributions, which allocate a mass probability (or a probability density) to any event relative to θ and/or a.

☐ From probability of events to previsions of gambles

Since each one is a precise model, each defines a unique linear prevision for each possible gamble. So, each $P(\cdot)$ or $P(\cdot|\cdot)$ can be assimilated to a linear prevision

☐ Domains of these linear previsions

Here, we always consider all possible gambles, so these linear previsions are each defined on the linear space of all gambles (on their respective domains).

Probability *vs.* Prevision (2)
Remarks

☐ **Remark on terms used**

- Random quantity, or variable = Gamble
- Expectation = Linear prevision

☐ **Previsions are more fundamental than probabilities**

- Precise world:

  Linear previsions $\Longleftrightarrow$ Probabilities

- Imprecise world:

  Linear previsions $\Longrightarrow$ Probabilities

☐ **See** (de Finetti, 1974-75; Walley, 1991)

Coherence of a standard Bayesian model

Coherence of these linear previsions

- If prior is proper, then $P(\boldsymbol{\theta})$ is coherent
- $P(\boldsymbol{a}|\boldsymbol{\theta})$ always coherent
- If prior is proper, then posterior is proper, and hence $P(\boldsymbol{\theta}|\boldsymbol{a})$ is coherent

Joint coherence (Walley, 1991, Thm. 7.7.2)

- The linear previsions, $P(\boldsymbol{\theta})$, $P(\boldsymbol{a}|\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\boldsymbol{a})$ are jointly coherent
- The generalized Bayes' rule, which extends a coherent lower prevision to its natural extension, reduces to Bayes' rule in the case of a linear prevision.

Class of coherent models

□ **One privileged way** of constructing coherent imprecise posterior probabilities

"... is to form the lower envelopes of a class of standard Bayesian priors and the corresponding class of standard Bayesian posteriors"
(Walley, 1991, p. 397)

□ **Lower envelope theorem** (id., Thm. 7.1.6)

The lower envelope of a class of separately coherent lower previsions, is a coherent lower prevision.

□ **Class of Bayesian models** (id., Thm. 7.8.1):

Suppose that $P_\gamma(\cdot)$, $P_\gamma(\cdot|\Theta)$ and $P_\gamma(\cdot|\mathcal{A})$ constitute a standard Bayesian model, for every $\gamma \in \Gamma$. Then their lower envelopes, $\underline{P}(\cdot)$, $\underline{P}(\cdot|\Theta)$ and $\underline{P}(\cdot|\mathcal{A})$ are coherent.

On the road to the IDM

□ Building an Imprecise Dirichlet model

- Class of Dirichlet priors
- A single precise sampling model
- Update each prior, using Bayes' theorem
- Class of Dirichlet posteriors
- Form the associated posterior lower prevision

□ Draw inferences from IDM

- Report lower probabilities of events of interest
- Report lower previsions of random quantities of interest
- Requires optimization: minimizing or maximizing

□ Which prior class?

□ Yielding which properties?

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 $s > 0$:

$$\mathcal{M}_0 \;=\; \{Diri(s\boldsymbol{t}) :\; \boldsymbol{t} \in \mathcal{T} = \mathcal{S}^\star(1, K)\}$$

Updating

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

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

Posterior IDM

Posterior uncertainty about $\boldsymbol{\theta}$, conditional on $\boldsymbol{a}$, is expressed by the set

$$\mathcal{M}_n \;=\; \{Diri(\boldsymbol{a} + s\boldsymbol{t}) :\; \boldsymbol{t} \in \mathcal{T} = \mathcal{S}^\star(1, K)\}.$$

Upper and lower probabilities

Prior L&U 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 L&U probabilities

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

Posterior uncertainty about B is expressed by

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

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

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

- **Derived parameter of interest** (real-valued)

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

Inferences about λ can be summarized by

- **L&U expectations**

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

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

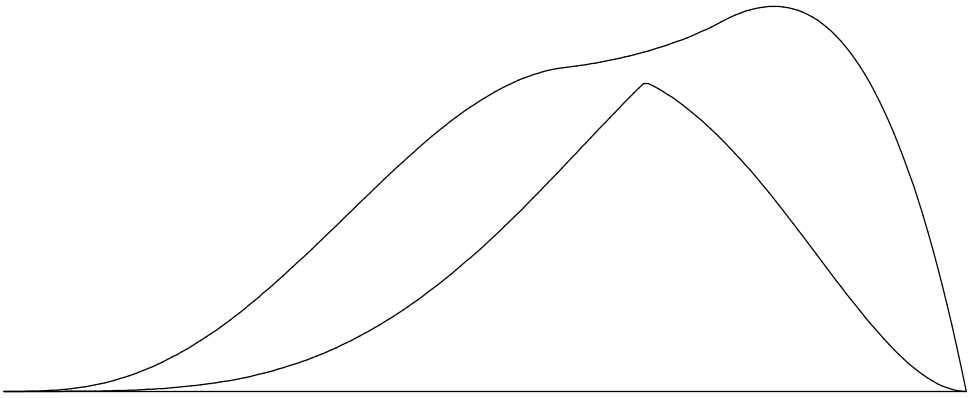
- **L&U cumulative distribution fonctions (cdf)**

$$\begin{aligned} \underline{F}_\lambda(u|\boldsymbol{a}) &= \underline{P}(\lambda \leq u|\boldsymbol{a}) \\ \overline{F}_\lambda(u|\boldsymbol{a}) &= \overline{P}(\lambda \leq u|\boldsymbol{a}) \end{aligned}$$

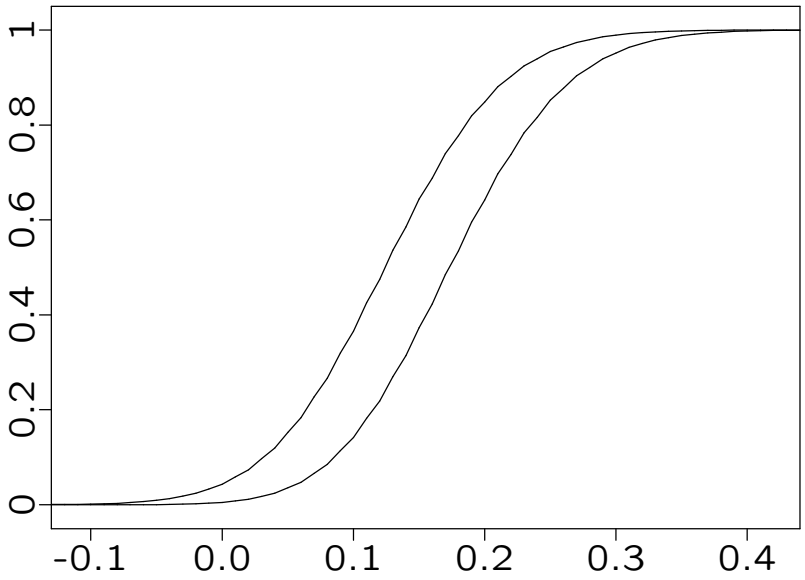
obtained by min-/maximization of $P_{\boldsymbol{st}}(\lambda \leq u|\boldsymbol{a})$ *w.r.t.* $\boldsymbol{t} \in \mathcal{S}^\star(1, K)$,

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

- **L&U df's**, $\lambda = \theta_k$



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



Optimization problems (1)

☐ Set or convex combinations?

The prior set, $\mathcal{M}_0$, and the posterior set, $\mathcal{M}_n$, of *Diri* distributions, are essentially used to define lower previsions $\underline{P}(\cdot)$ (taking lower envelopes). Each $\underline{P}(\cdot)$ is equivalent to its class of dominating linear previsions, which contains: the original set + all convex combinations.

☐ Optimization of $\mathbf{E}_{st}(\lambda)$ or $\mathbf{E}_{st}(\lambda|\boldsymbol{a})$

Since $E(\cdot)$ is linear, only requires optimization on the original set of Dirichlet's, $\mathcal{M}_0$ or $\mathcal{M}_n$.

☐ Optimization of $\mathbf{F}_{st,\lambda}(u)$, or $\mathbf{F}_{st,\lambda}(u|\boldsymbol{a})$

Similarly, since $(\lambda \leq u)$ is an event, $F(\cdot)$ is the probability of an event, *i.e.* the expectation of the corresponding indicator function. Hence, only the original set is required $\mathcal{M}_0$ or $\mathcal{M}_n$.

Optimization problems (2)

Optimization attained

- often by corners for $\boldsymbol{t} \in \mathcal{T}$, *i.e.* when some $t_k \to 1$, and all others tend to 0,
- but, not always

Some open questions

- Class of problems for which corners provide the optimization?
- The two previous min-/maximization problems, of the expectation and of the cdf of λ, have the same (swapping lower and upper) solution, in general?
- Or for some particular class of functions $\lambda = g(\boldsymbol{\theta})$ to be found?

Partial answers:

If $\lambda = g(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$, then YES.
Exact and approximate answers for more general λ in (Hutter, 2003).

Inferences about θ_k from the IDM

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

Expectations

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

Cdf's

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

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

Haldane + s extreme observations.

Extreme IDM's (1)

Two extremes

- $s \to 0$: Haldane's model
- $s \to \infty$: vacuous model

Haldane's model: $s \to 0$

- Unreasonable account of prior uncertainty
- Inferences over-confident with extreme data
- You learn too quickly!

Extreme IDM's (2)

Vacuous model: $s \to \infty$

- The IDM(s_{sup}) contains all IDM's with $s \leq s_{sup}$, *i.e.* all $Diri_{\boldsymbol{st}}$, $s \leq s_{sup}$, $\boldsymbol{t} \in \mathcal{T}$ At the limit, the IDM($s_{sup} \to \infty$) contains all Dirichlet's
- Hence, the IDM($s_{sup} \to \infty$) contains all mixtures (convex combinations) of Dirichlet's
- But, any distribution on Θ can be approximated by a finite convex mixture of Dirichlet's. So, the IDM($s_{sup} \to \infty$), contains all distributions on Θ
- Leads to vacuous statements for any gamble, and for both prior and posterior inferences
- You never learn anything!

Conclusions

- $s \to 0$: Too precise!
- $s \to \infty$: Too imprecise!

Hyperparameter 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

Hyperparameter s must be small

- If too high, inferences are too weak

Hyperparameter s must be large enough to

- Encompass objective Bayesian inferences: Haldane: $s > 0$; Perks: $s \geq 1$
- Encompass frequentist inferences

Suggested values: $s = 1$ or $s = 2$

IMPORTANT DISTRIBUTIONS

Relevant distributions

Parametric inference

- Dirichlet, any K
- Beta, $K = 2$

Predictive inference

- Dirichlet-Multinomial, any K
- Beta-Binomial, $K = 2$

Dirichlet distribution

- **Consider**

$$\begin{aligned} \boldsymbol{\theta} &= (\theta_1, \ldots, \theta_K) \quad \boldsymbol{\theta} \in \Theta = \mathcal{S}(1, K) \\ \boldsymbol{t} &= (t_1, \ldots, t_K) \quad \boldsymbol{t} \in \mathcal{T} = \mathcal{S}^\star(1, K) \end{aligned}$$

and $s > 0$, $\boldsymbol{\alpha} = s\boldsymbol{t}$

- **Dirichlet density**

$$\boldsymbol{\theta} \sim Diri(\boldsymbol{\alpha}) = Diri(s\boldsymbol{t})$$

$$\begin{aligned} h(\boldsymbol{\theta}) &\propto \theta_1^{\alpha_1} \ldots \theta_K^{\alpha_K - 1} \\ &= \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(s)} \theta_1^{\alpha_1} \ldots \theta_K^{\alpha_K - 1} \end{aligned}$$

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

$$(\theta_1, \theta_2) \sim Diri(\alpha_1, \alpha_2) \iff \theta_1 \sim Beta(\alpha_1, \alpha_2)$$

Basic properties

- $E(\theta_k) = t_k$
- s determines dispersion of distribution

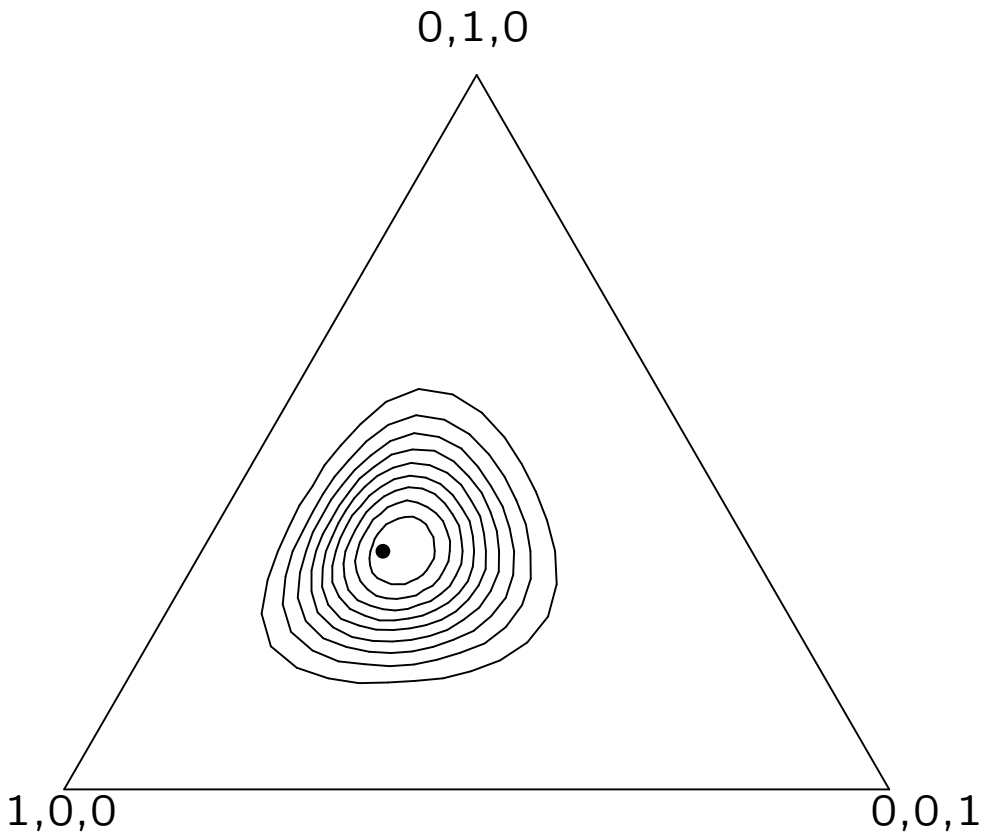
Examples of Dirichlet's

- **Example 1**

  $Diri(1, 1, \ldots, 1)$ 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 $(\theta_1, \ldots, \theta_5) \sim Diri(\alpha_1, \ldots, \alpha_5)$. Then,

□ Pooling property

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

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

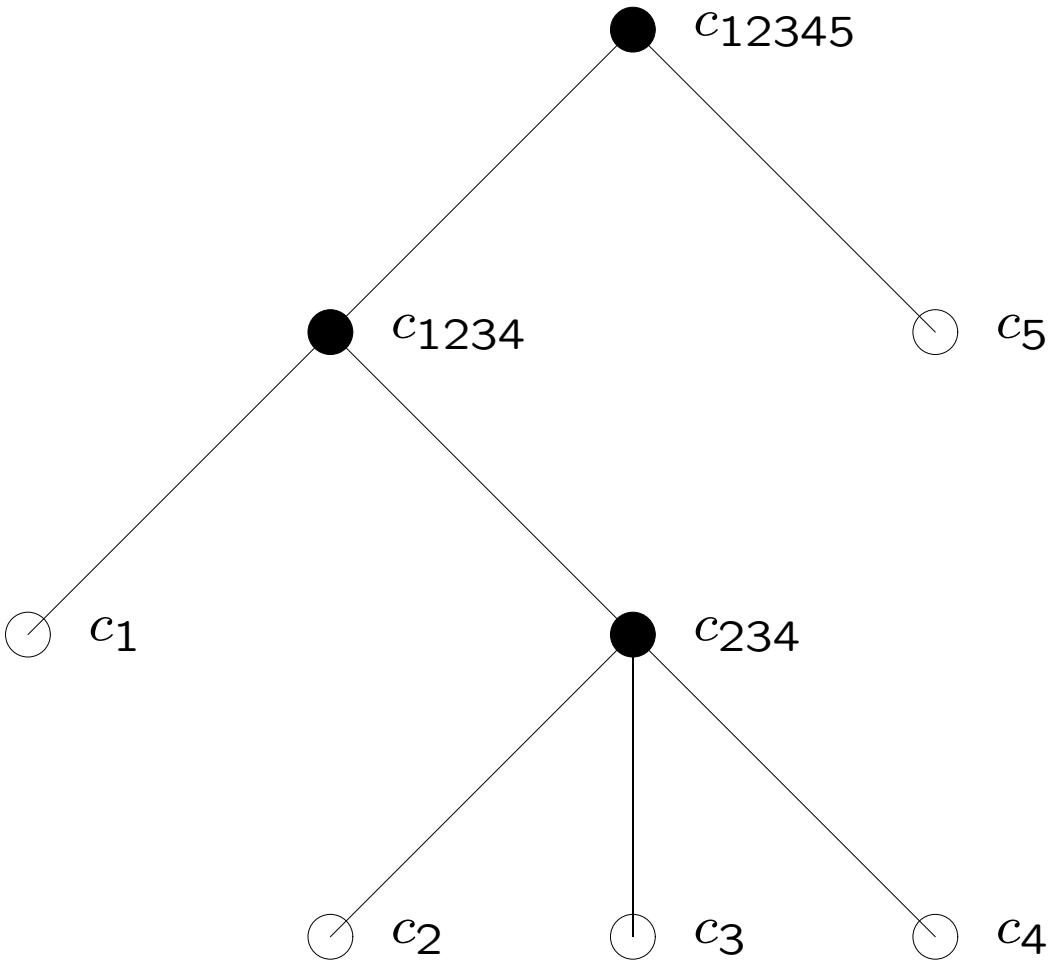
$$\boldsymbol{\theta}_T \sim Diri(\boldsymbol{\alpha}_T)$$

□ Restriction property

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

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

Tree representation of categories



"Node-cutting" a Dirichlet (1)

□ **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

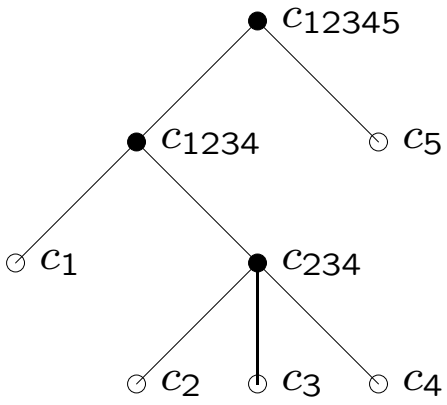
$$\begin{aligned} \boldsymbol{\theta}_{\overline{T}} &\sim Diri(\boldsymbol{\alpha}_{\overline{T}}) \\ \boldsymbol{\theta}_{\underline{T}} &\sim Diri(\boldsymbol{\alpha}_{\underline{T}}) \\ \boldsymbol{\theta}_{\overline{T}} &\perp\!\!\!\perp \boldsymbol{\theta}_{\underline{T}} \end{aligned}$$

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

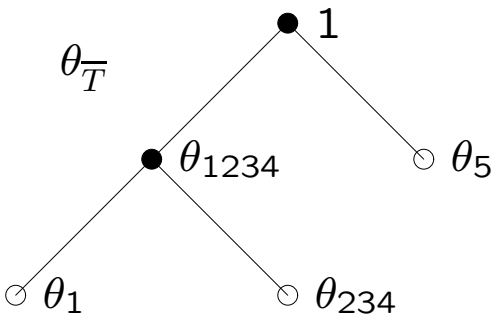
□ **Key** to computations of the Dirichlet.

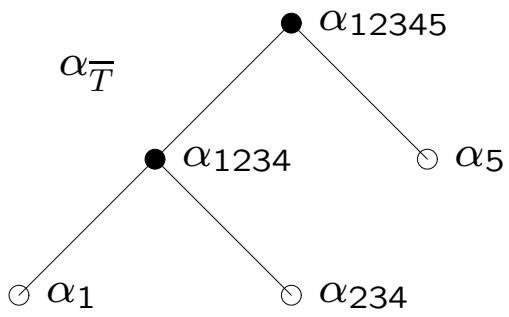
"Node-cutting" a Dirichlet (2)

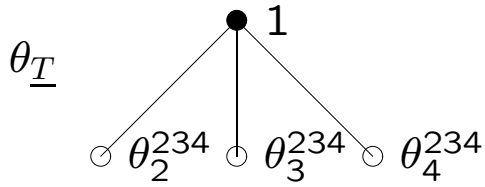
□ Set C and underlying tree T

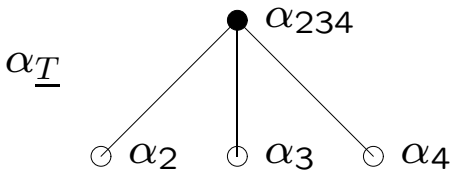


□ Cut at node c_{234}









Dirichlet-Multinomial distribution

Notation

$$\boldsymbol{a} \sim DiMn(n; \boldsymbol{\alpha})$$

for $\boldsymbol{a} = (a_1, \ldots, a_K)$, a_k positive integers,
with $\sum_k a_k = n$;
and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, α_k strictly positive reals,
with $\sum_k \alpha_k = s$

Probability distribution function

$$P(a_1, a_2) = \frac{\prod_k \binom{a_k + \alpha_k - 1}{a_k}}{\binom{n+s-1}{n}}$$

$$= \prod_k \frac{\Gamma(a_k + \alpha_k)}{a_k! \Gamma(\alpha_k)} \frac{n! \Gamma(s)}{\Gamma(n+s)}$$

Beta-Binomial distribution

Notation

$$(a_1, a_2) \sim BeBi(n; \alpha_1, \alpha_2)$$

for a_1 and a_2 positive integers, with $a_1 + a_2 = n$ and α_1 and α_2 strictly positive real-valued, with $\alpha_1 + \alpha_2 = s$

Probability distribution function

$$\begin{aligned} P(a_1, a_2) &= \frac{\binom{a_1+\alpha_1-1}{a_1}\binom{a_2+\alpha_2-1}{a_2}}{\binom{n+s-1}{n}} \\ &= \frac{\Gamma(a_1+\alpha_1)}{a_1!\Gamma(\alpha_1)} \frac{\Gamma(a_2+\alpha_2)}{a_2!\Gamma(\alpha_2)} \frac{n!\Gamma(s)}{\Gamma(n+s)} \end{aligned}$$

PROPERTIES OF THE IDM

Pooling categories

□ Pooling property of Dirichlet distributions

Any coarsening (or pooling) of categories, preserves both

- the Dirichlet form
- the value of s

All strengths and counts of the pooled categories are summed, whether prior or posterior, whether absolute or relative.

□ For the IDM

Thus, in the IDM, pooling categories induces an IDM with less categories, but with the same s.

Properties & principles

□ Prior ignorance about C and K

- Symmetry in the K categories
- Embedding pcple (EP) satisfied, due to the pooling property

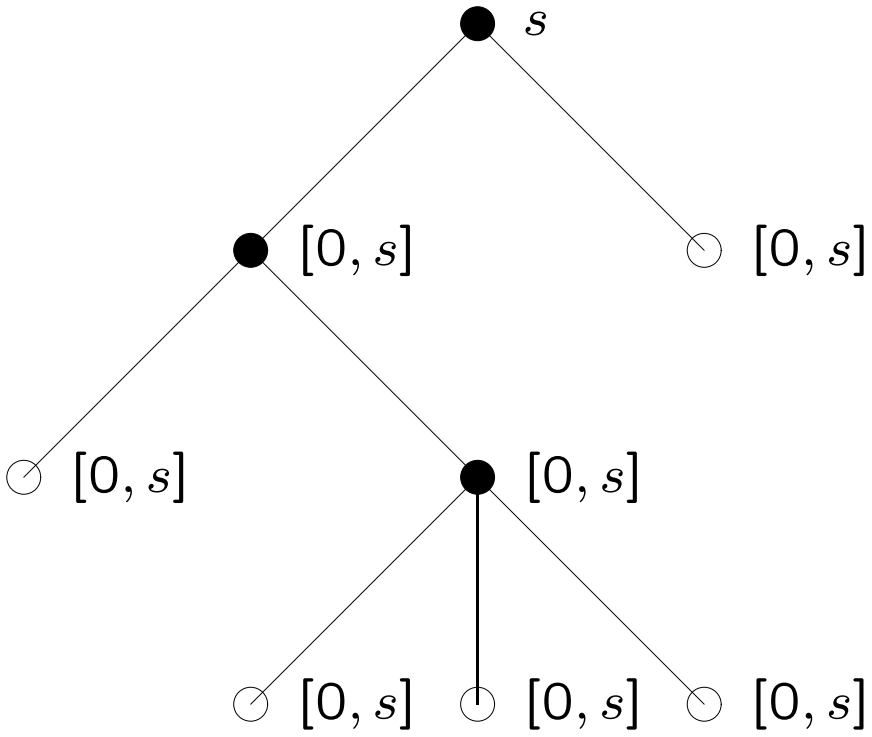
□ Prior near-ignorance about θ

- Near-ignorance properties: $E(\theta_k)$ and $F_{\theta_k}(u)$ are vacuous
- Many other events, or derived parameters, have vacuous prior probabilities, or previsions
- But not all, unless $s \to \infty$

□ Posterior inferences

- Satisfy coherence (CP)
- Satisfy the likelihood principle (LP)
- Representation invariance (RIP) is satisfied, for the same reason as EP is

Why does the IDM satisfy the EP and 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.

PREDICTIVE INFERENCE FROM THE IDM

Bayesian inference (recall)

□ **Apply Bayes' theorem once**

$$\left\{\begin{array}{c}\text{Prior } P(\boldsymbol{\theta}) \\ + \\ \text{Sampling } P(\boldsymbol{a}|\boldsymbol{\theta})\end{array}\right. \longrightarrow \left\{\begin{array}{c}\text{Posterior } P(\boldsymbol{\theta}|\boldsymbol{a}) \\ + \\ \text{Prior predictive } P(\boldsymbol{a})\end{array}\right.$$

□ **Apply Bayes' theorem a second time**

$$\left\{\begin{array}{c}\text{Prior' } P(\boldsymbol{\theta}|\boldsymbol{a}) \\ + \\ \text{Sampl.' } P(\boldsymbol{a}'|\boldsymbol{\theta},\boldsymbol{a})\end{array}\right. \longrightarrow \left\{\begin{array}{c}\text{Posterior' } P(\boldsymbol{\theta}|\boldsymbol{a}',\boldsymbol{a}) \\ + \\ \text{Post. pred. } P(\boldsymbol{a}'|\boldsymbol{a})\end{array}\right.$$

□ **Learning model** about

- unknown chances: $P(\boldsymbol{\theta})$ updated to $P(\boldsymbol{\theta}|\boldsymbol{a})$
- future data: $P(\boldsymbol{a})$ updated to $P(\boldsymbol{a}'|\boldsymbol{a})$

Bayesian prediction from a single Diri(α) prior

Dirichlet-multinomial prior

$$\boldsymbol{a} \sim DiMn(n; \boldsymbol{\alpha})$$

$$P(\boldsymbol{a}) = \frac{\prod_k \binom{a_k+\alpha_k-1}{a_k}}{\binom{n+s-1}{n}}$$

Dirichlet-multinomial posterior

$$\boldsymbol{a'}|\boldsymbol{a} \sim DiMn(n'; \boldsymbol{a}+\boldsymbol{\alpha})$$

$$P(\boldsymbol{a'}|\boldsymbol{a}) = \frac{\prod_k \binom{a'_k+a_k+\alpha_k-1}{a'_k}}{\binom{n'+n+s-1}{n'}}$$

Generalized binomial coefficients

$$\binom{m+x-1}{m} = \frac{\Gamma(m+x)}{m!\,\Gamma(x)}$$

for integer $m \geq 0$, and real $x > 0$

Beta-binomial marginals under a single Diri(α) prior

- **Beta-binomial marginal prior** for a_k

$$a_k \sim BeBi(n;\ \alpha_k, s-\alpha_k)$$

$$P(a_k) = \frac{\binom{a_k+\alpha_k-1}{a_k}\binom{n-a_k+s-\alpha_k-1}{n-a_k}}{\binom{n+s-1}{n}}$$

- **Beta-binomial marginal posterior** for a'_k

$$a'_k|\boldsymbol{a} \sim BeBi(n';\ a_k+\alpha_k, n-a_k+s-\alpha_k)$$

$$P(a'_k|\boldsymbol{a}) = \frac{\binom{a'_k+a_k+\alpha_k-1}{a'_k}\binom{n'-a'_k+n-a_k+s-\alpha_k-1}{n'-a'_k}}{\binom{n'+n+s-1}{n'}}$$

Prior predictive distribution under the IDM

□ **Prior prediction** about $\boldsymbol{a}$ and $\boldsymbol{f} = \boldsymbol{a}/n$

Prior uncertainty about $\boldsymbol{a}$ is described by a set of *DiMn* distributions:

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

□ **Vacuous L&U prior expectations** of a_k and f_k

$$\underline{E}(a_k) \;=\; 0 \qquad \overline{E}(a_k) \;=\; n$$
$$\underline{E}(f_k) \;=\; 0 \qquad \overline{E}(f_k) \;=\; 1$$

obtained as $t_k \to 0$ and $t_k \to 1$ respectively

□ **Vacuous L&U prior cdf's** of a_k

(Notation: $F_k(u) = P(a_k \leq u)$, for $u = 0, \cdots, n$)

$$\underline{F}_k(u) \;=\; 0 \quad \text{if } 0 \leq u < n$$
$$\overline{F}_k(u) \;=\; 1 \quad \text{if } 0 \leq u \leq n$$

obtained as $t_k \to 1$ and $t_k \to 0$ respectively

Posterior predictive distribution under the IDM (1)

□ **Posterior prediction** about $\boldsymbol{a'}|\boldsymbol{a}$ and $\boldsymbol{f'}|\boldsymbol{a}$

Posterior uncertainty about $\boldsymbol{a'}$, conditional on $\boldsymbol{a}$, is described by the corresponding set of updated *DiMn* distributions:

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

□ **L&U posterior expectations** of a'_k and f'_k

$$\underline{E}(a'_k|\boldsymbol{a}) \;=\; n' \,\frac{a_k}{n+s} \qquad \overline{E}(a'_k|\boldsymbol{a}) \;=\; n' \,\frac{a_k+s}{n+s}$$

$$\underline{E}(f'_k|\boldsymbol{a}) \;=\; \frac{a_k}{n+s} \qquad \overline{E}(f'_k|\boldsymbol{a}) \;=\; \frac{a_k+s}{n+s}$$

obtained as $t_k \longrightarrow 0$ and $t_k \longrightarrow 1$ respectively

Posterior predictive distribution under the IDM (2)

□ **L&U posterior cdf's** of a'_k
(Notation: $F_k(u|\boldsymbol{a}) = P(a'_k \leq u|\boldsymbol{a})$, for $u = 0, \cdots, n'$)

$$\underline{F}_k(u|\boldsymbol{a}) = \sum_{a'_k=0}^{u} \frac{\binom{a'_k+a_k+s-1}{a'_k}\binom{n'-a'_k+n-a_k-1}{n'-a'_k}}{\binom{n'+n+s-1}{n'}}$$

$$\overline{F}_k(u|\boldsymbol{a}) = \sum_{a'_k=0}^{u} \frac{\binom{a'_k+a_k-1}{a'_k}\binom{n'-a'_k+n-a_k+s-1}{n'-a'_k}}{\binom{n'+n+s-1}{n'}}$$

obtained as $t_k \to 1$ and $t_k \to 0$ respectively

An alternative view, the IDMM

Imprecise Dirichlet-multinomial model

Imprecise Dirichlet-multinomial model (IDMM) proposed by Walley & Bernard (1999), as a model for statistical inference about future observations $\boldsymbol{a'} = (a'_1, \ldots, a'_K)$, from observed data $\boldsymbol{a}$

- Hypothesis of exchangeability of all possible sequences of $n^* = n + n'$ observations; leads to a multi-hypergeometric likelihood
- Prior uncertainty about $\boldsymbol{a}^* = \boldsymbol{a} + \boldsymbol{a'}$ is described by a set of *Dirichlet-multinomial* distributions, $\{DiMn(n^*;\ s\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{S}^\star\}$ for any finite n^*

Equivalence with the IDM

The IDMM yields exactly the same predictive prior and posterior sets of *DiMn* distributions as the IDM for both $\boldsymbol{a}$ and $\boldsymbol{a'}|\boldsymbol{a}$.

The only difference is that the IDM requires unobservable parameters, $\boldsymbol{\theta}$, for making predictions, whereas the IDMM directly models observables.

Links between IDM and IDMM

Parametric and predictive inference

In general, in both precise Bayesian models and in the IDM,

- $\boldsymbol{\theta}$, $\boldsymbol{\theta}|a$ yields $\boldsymbol{f}$, $\boldsymbol{f'}|a$ (from Bayes' theorem)
- $\boldsymbol{f}$, $\boldsymbol{f'}|a$ yields $\boldsymbol{\theta}$, $\boldsymbol{\theta}|a$ (as $n' \to \infty$)

Equivalence between IDM and IDMM

- The IDM and the IDMM are equivalent, if we assume that n' can tend to infinity
- Any IDMM statement about $\boldsymbol{f'}$ which is independent of n' is also a valid IDM statement about $\boldsymbol{\theta}$

Two views of the IDMM

- The IDMM is the predictive side of the IDM
- The IDMM is a model of its own

Pooling categories

☐ **Pooling** categories c_k and c_l into c_j

$$\begin{aligned} a_j &= a_k + a_l \\ a'_j &= a'_k + a'_l \\ \alpha_j &= \alpha_k + \alpha_l \end{aligned}$$

☐ **Then**

- Each $DiMn_K$, prior or posterior, is transformed into a $DiMn_{K-1}$ where c_j replaces c_k and c_l, with these summed counts or strengths.
- Recursively, for any pooling in $J < K$ categories, both the *DiMn* form and the value of s are preserved.

☐ **Thus, in the IDMM,**

L&U prior and posterior probabilities for any event involving pooled counts with $J < K$ categories are invariant whether we

- Pool first, then apply IDMM
- Apply IDMM first, then pool

Properties of prior ignorance and posterior inferences

Prior ignorance

- Symmetry in the K categories
- Embedding pcple (EP) satisfied, due to the pooling property
- Near-ignorance properties: $E(a_k)$ and $F_k(.)$ are vacuous

Posterior inferences

- Satisfy coherence (CP)
- Satisfy the likelihood principle (LP)
- Representation invariance (RIP) satisfied, for the same reason as EP

Frequentist prediction

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

- Considers binomial or hypergeometric data ($K = 2$), $\boldsymbol{a} = (a_1, n - a_1)$.
- Studies the prediction about n' future observations, $\boldsymbol{a'} = (a'_1, n' - a'_1)$.
- Derives lower and upper confidence limits (frequentist) for a'_1.
- Compares these confidence limits to credibility limits (Bayesian) from a Beta prior.

□ **Main result**

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

□ **IDM with $s = 1$!**

These two *Beta* priors are the most extreme priors under the IDM with $s = 1$

Towards the IDM?
(Thatcher, 1964)

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)

Comments on predictive inference

□ Predictive approach is more fundamental (see, Geisser, 1993)

- Finite population & data
- Models observables only, not hypothetical parameters
- Relies on the exchangeability assumption only.
- Pearson (1920) considered predictive inference as "the fundamental problem of practical statistics"

□ Predictive approach is easier to understand, more natural

□ For the IDMM, in particular

- Gives the IDM as a limiting case as $n' \to \infty$
- Covers sampling with replacement from a finite population

THE RULE OF SUCCESSION

Rule of succession problem

Problem

- Prediction about the next observation
- Immediate prediction

A solution to it

- Called a rule of succession
- So many rules for such an (apparently) simple problem!

Highly debated problem

- Very early problem in Statistics
- Laplace computing the probability that the sun will rise tomorrow

Two types of problems / solutions

- Prior rule, before observing any data
- Posterior rule, after observing some data

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?"

Two problems of predictive inference

- Prior prediction, before observing any item
- Posterior prediction, after observing n items

Inference from a state of prior ignorance

about the proportions of the various colours

Notation

Event, elementary or combined

Let B_j be the event that the next observation is of type c_j, where c_j is a subset of C with J elements

$$1 \leq J \leq K$$

If $J = 1$, then $c_j = c_k$ is an elementary category
If $J > 1$, then c_j is a combined category

Define

The observed count and frequency of c_j

$$a_j = \sum_{k \in j} a_k \qquad f_j = \sum_{k \in j} f_k$$

The prior strength, and relative strength, of c_j from a *Diri*($\boldsymbol{\alpha}$) prior

$$\alpha_j = \sum_{k \in j} \alpha_k \qquad t_j = \sum_{k \in j} t_k$$

Rule of succession under a Diri(α)

Bayesian rule of succession

The rule of succession obtained from a single Dirichlet distribution, $Diri(\boldsymbol{\alpha}) = Diri(s\boldsymbol{t})$, is

$$\begin{aligned} P(B_j|\boldsymbol{a}) &= \frac{a_j + \alpha_j}{n + s} \\ &= \frac{nf_j + st_j}{n + s} \end{aligned}$$

The prior prediction, obtained for $n = a_j = 0$, is

$$P(B_j) = t_j$$

Generally

If data are multinomial, with parameters $\boldsymbol{\theta}$ and $\theta_j = \sum_{k \in j} \theta_k$, then

$$\begin{aligned} P(B_j) &= E(\theta_j) \\ P(B_j|\boldsymbol{a}) &= E(\theta_j|\boldsymbol{a}) \end{aligned}$$

Prior rule of succession under the IDM

Prior rule of succession

The L&U prior probabilities of B_j are vacuous:

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

obtained as $t_j \to 0$ and $t_j \to 1$ respectively

Prior ignorance

Prior imprecision is maximal, L&U probabilities are vacuous:

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

irrespectively of s

Posterior rule of succession under the IDM

Posterior rule of succession

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

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

obtained as $t_j \to 0$ and $t_j \to 1$ respectively

Posterior imprecision

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

L&U probabilities and f_j

The interval always contains $f_j = a_j/n$. The L&U probabilities both converge to f_j as n increases.

Rule independent from C, K and J

Rule of succession and imprecision

☐ Degree of imprecision about B_j

- Prior state: imprecision is maximal

$$\Delta(B_j) = 1$$

- Posterior state:

$$\Delta(B_j|\boldsymbol{a}) = \frac{s}{n+s}$$

☐ 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.

Objective Bayesian models

☐ Bayesian rule of succession

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

$$P(B_j|\boldsymbol{a}) = \frac{a_j + \alpha_j}{n+s} = \frac{nf_j + s\frac{J}{K}}{n+s}$$

☐ Objective Bayesian rules: $P(B_j|\boldsymbol{a}) =$

Haldane	a_j/n
Perks	$(a_j + J/K)/(n+1)$
Jeffreys	$(a_j + J/2)/(n + K/2)$
Bayes	$(a_j + J)/(n+K)$

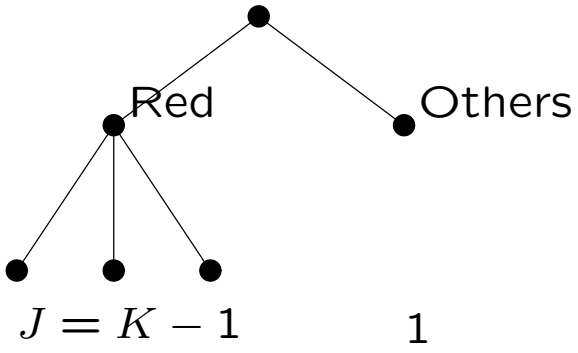
☐ Dependence on K and J except Haldane

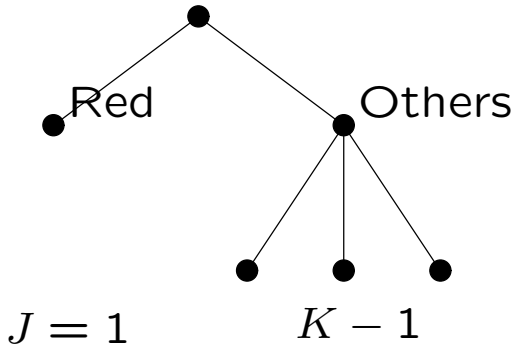
☐ Particular case $J = 1$, $K = 2$

If $a_j = n/2$, *i.e.* $f_j = 1/2$, each Bayesian rule leads to $P(B_j|\boldsymbol{a}) = 1/2$, whether $n = 0$, or $n = 10$, 100 or 1000.

Categorization arbitrariness

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





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

☐ Bayesian rules

Yield intervals when arbitrariness is introduced

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

Frequentist rule of succession

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

- Studies the particular case of immediate prediction

□ **Main result** (recall)

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

□ **Frequentist "rule of succession"**

When reinterpreted as Bayesian rules of succession, the lower and upper confidence limits respectively correspond to:

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

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

IMPRECISE BETA MODEL

IBM: Relevant papers

□ **IBM:** IDM with $K = 2$

□ **Imprecise Beta Model** (Walley, 1991)

Several models for binomial data, with classes of priors for θ of the form
$Beta(st, s(1-t))$ with $s_1 \leq s \leq s_2$, $t_1 \leq t \leq t_2$

□ **Bernoulli process** (Bernard, 1996, Am.Stat.)

Considers binary data, and compares several objective methods of inference, either frequentist or Bayesian, for one-sided tests about θ

□ **Frequentist inference and likelihood principle** (Walley, 2002, JSPI)

More general paper showing how imprecise probability models can be designed to encompass both frequentist and objective Bayesian models, and thus "reconcile frequentist properties with the likelihood principle"

Bernoulli process, Frequentist *vs.* Bayesian

☐ Data from a Bernoulli process

We observe a sequence of binary data (success / failure), *e.g.* the sequence

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

with unknown fixed chances of success, $\theta = \theta_{\text{suc}}$, and of failure, $1 - \theta = \theta_{\text{fai}}$.

Composition of the $n = 10$ observations is

$$a = a_{\text{suc}} = 8, \quad b = a_{\text{fai}} = 2$$

☐ Testing a one-sided hypothesis about θ

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

☐ In our example:

We observe a frequency of successes $f = 8/10$, can we conclude that $\theta > \theta_0 = 1/2$?

☐ **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 is greater than $f = \frac{8}{10}$, and add up their probabilities under H_0 (yielding "the" p-value of the test).

☐ **"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: $p_{inc} = P(F \geq f|H_0)$
- Exclusive: $p_{exc} = P(F > f|H_0)$
- Mid-P convention: $p_{mid} = (p_{exc} + p_{inc})/2$

Frequentist results

□ **Example:** p-value for a-rule and exclusive convention

$$PF_{ab}^{aE} = PF(F > f \mid H_0, a\text{-rule})$$

where $F = \frac{A}{A+B}$ is the frequency of a possible dataset, and $f = \frac{a}{a+b}$ is the frequency for the observed data.

For $a = 8$, $b = 2$ and $\theta_0 = 1/2$, we find

$$PF_{ab}^{aE} = \frac{20}{1024} = 0.020$$

□ **Frequentist conclusion**
The p-value, 0.020, is too small (smaller than *e.g.* the level 0.05). Hence H_0 is rejected, and H_1 is accepted: $\theta > 1/2$ at level 0.05

□ **But,** for the n-rule and the inclusive convention

$$PF_{ab}^{nI} = \frac{56}{1024} = 0.055$$

The p-value, 0.055, is too large, and H_0 cannot be rejected at level 0.05.

Frequentist p-values

□ Various p-values on our example

		Convention		
		Exc.	Mid-P	Inc.
Rule	n-rule	11	67/2	56
	a-rule	20	76/2	56
	b-rule	11	31/2	20

Table gives: $1024 \times p$-value

□ Range of p-values

These p-values range from

$$PF_{ab}^{nE} = \frac{11}{1024} \quad \text{and} \quad PF_{ab}^{nI} = \frac{56}{1024}$$

This is a general fact, always true for any (a, b) and any reference value θ_0 for θ.

Relations between p-values

□ **Define**

$$PF_{a,b} = PF^{aE}_{a,b}$$

For each couple of positive integers, (a, b), $PF_{a,b}$ can be computed recursively by

$$\begin{array}{rcl} PF_{a,0} = 0 & & PF_{0,b} = 1 \\ PF_{a+1,b+1} & = & \theta_0\, PF_{a,b+1} + (1-\theta_0)\, PF_{a+1,b} \end{array}$$

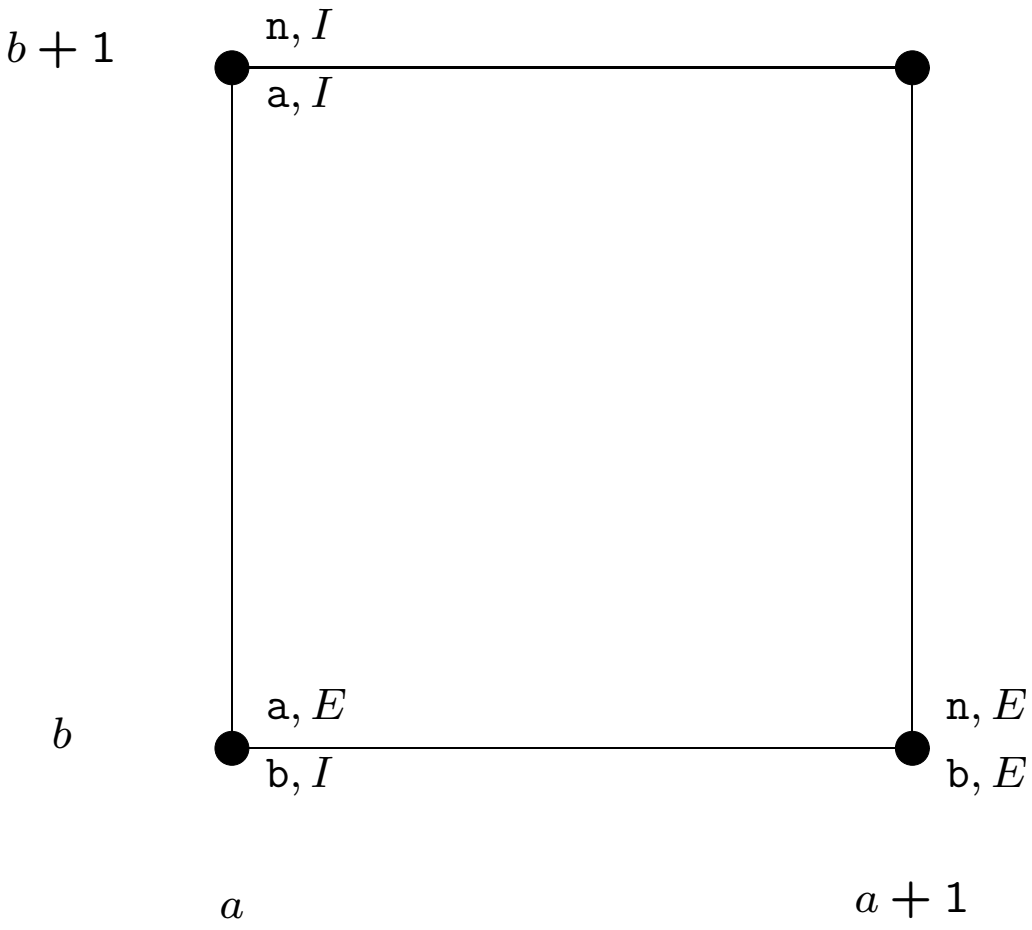
□ **Then** using a relation between binomial cdf and negative-binomial cdf,

$$\begin{array}{rcl} PF^{nI}_{a,b} & = & PF_{a,b+1} \\ PF^{nE}_{a,b} & = & PF_{a+1,b} \\ PF^{aI}_{a,b} & = & PF_{a,b+1} \\ PF^{aE}_{a,b} & = & PF_{a,b} \\ PF^{bI}_{a,b} & = & PF_{a,b} \\ PF^{bE}_{a,b} & = & PF_{a+1,b} \end{array}$$

Also, each mid-P p-value is intermediate between the corresponding inclusive and exclusive p-values

Frequentist levels map

- **Part of the map** giving $PF_{a,b}$ for any (a, b)



- **Summary of all frequentist levels**

$$PF_{a+1,b} \leq \text{all } PF \leq PF_{a,b+1}$$

Objective Bayesian approach

Principle

From the prior, $\theta \sim Beta(\alpha, \beta)$, the corresponding posterior is $\theta|(a,b) \sim Beta(a+\alpha, b+\beta)$, with density $h(\theta|(a,b),(\alpha,\beta))$, which yields

$$\begin{aligned} PB_{a,b}^{\alpha,\beta} &= P(H_0 \mid (a,b),(\alpha,\beta)) \\ &= \int_0^{\theta_0} h(\theta \mid (a,b),(\alpha,\beta))\, d\theta \end{aligned}$$

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

Objective Bayesian tests

Define

$$PB_{a,b} = PB_{a,b}^{0,0}$$

i.e. the posterior probability of H_0 (Bayesian level) under the Haldane's prior

Then,

$$PB_{a,b}^{\alpha,\beta} = PB_{a+\alpha,b+\beta}$$

Various Bayesian levels for H_0 ranging from

$$PB_{1,0} = \frac{11}{1024} \text{ to } PB_{0,1} = \frac{56}{1024}$$

including $\frac{20}{1024}$ (Haldane), $\frac{26.66}{1024}$ (Jeffreys, n-rule), $\frac{67/2}{1024}$ (Bayes-Laplace)

Summary of all objective Bayesian levels

From monotonicity considerations,

$$PB_{a+1,b} \leq \text{all } PB \leq PB_{a,b+1}$$

Link between frequentist and Bayesian

☐ Theorem (Bernard, 1996, Eq.(11))

For any (a, b) and θ_0,

$$PB_{a,b} \;=\; PF_{a,b}$$

☐ Consequence

$$PB_{a+1,b} \leq \text{ all } PB \text{ and } PF \leq PB_{a,b+1}$$

For any observed (a, b), and reference value θ_0, all frequentist levels considered, as well as all objective Bayesian levels, for the one-sided test of $H_0 : \theta \leq \theta_0$ are within the interval

$$[PB_{a+1,b} \,;\, PB_{a,b+1}]$$

whose bounds are obtained with the two extreme priors

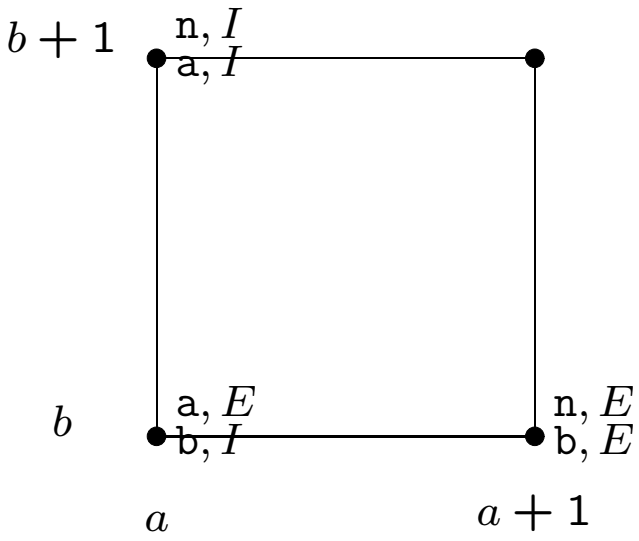
$$(\alpha = 1, \beta = 0) \quad (\alpha = 0, \beta = 1)$$

☐ Ignorance zone

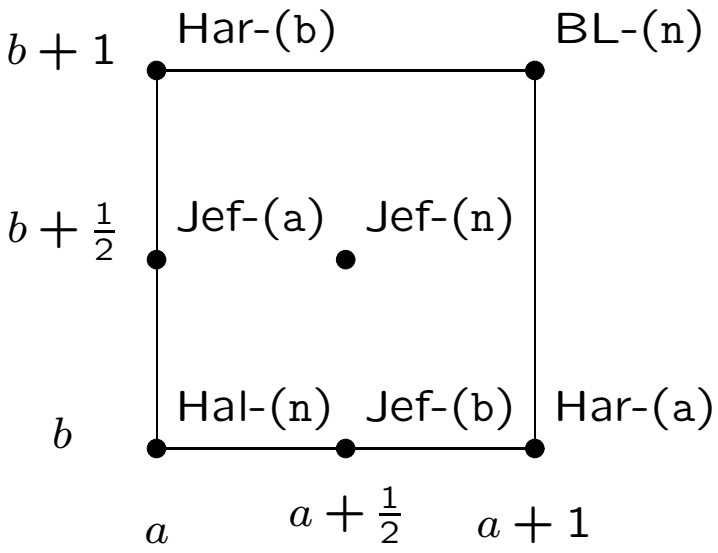
The bounds of this ignorance zone correspond to the *Imprecise Beta Model* (IBM) with $s = 1$.

Frequentist and Bayesian levels maps

Frequentist levels map (discrete)



Bayesian levels map (continuous)



Generalisation (Walley, 2002)

□ Monotone stopping rules

- Consider all binary sequences, each sequence yielding observed counts (a, b)
- Define the stopping region:
  $SR = \{(a, b); \text{process stopped at } (a, b)\}$
- The stopping rule is monotone *iff*

$$\text{If } (a, b) \in SR \text{ then } (a+1, b) \in SR \text{ and } (a, b+1) \in SR$$

□ Hypotheses H_0 considered

One-sided, and equi-tailed two-sided hypotheses

□ Walley shows that

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

□ Strong argument for $s \geq 1$ for the IDM

TWO BY TWO CONTINGENCY TABLES

Independence in a 2×2 contingency table

Data

	$j1$	$j2$
$i1$	a_{11}	a_{12}
$i2$	a_{21}	a_{22}

	$j1$	$j2$
$i1$	8	4
$i2$	2	5

Problem

Positive association between A and B?

Derived parameter: contingency coefficient

$$\rho = \frac{\theta_{11}}{\theta_{1.}\theta_{.1}} \qquad r_{obs} = 1.26$$

Hypothesis to be tested:

$$H_0 : \rho \leq 1 \quad \text{vs.} \quad H_1 : \rho > 1$$

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 $\boldsymbol{a}$ with the same margins than those observed.

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

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

The p-value of the test is defined as,

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

where "more extreme data" means all $\boldsymbol{a}$ 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 $H_0 : \rho \geq 1$ *vs.* $H_1 : \rho < 1$.

Bayesian & Imprecise models

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

Haldane, Perks, Jeffreys, Bayes-Laplace

- **IBM** (precisely product-IBM)

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

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

- **IDM**, with $s = 1$ or $s = 2$

- **Relations** between models

$$\begin{aligned}
&\underline{P}[\text{IDM}_2] \leq \underline{P}[\text{IBM}] = p_{exc} \leq \underline{P}[\text{IDM}_1] \\
&\leq PB[\text{Hal}],\ PB[\text{Per}],\ PB[\text{Jef}],\ PB[\text{BL}] \leq \\
&\overline{P}[\text{IDM}_1] \leq \overline{P}[\text{IBM}] = p_{inc} \leq \overline{P}[\text{IDM}_2]
\end{aligned}$$

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$(s=2)$

LARGE n AND POSTERIOR IMPRECISION

Large n, Bayesian models and IDM

☐ Claim by Bayesians

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

This claim is also sometimes present in 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 multivariate binary data (Bernard, 2001)

Inference about a single chance θ

Problem

- Observed counts $\boldsymbol{a} = (a_1, a_2)$, $n = a_1 + a_2$
- Test $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$

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

$$\begin{aligned} \underline{P}(\theta \leq \theta_0|\boldsymbol{a}) &= P(A_1 > a_1|H_0, n) \\ \overline{P}(\theta \leq \theta_0|\boldsymbol{a}) &= P(A_1 \geq a_1|H_0, n) \end{aligned}$$

$$\begin{aligned} \Delta(\theta \leq \theta_0|\boldsymbol{a}) &= P(A_1 = a_1|H_0, n) \\ &= \binom{n}{a_1}\theta_0^{a_1}(1-\theta_0)^{a_2} \end{aligned}$$

Example: $a_1 = 0$, $a_2 = 100$, $\theta_0 = 0.001$

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

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

Association in 2×2 tables

- **Example** $n = 115$

	$i1$	$i2$
$i1$	0	4
$i2$	4	107

- **Fisher's test:** $H_0 : \rho \geq 1$ *vs.* $H_1 : \rho < 1$

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

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

  Haldane: $P(H_1|\boldsymbol{a}) = 0$
  Perks: $P(H_1|\boldsymbol{a}) = 0.350$
  Jeffreys: $P(H_1|\boldsymbol{a}) = 0.571$
  Bayes: $P(H_1|\boldsymbol{a}) = 0.802$

- **IDM answers**

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

- **Why?** Independence is compatible with data (despite $a_{11} = 0$), 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 $P(\boldsymbol{a}|H_0)$ is high.

□ **Consequences for the IDM**

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

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 $\boldsymbol{a} = (a_1, \ldots, a_K)$.

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,

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

Inferences from the IDM

Prior expectations

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

Posterior expectations

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

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

$$\mathit{L\text{-}Diri}(\boldsymbol{y}, \boldsymbol{a} + \boldsymbol{\alpha} = (a_1 + n, a_2, \ldots, a_K))$$

$$\mathit{L\text{-}Diri}(\boldsymbol{y}, \boldsymbol{a} + \boldsymbol{\alpha} = (a_1, \ldots, a_{K-1}, \ldots, a_K + n))$$

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.

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.

IDM's nested according to s

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

$$Int[s_1] \subset Int[s_2]$$

Learning parameter

Number of additional observations

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

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

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 $s \geq K$, but they produce unreasonable inferences when K large (categorization arbitrariness, inference on a mean).
- "reference priors": 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.

COMPUTATIONAL ASPECTS

Computational aspects

General problem

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

- Simple (and identical) solution to both problems when $g(.)$ is linear: $t_k \rightarrow 1$ 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 $t_k \rightarrow 1$ for some k provides the solution.
- Is saying $t_k \rightarrow 1$ enough to specify the optimization 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?

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)

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 θ 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.