Predictive inference: From Bayesian inference to Imprecise Probability

Jean-Marc Bernard
University Paris Descartes
CNRS UMR 8069

Third SIPTA School on
Imprecise Probabilities

Montpellier, France
7 July 2008

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)

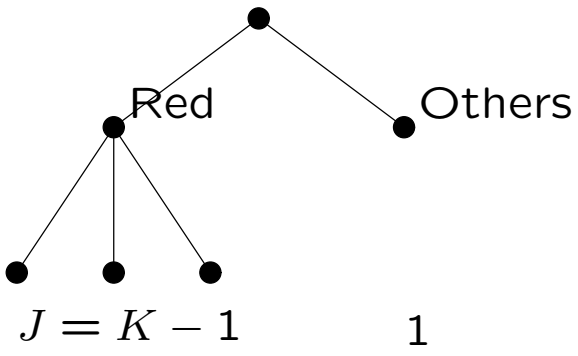
Categories

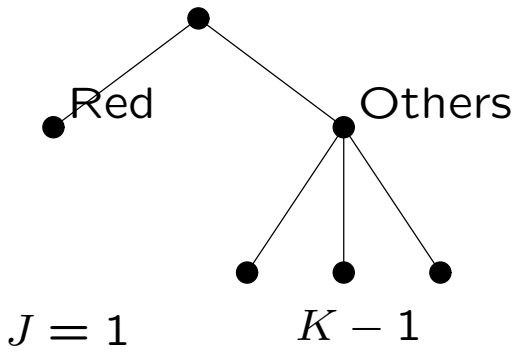
- Set of K of categories or types

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

- 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

$$\begin{array}{ccc} I & \to & C = \{c_1, \ldots, c_K\} \\ i & \mapsto & c_k \end{array}$$

- 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 (1)

☐ Inference about what?

- Predictive inference: About future counts or frequencies in n' future observations

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

$n' \geq 1$ Predictive inference (general)
$n' = 1$ Immediate prediction

- Parametric inference: About true/parent counts or frequencies (parameters) in population of

... size $N < \infty$

$$\begin{aligned} \boldsymbol{A} &= (A_1, \ldots, A_K) \\ \boldsymbol{\theta} &= (\theta_1, \ldots, \theta_K) = \boldsymbol{A}/N \end{aligned}$$

... size $N = \infty$

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \quad \sum_k \theta_k = 1$$

Statistical inference problems (2)
Prior *vs.* posterior inferences

☐ Prior inferences

- $n = 0$ (no data yet)
- Unconditional
- Describes prior uncertainty about $\boldsymbol{f'}$ or $\boldsymbol{\theta}$
- Issue: formalize prior ignorance

☐ Posterior inferences

- $n \geq 1$ (data $\boldsymbol{a}$ are available)
- Conditional on $\boldsymbol{a}$
- Describes what can be infered about $\boldsymbol{f'}$ or $\boldsymbol{\theta}$ from the prior state + the knowledge of $\boldsymbol{a}$

Relating past & future data (1)
Random sampling

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

Finite/infinite population

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

Stopping rule

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

These elements define a sampling model

Relating past & future data (2) 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

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

□ Equivalence with MHyp sampling

Induced $P(\boldsymbol{a}|\boldsymbol{a}^*)$ is the same as if data with counts $\boldsymbol{a}$ were obtained from random sampling from a population having counts $\boldsymbol{a}^* = \boldsymbol{a} + \boldsymbol{a'}$

□ **Direct link**: No need to invoke unknown parameters $\boldsymbol{\theta}$ of a larger population

A statistical challenge

☐ Model prior ignorance

- Model prior ignorance about $\boldsymbol{\theta}$, or $\boldsymbol{a}$ and $\boldsymbol{a}^*$
- Arbitrariness of C and K, both may vary as data items are observed
- Model prior ignorance about both the set C and the number K of categories

☐ Make reasonable posterior 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): Posterior inferences should not depend on refinements or coarsenings of categories

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 *e.g.* $\boldsymbol{a}|\boldsymbol{\theta}$
- 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
- Conjugate priors: Dirichlet for multinomial data, Dirichlet-multinomial for multiple-hypergeometric data
- Depend on the sampling model through the likelihood function only

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 parametric 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.

The IDMM in brief

□ **Model for predictive inference** for categorical data
Proposed by Walley, Bernard (1999), also partly studied in (Walley, 1996).
Inference about future data $\boldsymbol{a'} = (a'_1, \ldots, a'_K)$ from observed data $\boldsymbol{a} = (a_1, \ldots, a_K)$, categorized in K categories C.

□ **Two alternative, equivalent views**

- A predictive model derived from the parametric IDM
- A model of its own, modeling only observables: available data $\boldsymbol{a}$ and future data $\boldsymbol{a'}$

□ **Imprecise probability model**
Prior uncertainty about $\boldsymbol{a}$ expressed by a set of Dirichlet-multinomial distributions.
Posterior uncertainty about $\boldsymbol{a'}|\boldsymbol{a}$ then described by a set of updated Dirichlet-multinomial distributions.

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

Outline

1. Introduction
2. Bayesian approach to inference
3. Important distributions
4. Objective Bayesian models
5. From Bayesian to imprecise probability models

6. Definition of the IDM & the IDMM
7. Predictive inferences from the IDMM
8. The rule of succession
9. Conclusions

References

THE BAYESIAN APPROACH

Bayesian inference

☐ **Focus on the Bayesian approach** since

- Bayesian, precise: a single Dirichlet prior on $\boldsymbol{\theta}$ yields a single Dirichlet posterior on $\boldsymbol{\theta}|\boldsymbol{a}$ (PDM)
- IP-model: a prior set of Dirichlet's yields a posterior set of Dirichlet's (IDM)

☐ $\cdots$ **and for predictive inferences** since

- Bayesian, precise: a single Dirichlet-Multinomial (*DiMn*) prior on $\boldsymbol{a}^*$ yields a single *DiMn* posterior on $\boldsymbol{a}'|\boldsymbol{a}$ (PDMM)
- IP-model: a prior set of *DiMn*'s yields a posterior set of *DiMn*'s (IDMM)

☐ **Goal**

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

Three sampling models

Multinomial data

- Random sampling
- Infinite population, $N = \infty$
- Data have a multinomial (*Mn*) likelihood

Multiple-hypergeometric data

- Random sampling
- Finite population, $N < \infty$
- Data have a multiple-hypergeometric (*MHyp*) likelihood

Exchangeable data

- Data $\boldsymbol{a}$ generated by an exchangeable process with counts $\boldsymbol{a}^* = \boldsymbol{a} + \boldsymbol{a}'$
- Data have a *MHyp* likelihood too

Hypotheses

- Set C, and number of categories, K, are considered as known and fixed

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 (1): 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 $P(\boldsymbol{a'}|\boldsymbol{a})$

Bayesian inference (2)

□ 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 $\theta|a$

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

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 $\boldsymbol{a}+\boldsymbol{\alpha}$

Equivalence of 3 models for predictive inference

□ Multinomial + Dirichlet model

$$\begin{cases} \boldsymbol{\theta} \sim \textit{Diri} \text{ (Prior)} \\ \boldsymbol{a}|\boldsymbol{\theta} \sim \textit{Mn} \text{ (Samp.)} \\ \boldsymbol{a'}|\boldsymbol{\theta}, \boldsymbol{a} \sim \textit{Mn} \text{ (Samp.)} \end{cases} \longrightarrow \begin{cases} \boldsymbol{a} \sim \textit{DiMn} \\ + \\ \boldsymbol{a'}|\boldsymbol{a} \sim \textit{DiMn} \end{cases}$$

□ M.-Hypergeometric + DiMn model

$$\begin{cases} \boldsymbol{A} \sim \textit{DiMn} \text{ (Prior)} \\ \boldsymbol{a}|\boldsymbol{A} \sim \textit{MHyp} \text{ (Samp.)} \\ \boldsymbol{a'}|\boldsymbol{A}, \boldsymbol{a} \sim \textit{MHyp} \text{ (Samp.)} \end{cases} \longrightarrow \begin{cases} \boldsymbol{a} \sim \textit{DiMn} \\ + \\ \boldsymbol{a'}|\boldsymbol{a} \sim \textit{DiMn} \end{cases}$$

□ Exchangeability + DiMn model

$$\begin{cases} \boldsymbol{a}^* \sim \textit{DiMn} \text{ (Prior)} \\ \boldsymbol{a}|\boldsymbol{a}^* \sim \textit{MHyp} \text{ (Samp.)} \\ \boldsymbol{a'}|\boldsymbol{a}^*, \boldsymbol{a} \sim \textit{MHyp} \text{ (Samp.)} \end{cases} \longrightarrow \begin{cases} \boldsymbol{a} \sim \textit{DiMn} \\ + \\ \boldsymbol{a'}|\boldsymbol{a} \sim \textit{DiMn} \end{cases}$$

Bayesian answers to inference (1)
Parametric problems

- **Prior uncertainty**: $P(\boldsymbol{\theta})$

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

For drawing all inferences, from observed data to unknown parameters

- **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: $h(\lambda|\boldsymbol{a})$
  - Expectation, variance: $E(\lambda|\boldsymbol{a})$, $Var(\lambda|\boldsymbol{a})$
  - 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

Bayesian answers to inference (2)
Predictive problems

- **Prior uncertainty**: $P(\boldsymbol{a})$ or $P(\boldsymbol{f})$

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

For drawing all inferences, from observed data to future data

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

IMPORTANT DISTRIBUTIONS

Relevant distributions

- **Parametric inference** on infinite population
  - Dirichlet (*Diri*), any K
  - Beta (*Beta*), $K = 2$
- **Predictive inference** on future n' data
  - Dirichlet-Multinomial (*DiMn*), any K
  - Beta-Binomial (*BeBi*), $K = 2$
- **Links**

	n'	$n' \to \infty$
$K = 2$	*BeBi*	*Beta*
K	*DiMn*	*Diri*

Beta distribution

□ **Consider** the variable

$$\theta \in [0,1]$$

and the hyper-parameters

$$\alpha_1 > 0,\ \alpha_2 > 0$$

or $s = \alpha_1 + \alpha_2,\ t_1 = \alpha_1/s,\ t_2 = \alpha_2/s$,
with $s > 0,\ t_1 > 0,\ t_2 > 0,\ t_1 + t_2 = 1$

□ **Beta density**

$$\theta \sim Beta(\alpha_1, \alpha_2) = Beta(st_1, st_2)$$

$$\begin{aligned} h(\theta) &= \frac{\Gamma(s)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\,\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \\ &\propto \theta_1^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \end{aligned}$$

□ **Expectation and variance**

$$\begin{aligned} E(\theta) &= \alpha_1/s = t_1 \\ Var(\theta) &= \frac{\alpha_1\alpha_2}{s^2(s+1)} = \frac{t_1 t_2}{s+1} \end{aligned}$$

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

Dirichlet density

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

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

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

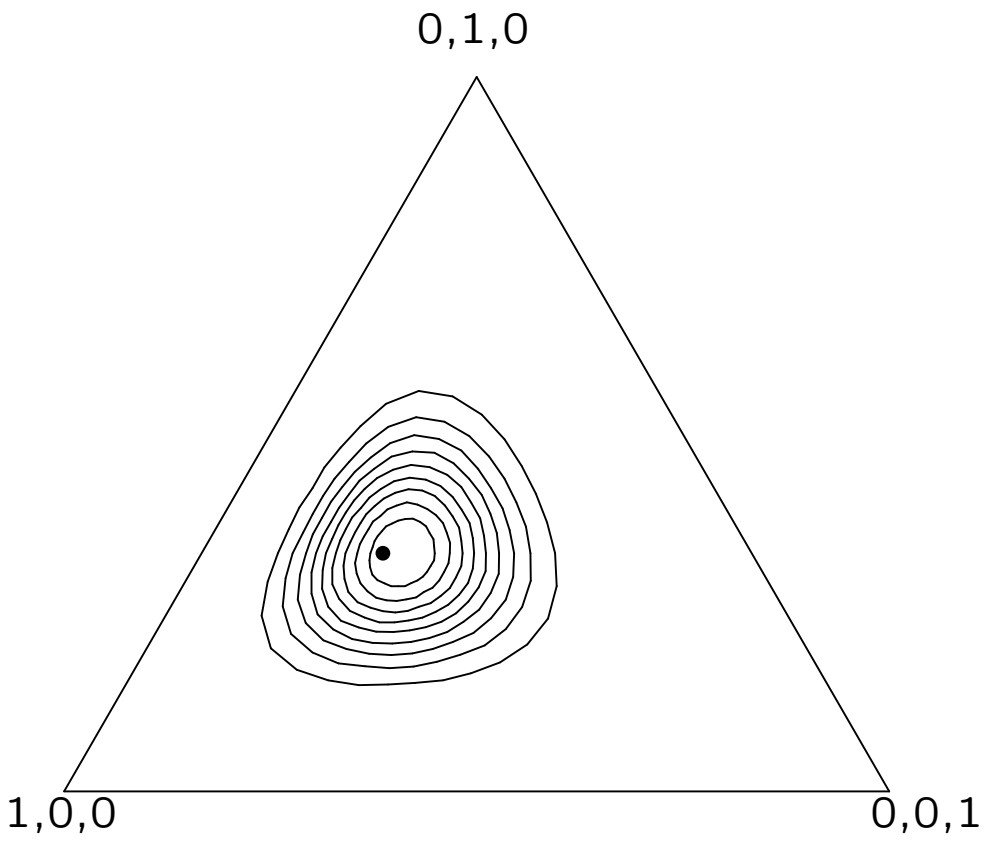
Examples of Dirichlet's

Example 1

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

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, $\theta_{234} = \theta_2 + \theta_3 + \theta_4$, and strengths, $\alpha_{234} = \alpha_2 + \alpha_3 + \alpha_4$.

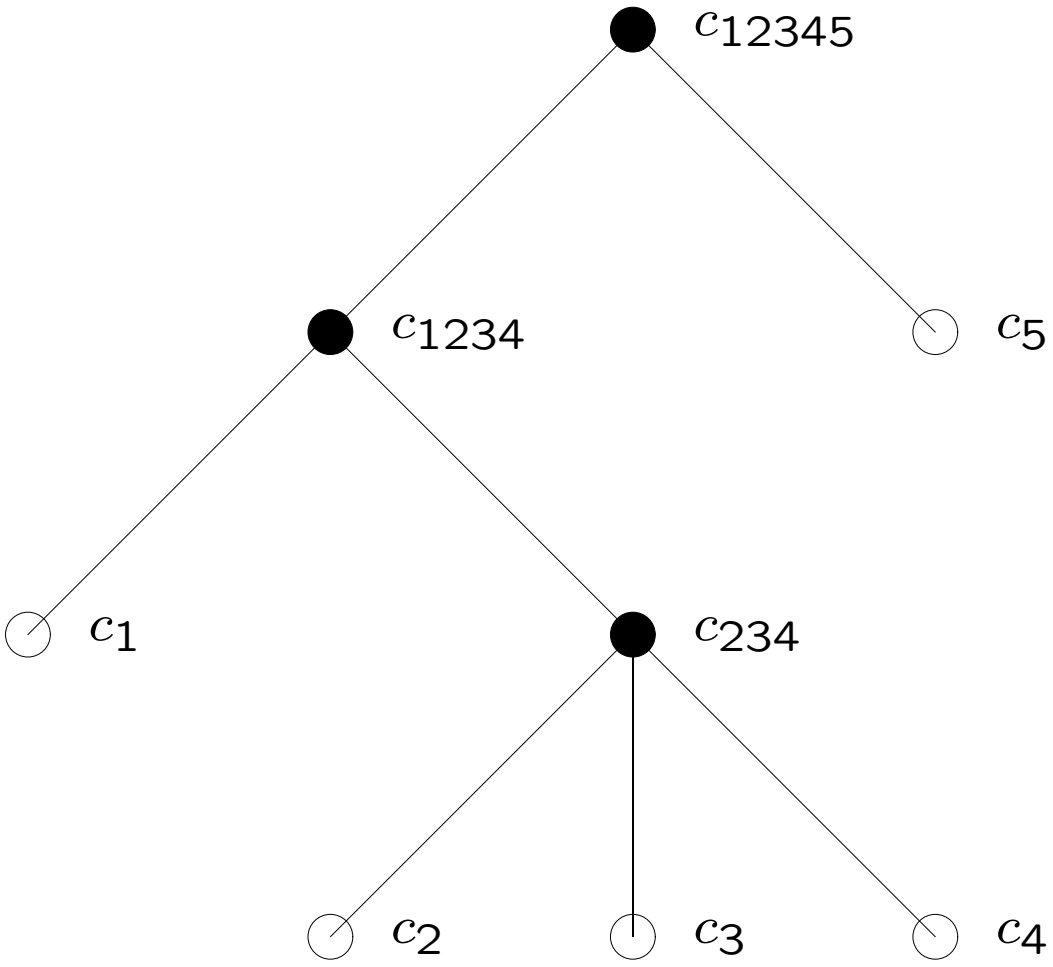
- **Restriction property**

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

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

- **Generalizes** to any tree underlying the set C.

Tree representation of categories



Beta-Binomial distribution (1)

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 $\alpha_1 > 0$ and $\alpha_2 > 0$, with $\alpha_1 + \alpha_2 = s$

Probability distribution function

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

Beta-Binomial distribution (2)

- **Expectation & variance** of a_1 and $f_1 = a_1/n$

$$
\begin{aligned}
E(a_1) &= n\frac{\alpha_1}{s} = nt_1 \\
E(f_1) &= t_1 \\
Var(f_1) &= \frac{t_1(1-t_1)}{s+1} \times \frac{n+s}{n}
\end{aligned}
$$

where $t_1 = \alpha_1/s$, $1 - t_1 = t_2 = \alpha_2/s$

- **Convergence** of distribution of f_1

$$t_1 \rightarrow Beta(\alpha_1, \alpha_2)$$

when $n \rightarrow \infty$

Dirichlet-Multinomial distribution

Notation

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

for $\boldsymbol{a} = (a_1, \ldots, a_K)$, a_k positive ints, $\sum_k a_k = n$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\alpha_k > 0$, $\sum_k \alpha_k = s$

Probability distribution function

$$\begin{aligned} P(\boldsymbol{a}) &= \frac{\prod_k \binom{a_k+\alpha_k-1}{a_k}}{\binom{n+s-1}{n}} \\ &= \frac{n!\Gamma(s)}{\Gamma(n+s)} \prod_k \frac{\Gamma(a_k+\alpha_k)}{a_k!\Gamma(\alpha_k)} \\ &= \binom{n}{\boldsymbol{a}} \frac{\prod_k \alpha_k^{[a_k]}}{s^{[n]}} \end{aligned}$$

Mathematical functions or coefficients

Binomial coefficient

$$\binom{n}{a} = \frac{n!}{a!(n-a)!}$$

for n, a integers, $n \geq a$

Multinomial coefficients

$$\binom{n}{\boldsymbol{a}} = \frac{n!}{a_1! \cdots a_k!}$$

for $\boldsymbol{a} = (a_1, \ldots, a_K)$ integers, $\sum_k a_k = n$

Generalized binomial coefficients

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

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

Ascending factorial (from Appell ?)

$$x^{[m]} = x(x+1)\cdots(x+m-1), \quad x^{[0]} = 1$$

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

OBJECTIVE BAYESIAN MODELS

Objective Bayesian models

☐ Priors proposed for objective inference

Idea: $\boldsymbol{\alpha}$ expressing prior ignorance about $\boldsymbol{\theta}$ or $\boldsymbol{a}^*$ (Kass & Wasserman, 1996; Bernard, 1996)

☐ For direct (Mn or MHyp) 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, uniform: $\alpha_k = 1$ $(s = K)$
- Berger-Bernardo reference priors

☐ For negative (Mn or MHyp) 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 ($P(\boldsymbol{\theta})$ or $P(\boldsymbol{a}^*)$) chosen independently of sampling model ($P(\boldsymbol{a}|\boldsymbol{\theta})$ or $P(\boldsymbol{a}|\boldsymbol{a}^*)$). 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

□ 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
- 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 $a_k = n = 1$, essentially says that $\theta_k = 1$, or that $a'_k = n'$, with probability 1.
  If $a_k = 0$, essentially says that $\theta_k = 0$, or that $a'_k = 0$ for any n', with probability 1.
- Doesn't give a reasonable account of uncertainty.

☐ Limit case of the ID(M)M

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 $P(\boldsymbol{\theta}|\boldsymbol{a})$

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 = Gamble
- Expectation = Prevision

☐ **Previsions of gambles are more fundamental than probabilities of events**

- Precise world:

$$\text{Previsions} \Longleftrightarrow \text{Probabilities}$$

- Imprecise world:

$$\text{Previsions} \Longrightarrow \text{Probabilities}$$

☐ **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
- This is assured by generalized Bayes' rule, which reduces to Bayes' rule/theorem in the case of linear previsions.

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.

Towards the IDM & the IDMM

□ Building an Imprecise Dirichlet model

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

□ . . . or an Imprecise Dirichlet-multinomial model

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

The IDM & IDMM

Class of priors for the IDM & the IDMM

□ **Models proposed** by Walley (1996) for the IDM, and by Walley, Bernard (1999) for the IDMM.

□ **Which prior class?**

Chosing a *Diri* or a *DiMn* prior amounts to chosing prior strengths

$$\begin{aligned} \boldsymbol{\alpha} &= (\alpha_1, \ldots, \alpha_K) \\ &= s\,\boldsymbol{t} \\ &= s\,(t_1, \ldots, t_K) \end{aligned}$$

In the IDM or the IDMM

- Fix the total prior strength s
- Let $\boldsymbol{t}$ take all possible values in $\mathcal{T} = \mathcal{S}^\star(1, K)$

□ **Yielding which properties?**

- Nice properties for modeling prior ignorance
- Satisfy several desirable principles

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 $s > 0$:

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

Posterior IDM

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

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

Updating

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

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

Prior and posterior IDMM

Prior IDMM

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

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

Posterior IDMM

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

$$\mathcal{M}_n \;=\; \{DiMn(n'; \boldsymbol{a} + s\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{T} = \mathcal{S}^\star(1, K)\}.$$

Updating

Similarly, each *DiMn* distribution on $\boldsymbol{a}^*$ in the set $\mathcal{M}_0$ is updated into another *DiMn* on $\boldsymbol{a'}|\boldsymbol{a}$ in the set $\mathcal{M}_n$.

Counts / frequencies

Prior on $\boldsymbol{a}^*$ or $\boldsymbol{f}^*$, posterior on $\boldsymbol{a'}|\boldsymbol{a}$ or $\boldsymbol{f'}|\boldsymbol{a}$.

Drawing inferences from the IDM or IDMM

□ Events, indicator functions

- Compute lower & upper (L&U) probabilities of events of interest
- Substantial conclusion if lower probability is sufficiently large

□ Random quantities

- Compute L&U cumulative distribution functions (cdf)
- Compute L&U expectations
- Compute L&U variances
- Compute L&U credible limits
- Compute (conservative) credible interval having a fixed (*e.g.* 0.95) lower probability

□ Optimization problems:

minimizing and maximizing

L&U probabilities of an event

Prior L&U probabilities

Consider an event B relative to $\boldsymbol{f'}$, and $P_{\boldsymbol{st}}(B)$ the prior probability obtained from the distribution $DiMn(n'; s\boldsymbol{t})$ 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 $P_{\boldsymbol{st}}(B)$ *w.r.t.* $\boldsymbol{t} \in \mathcal{S}^\star(1, K)$.

Posterior L&U probabilities

Denote $P_{\boldsymbol{st}}(B|\boldsymbol{a})$ the posterior probability of B obtained from the prior $DiMn(n'; s\boldsymbol{t})$ in $\mathcal{M}_0$, *i.e.* the posterior $DiMn(n'; \boldsymbol{a} + s\boldsymbol{t})$ 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 $P_{\boldsymbol{st}}(B|\boldsymbol{a})$ *w.r.t.* $\boldsymbol{t} \in \mathcal{S}^\star(1, K)$.

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

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

$$\lambda = g(\boldsymbol{f'}) \;=\; \begin{cases} f'_k \\ \sum_k y_k f'_k \\ f'_i / f'_j \\ \textit{etc.} \end{cases}$$

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.* $\boldsymbol{t} \in \mathcal{S}^\star(1, K)$,

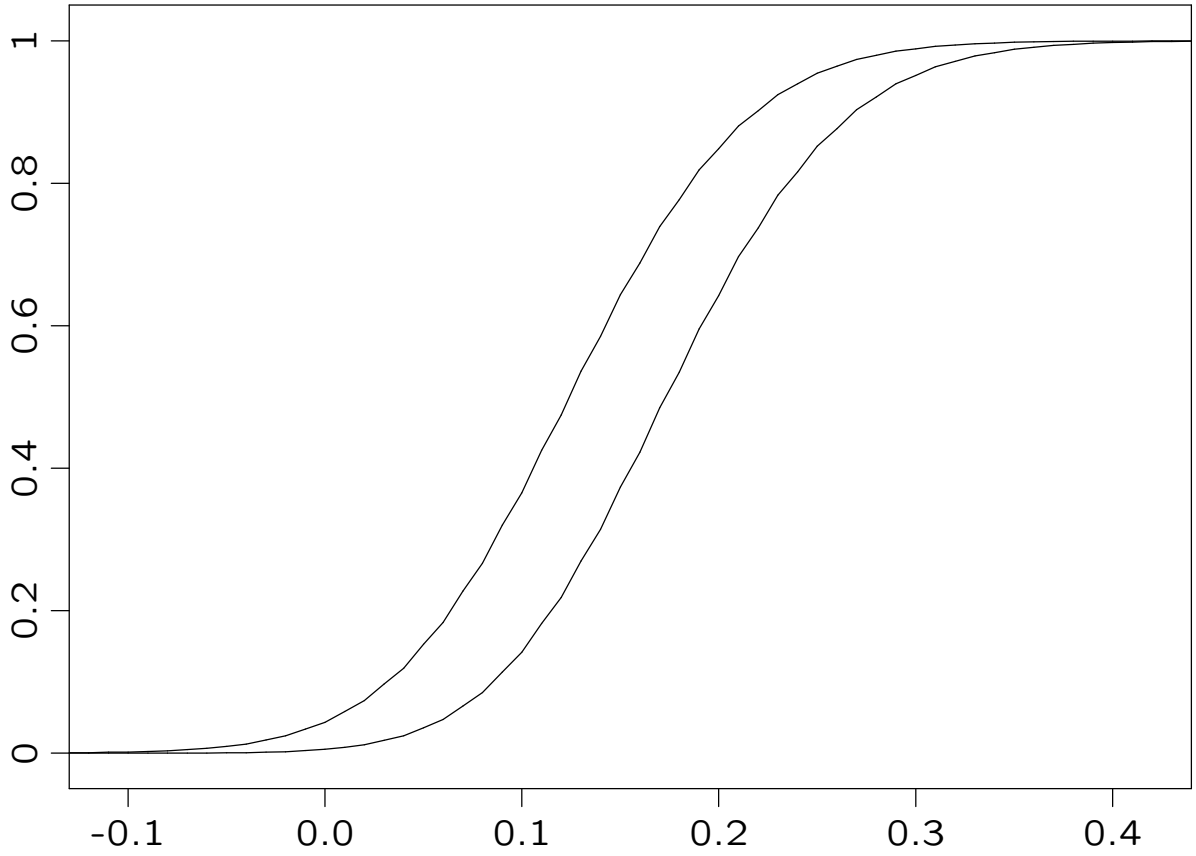
Example of L&U cdf's

Example from Walley, Bernard (1999)

Data $\boldsymbol{a} = (2, 12, 46, 6, 0)$ with $n = 66$ and $K = 5$.
Prediction for $n' = 384$ (*i.e.* $n^* = 450$), on

$$\begin{aligned} \lambda = g(\boldsymbol{f}^*) &= 2f_1^* + f_2^* - f_4^* - 2f_5^* \\ &= \frac{384}{450}\, g(\boldsymbol{f'}) + \frac{66}{450}\, g(\boldsymbol{f}) \end{aligned}$$

L&U cdf's of λ



Optimization problems

☐ Set or convex combinations?

The prior & posterior sets, $\mathcal{M}_0$ and $\mathcal{M}_n$, of *Diri* or *DiMn* distributions, are used to define lower previsions $\underline{P}(\cdot)$ (by taking lower envelopes). Each $\underline{P}(\cdot)$ is equivalent to the class of its dominating linear previsions, which contains also all convex combinations of these *Diri* or *DiMn* distributions.

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

☐ 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

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 ID(M)M's (1)

Ignorance vs. Near-ignorance

- Ignorance in the IP theory: vacuous probabilistic statements
- Complete ignorance: ignorance about all gambles and events
- Near-ignorance: ignorance about some gambles and/or events

Two extremes

- $s \to 0$: Haldane's model, precise
- $s \to \infty$: vacuous model, maximally imprecise

Haldane's model: $s \to 0$

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

Extreme ID(M)M's (2)

☐ Vacuous model: $s \to \infty$

- The IDM(s_{sup}) contains all IDM's with $s \leq s_{sup}$, *i.e.* all $Diri_{s\boldsymbol{t}}$, $s \leq s_{sup}$, $\boldsymbol{t} \in \mathcal{T}$. 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

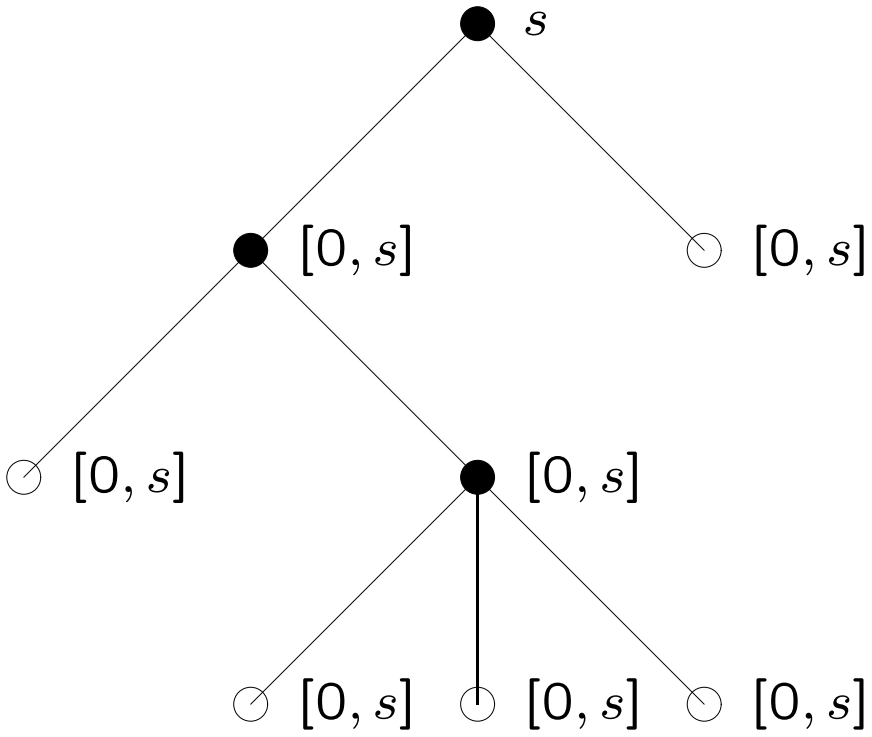
- If too high, inferences are too weak

Hyperparameter s must be large enough to

- Encompass objective Bayesian inferences:
  Haldane: $s > 0$; Perks: $s \geq 1$
  Other solutions? Problem: $s \geq K/2$ or $\geq K$
- Encompass frequentist inferences

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

Why does the ID(M)M satisfy the EP and RIP?



- *Diri* or *DiMn* distributions compatible with any tree. But, under a PDM or PDMM, total prior strength s scatters when moving down the tree
- In the IDM or IDMM, all allocations of s to the nodes are possible (due to imprecision)
- Each sub-tree inheritates the same s

PREDICTIVE INFERENCE FROM THE ID(M)M

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

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

$$\begin{aligned} P(\boldsymbol{a}) &= \prod_k \binom{a_k + \alpha_k - 1}{a_k} \Big/ \binom{n + s - 1}{n} \\ &= \binom{n}{\boldsymbol{a}} \frac{{\alpha_1}^{[a_1]} \cdots {\alpha_K}^{[a_K]}}{s^{[n]}} \end{aligned}$$

Dirichlet-multinomial posterior

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

$$\begin{aligned} P(\boldsymbol{a}'|\boldsymbol{a}) &= \prod_k \binom{a'_k + a_k + \alpha_k - 1}{a'_k} \Big/ \binom{n' + n + s - 1}{n'} \\ &= \binom{n'}{\boldsymbol{a}'} \frac{(a_1 + \alpha_1)^{[a'_1]} \cdots (a_K + \alpha_K)^{[a'_K]}}{(n + s)^{[n']}} \end{aligned}$$

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

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

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

$$\begin{aligned} P(a_k) &= \frac{\binom{a_k+\alpha_k-1}{a_k}\binom{n-a_k+s-\alpha_k-1}{n-a_k}}{\binom{n+s-1}{n}} \\ &= \binom{n}{a_k}\frac{{\alpha_k}^{[a_k]}(s-\alpha_k)^{[n-a_k]}}{s^{[n]}} \end{aligned}$$

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

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

$$\begin{aligned} P(a'_k|\boldsymbol{a}) &= \frac{\binom{a'_k+a_k+\alpha_k-1}{a'_k}\binom{n'-a'_k+n-a_k+s-\alpha_k-1}{n'-a'_k}}{\binom{n'+n+s-1}{n'}} \\ &= \binom{n'}{a'_k}\frac{(a_k+\alpha_k)^{[a'_k]}(n-a_k+s-\alpha_k)^{[n'-a'_k]}}{(n+s)^{[n']}} \end{aligned}$$

Prior predictive distribution under the IDMM

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

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

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

□ **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 $t_k \longrightarrow 0$ and $t_k \longrightarrow 1$ 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 $t_k \longrightarrow 1$ and $t_k \longrightarrow 0$ respectively

Posterior predictive distribution under the IDMM (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:

$$\mathcal{M}_n \;=\; \{DiMn(n';\, \boldsymbol{a} + s\boldsymbol{t}) : \boldsymbol{t} \in \mathcal{S}^{\star}(1, K)\}$$

□ **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 IDMM (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

□ **L&U posterior exp. & cdf's** are obtained using either

$$BeBi(n';\ a_k, n - a_k + s)$$
$$\text{or}\quad BeBi(n';\ a_k + s, n - a_k)$$

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 all absolute strengths obtained by summation.
- Recursively, for any pooling in $J < K$ categories, the *DiMn* form and the value of s are both 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(s)
- Apply IDMM(s) first, then pool

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 a & f

- Near-ignorance properties: L&U exp. $E(a_k)$, $E(f_k)$ and cdf's $F_{a_k}(.)$, $F_{f_k}(.)$ 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

Frequentist prediction

□ **"Bayesian and confidence limits for predictions"** (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 IDMM?
(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)

THE RULE OF SUCCESSION

Rule of succession problem

Problem $P(a'|a)$ for $n' = 1$

- Prediction about the next observation
- Also called 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 PDMM

Bayesian rule of succession

The rule of succession obtained from a PDMM, with hyper-parameters $\boldsymbol{\alpha} = s\boldsymbol{t}$, 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

Denoting $f'_j = \sum_{k \in j} f'_k$, the future frequencies in n' data, and possibly $\theta_j = \sum_{k \in j} \theta_k$, the population frequencies, then

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

Prior rule of succession under the IDMM

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 IDMM

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 *DiMn* distribution, *DiMn*$(n'; \boldsymbol{\alpha})$ with $n' = 1$ and $\alpha_k = s/K$, is

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

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	$(a_j + J)/(n + K)$

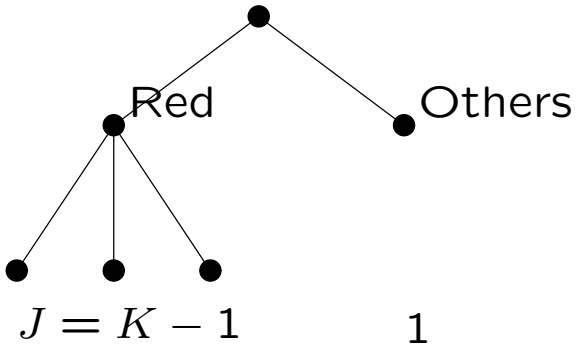
Dependence on K and J except Haldane

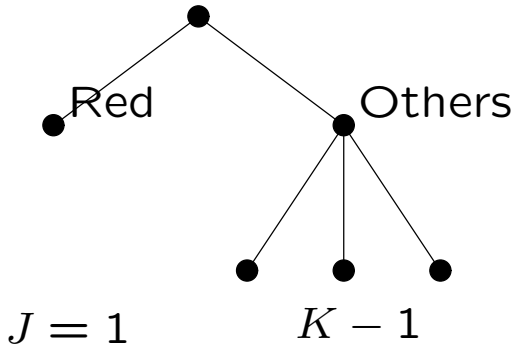
Particular case $J = 1$, $K = 2$

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** (reminder)

- 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 $s = 1$.

CONCLUSIONS

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 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

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

□ **About Dirichlet's**

(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.

□ **Same arguments hold** for *DiMn* distributions

Links between IDM and IDMM

Parametric and predictive inference

In general, in both precise Bayesian models and in the ID(M)M,

- $\boldsymbol{\theta}$, $\boldsymbol{\theta}|\boldsymbol{a}$ yield $\boldsymbol{f}$, $\boldsymbol{f'}|\boldsymbol{a}$ (from Bayes' theorem)
- $\boldsymbol{f}$, $\boldsymbol{f'}|\boldsymbol{a}$ yield $\boldsymbol{\theta}$, $\boldsymbol{\theta}|\boldsymbol{a}$ (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

Fundamental properties of the ID(M)M

Principles

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

ID(M)M *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.

REFERENCES

Berger, J. O., Bernardo, J. M. (1992), "Ordered Group Reference Priors with Application to the Multinomial Problem", *Biometrika*, 79, no. 1, pp. 2–5–37.

Bernard, J.-M. (1996), "Bayesian Interpretation of Frequentist Procedures for a Bernoulli Process", *The American Statistician*, 50, no. 1, 7–13.

Bernard, J.-M. (1997), "Bayesian Analysis of Tree-Structured Categorized Data", *Revue Internationale de Systémique*, Vol. 11 no. 1, pp. 1–1–29.

Fang, K. T., Kotz, S., Ng, K. W. (1990), *Symetric Multivariate and Related Distributions*, New-York: Chapman and Hall.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Monographs on Statistics and Applied Probability 55, New-York: Chapman & Hall.

Haldane, J. B. S. (1948), "The Precision of Observed Values of Small Frequencies," *Biometrika*, 35, pp. 2–97–300.

Hutter, M. (2003). Robust Estimators under the Imprecise Dirichlet Model. In Bernard J.-M., Seidenfeld T., Zaffalon M. (Eds), *ISIPTA'03: Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano, Switzerland*, Proceedings in Informatics 18, Waterloo, Ontario, Canada: Carleton Scientific, pp. 274–289.

Jeffreys, H. (1946), "An Invariant Form for the Prior Probability in Estimation Problems", *Proceedings of the Royal Society of London*, Ser. A, 186, pp. 4–53–461.

Jeffreys, H. (1961), *Theory of Probability*, 3rd ed., Oxford: Clarendon Press.

Kass, R. E., Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules", *Journal of the American Statistical Association*, Vol. 91, no. 435, pp. 1–343–1370.

Perks, F. J. A. (1947), "Some Observations on Inverse Probability Including a New Indifference Rule (with discussion)," *Journal of the Institute of Actuaries*, 73, pp. 2–85–334.

Thatcher, A. R. (1964), "Relationships Between Bayesian and Confidence Limits for Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, pp. 1–76–210.

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, Monographs on Statistics and Applied Probability 42, London: Chapman & Hall.

Walley, P. (1996), "Inferences from Multinomial Data: Learning about a Bag of Marbles", *Journal of the Royal Statistical Society*, Ser. B., 58 no. 1, pp. 3–--57.

Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. Journal of Statistical Planning and Inference, Vol. 105, no. 1, pp. 35–65.

Walley, P., Bernard, J.-M. (1999), "Imprecise Probabilistic Prediction for Categorical Data", Technical Report CAF-9901, Laboratoire Cognition et Activités Finalisées, University Paris 8, Saint-Denis, France, January 1999.