Statistics and Imprecise Probabilities

Thomas Augustin

Department of Statistics, LMU Munich

SIPTA Summer School 2022
August 17^{th}, 2022

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

LMU Munich

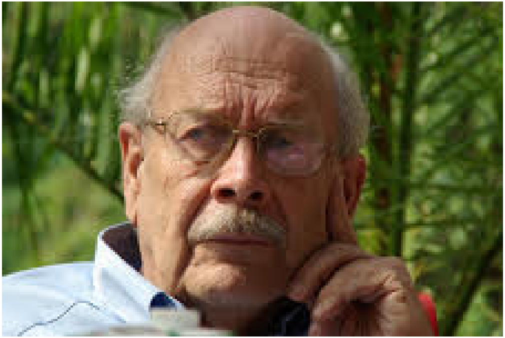
- one of largest German universities
- $\approx$ 50.000 students

- Department of Statistics
- Bachelor (Major and Minor) and Master programme in *Statistics and Data Science*
- PhD study

The Department of Statistics at LMU

- founded 1973/74 as Department for Statistics and Philosophy of Science (Weichselberger, Stegmüller)
- philosophy of Science: predecessor institute for Munich Center for Mathematical Philosophy (Hartmann, Leitgeb, List)
- major research focus of the department of statistics changing over time
  - foundations of statistics (Ferschl, Schneeweiß, Weichselberger)
  - advanced statistical regression modelling (CRC, Fahrmeir, Tutz)
  - statistical machine learning and data science (Munich Center for Machine Learning (MCML), Bischl, Kreuter)

Kurt Weichselberger (1929-2016)

 [1]

See also Augustin & Seising (2018, IJAR)

[1] Photo kindly provided by Weichselberger's family

Foundations of Statistics and Their Applications

https://www.foundstat.statistik.uni-muenchen.de/index.html

[Aug 16th, 2022]

- Thomas Augustin
- Hannah Blocher
- Dominik Kreiß
- Christoph Jansen
- Gilbert Kiprotich
- Malte Nalenz
- Julian Rodemann
- Georg Schollmeyer

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Some Selected Further Reading I: Classical Work

- C. Manski (2003): *Partial identification of probability distributions.* Springer, New York.
- D. Ríos Insua and F. Ruggeri (eds.) (2000): *Robust Bayesian Analysis.* Springer, Berlin.
- P. Walley (1991): *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall, London.
- P. Walley (1996): Inferences from multinomial data: Learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society, Series B, 58:3–34.*
- K. Weichselberger (2001): *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept.* Physica, Heidelberg.[2]
- Biannual ISIPTA Proceedings: `www.sipta.org[Aug16th,2022]`

[2]in German; Elementary Foundations of a more General Calculus of Probability I: Interval Probability as a Comprehensive Concept.

Some Selected Further Reading II: Review Papers I

- T. Augustin (2022): Statistics with imprecise probabilities: a short survey. In: L. Aslett, F. Coolen, J. De Bock (eds.) *Uncertainty in Engineering: Introduction to Methods ans Applications*. Springer, Cham, pp. 67-79.
- T. Augustin, G. Walter, F. Coolen, (2014): Statistical inference. In: T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes (eds.). *Introduction to Imprecise Probabilities*. Wiley, Chichester, pp. 135–188.
- S. Bradley. Imprecise probabilities (2019): In Edward N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). Standford University.[3]

[3] https://plato.stanford.edu/entries/imprecise-probabilities/[Aug16th,2022]

Some Selected Further Reading III: Review Papers II

- F. Molinari (2020): Microeconometrics with partial identification. In: S. Durlauf, L. Hansen, J. Heckman and R. Matzkin (eds.) *Handbook of Econometrics, Vol. 7A*, pp. 355–486.
- B. Ristic, C. Gilliam, M. Byrne and A. Benavoli (2020): A tutorial on uncertainty modeling for machine reasoning. *Information Fusion* 55:30–44.

For the statistical background, see, for instance,

- B. Efron and T. Hastie (2016): *Computer Age Statistical Inference.* Cambridge UP.

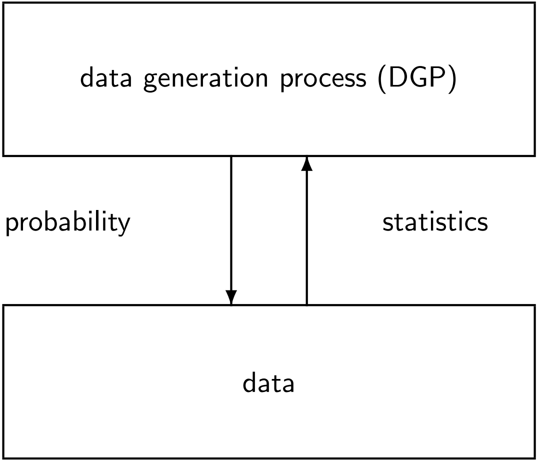
Table of contents

1. Introduction and Background
2. **Statistics in a Nutshell**
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Statistics

- inference, reasoning, learning, modelling
- here not: data production (mainly official statistics)

Statistics as Inverted Probability



Sample, Statistical Model

- **sample**: random vector/matrix $X = (X_1, \dots, X_n)$ on some space $\mathcal{X}$
- sample size n
- joint probability measure $p(\cdot)$ as a model for the data generation process DGP
- capital letter X: random, describing potential observation; small letter x: fixed value, standing for realization, concrete observation

Sample, Statistical Model

- **sample**: random vector/matrix $X = (X_1, \dots, X_n)$ on some space $\mathcal{X}$
- sample size n
- joint probability measure $p(\cdot)$ as a model for the data generation process DGP
- capital letter X: random, describing potential observation; small letter x: fixed value, standing for realization, concrete observation
- parametric modelling: $p(\cdot)$ is known up to some aspects $\longrightarrow$ **parameter** ϑ (low dimensional, ("natural parametrization")) with values in some **parameter space** Θ
- Thus inference on $p(\cdot)$ is described as inference on ϑ
- Basic ingredients of a **statistical model**: $\mathcal{X}$ and $(p_\vartheta(\cdot))_{\vartheta \in \Theta}$

Sample, Statistical Model

- **sample**: random vector/matrix $X = (X_1, \ldots, X_n)$ on some space $\mathcal{X}$
- sample size n
- joint probability measure $p(\cdot)$ as a model for the data generation process DGP
- capital letter X: random, describing potential observation; small letter x: fixed value, standing for realization, concrete observation
- parametric modelling: $p(\cdot)$ is known up to some aspects $\longrightarrow$ **parameter** ϑ (low dimensional, ("natural parametrization")) with values in some **parameter space** Θ
- Thus inference on $p(\cdot)$ is described as inference on ϑ
- Basic ingredients of a **statistical model**: $\mathcal{X}$ and $(p_\vartheta(\cdot))_{\vartheta\in\Theta}$

  $p_\vartheta(\cdot)$ has density/probability mass function

$$f(x\|\vartheta)$$

Typical Situations

This comprises most of the model classes considered in statistics, where $X_1, \ldots, X_i, \ldots, X_n$ are describing

- independently and identically distributed **repetitions**
- independently and identically distributed repetitions split in pairs $X_i = (Y_i^T, Z_i^T)^T$ where $p_\vartheta(\cdot)$ is constructed from the modelled conditional distributions of Y_i given Z_i: **regression models** with covariates Z_i and dependent variable Y_i
- longitudinally dependent observations: **panel study, time series, stochastic process in discrete time** i

Inference Tasks

- **testing hypotheses** on ϑ: decide between potentially underlying DGPs
- **estimation** of ϑ: give a (vector of) values for a (multivariate) characteristic of the underlying DGP
- **interval estimation**: give range with some guaranteed coverage
- **decision making** with data coming from the underlying DGP
- **predictive**: characterize underlying distribution by making statements on the properties of observations not yet seen

Inference Paradigms

See, e.g., Barnett (1999[3], Wiley), Efron & Hastie (2016, Cambridge UP) for textbooks, and `http://bff-stat.org/` for recent developments

Inference Paradigms

- frequentist

Inference Paradigms

- frequentist
- likelihood

Inference Paradigms

- frequentist
- likelihood
- Bayesian

Inference Paradigms

- frequentist
- likelihood
- Bayesian
- fiducial inference, also called Fisherian inference

Frequentist Inference

- search for a procedure that behaves well under infinitely many virtual repetitions of the underlying "experiment":
- unknown, but fixed true parameter values

Likelihood Inference

- After having seen the data, reinterpret $f(x||\vartheta)$ as a function in ϑ.
- It expresses the likelihood/plausibility that x has been produced by the model with ϑ as the truly underlying parameter

Bayesian Inference

- subjective probability: express YOUR uncertainty by a probability
- assign a probability on the parameter: **prior distribution** (density/probability mass function $\pi(\cdot)$)
- update the prior in the light of the sample by Bayes rule: **posterior distribution** (density/probability mass function $\pi(\cdot|x)$

$$\pi(\vartheta|x) = f(x||\vartheta) \cdot \pi(\vartheta)$$

Bayesian Inference

- subjective probability: express YOUR uncertainty by a probability
- assign a probability on the parameter: **prior distribution** (density/probability mass function $\pi(\cdot)$)
- update the prior in the light of the sample by Bayes rule: **posterior distribution** (density/probability mass function $\pi(\cdot|x)$

$$\pi(\vartheta|x) = f(x||\vartheta) \cdot \pi(\vartheta)$$

Prior knowledge: $\pi(\vartheta)$
sampling distribution $f(x|\vartheta)$
$+$observation x

$\Longrightarrow$ current knowledge $\pi(\vartheta|x)$: posterior

Fiducial Inference, also called Fisherian Inference

- "posteriors without priors", relation to logical probability
- "[. . .] an attempt to eat the Bayesian omelette without breaking the Bayesian eggs" (Savage 1961, Proc 4th Berkeley)
- "Fiducial inference stands as R. A. Fisher's one great failure." (Zabell, 1992, StatSc, p. 369)
- intensive discussion inspiring quite productive rescue attempts, including Dempster (1967, AnnMathStat), Seidenfeld (1979, Reidel), Hampel (2006, Ahlswede et al.), Weichselberger (2009, ISIPTA Tut), Martin & Liu (2015, Chapman & Hall).

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

First Inquires on the Classical Approaches

- Are infinite repetitions stable over time?

- How do we get the concrete form of the probabilities involved?
- Do small differences in the modelling matter?
- Can “wrong choices” be detected? If so what to do?

Is it a Good Idea to Bring in Subjective Information into Statistical Inference?

?

General Aspects and some Caveats of Bayesian Inference

- For $n \longrightarrow \infty$ full weight on the sample, irrespective of prior: "asymptotic objectivity". Asymptotically, the posterior concentrates around the true parameter value.
- For finite (not very large n), the parameters of the prior have to be specified by the researcher, and this choice substantially influences the result.

General Aspects and some Caveats of Bayesian Inference (Continued)

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

General Aspects and some Caveats of Bayesian Inference (Continued)

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

  "Bayesian methods are increasingly used in proof-of-concept studies. An important benefit of these methods is the potential to use informative priors, thereby reducing sample size. This is particularly relevant for treatment arms where there is a substantial amount of historical information such as placebo and active comparators." (Mutsvar, Tytgat & [Ros.] Walley, 2016, Pharmaceut-Statist, p. 28)

General Aspects and some Caveats of Bayesian Inference (Continued)

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

- But is the knowledge truly precise enough?
- How to express ignorance?
- How to express valuable *partial* knowledge?

General Aspects and some Caveats of Bayesian Inference (Continued)

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

- But is the knowledge truly precise enough?
- How to express ignorance?
- How to express valuable *partial* knowledge?

- What to do under prior-data conflict? What would one hope for?

Excursus: Uniform Priors as Noninformative Priors I?

Consider a parameter $\vartheta \in [0;1]$, for instance the success probability in i.i.d. Bernoulli trials.

Excursus: Uniform Priors as Noninformative Priors I?

Consider a parameter $\vartheta \in [0;1]$, for instance the success probability in i.i.d. Bernoulli trials.
If you do not have knowledge on ϑ, you also do not have knowledge on ϑ^2.

Excursus: Uniform Priors as Noninformative Priors I?

Consider a parameter $\vartheta \in [0;1]$, for instance the success probability in i.i.d. Bernoulli trials.
If you do not have knowledge on ϑ, you also do not have knowledge on ϑ^2.
If you knew something about ϑ^2, you would know something about $\sqrt{\vartheta^2} = \vartheta$.

Excursus: Uniform Priors as Noninformative Priors I?

Consider a parameter $\vartheta \in [0;1]$, for instance the success probability in i.i.d. Bernoulli trials.
If you do not have knowledge on ϑ, you also do not have knowledge on ϑ^2.
If you knew something about ϑ^2, you would know something about $\sqrt{\vartheta^2} = \vartheta$.
If the prior distribution for θ (random quantity U) and the prior distribution for θ^2 (random quantity U^2) are both uniform, then $\mathbb{E}(U) = \mathbb{E}(U^2) = 0.5$, leading to the contradiction

$$1/12 = \mathbb{V}(U) = \mathbb{E}(U^2) - (\mathbb{E}(U))^2 = 0.5 - 0.5^2 = 0.25$$

Excursus: Uniform Priors as Noninformative Priors I?

Indeed, for the distribution function of $Y = U^2$ with uniformly U one obtains

$$F_Y(y) = P(Y \leq y) = P(U^2 \leq y) \stackrel{4}{=} P(U \leq \sqrt(y)) = [u]_0^{\sqrt(y)} = \sqrt(y).$$

Therefore, the density $f_y(y)$ has the form

$$f_y(y) = \frac{d\,F_Y(y)}{d\,y} = \frac{d\,y^{0.5}}{d\,y} = 0.5y^{-0.5} = 0.5\frac{1}{\sqrt{y}},$$

in particular U^2 is not uniformly distributed.

[4] suppU=[0,1]

Excursus: Uniform Priors as Noninformative Priors I?

Indeed, for the distribution function of $Y = U^2$ with uniformly U one obtains

$$F_Y(y) = P(Y \le y) = P(U^2 \le y) \overset{4}{=} P(U \le \sqrt(y)) = [u]_0^{\sqrt(y)} = \sqrt(y)\,.$$

Therefore, the density $f_y(y)$ has the form

$$f_y(y) = \frac{d\,F_Y(y)}{d\,y} = \frac{d\,y^{0.5}}{d\,y} = 0.5y^{-0.5} = 0.5\frac{1}{\sqrt{y}}\,,$$

in particular U^2 is not uniformly distributed.
Classical solution (?): Jaffray priors

[4] suppU=[0,1]

Prior-data conflict

What to do if prior information and (outlier-free) sample information are conflicting (and the sample is too small to rule out the effect of the prior distribution)?

Prior-data conflict

What to do if prior information and (outlier-free) sample information are conflicting (and the sample is too small to rule out the effect of the prior distribution)?

> *"[...] if we can show that the observed data is surprising in light of the sampling model and the prior, then we must be at least suspicious about the validity of the inferences drawn [...]." (Evans & Moshonov, 2006, BayesianAnal, p. 893)*

Prior-data conflict

What to do if prior information and (outlier-free) sample information are conflicting (and the sample is too small to rule out the effect of the prior distribution)?

> *"[...] if we can show that the observed data is surprising in light of the sampling model and the prior, then we must be at least suspicious about the validity of the inferences drawn [...]." (Evans & Moshonov, 2006, BayesianAnal, p. 893)*

- How to be cautious within a classical probability?

- How to be cautious within the classical probability calculus?

?

- How to be cautious within the classical probability calculus?

?

- Conflicting information goes beyond variability, and thus can not be captured by the variance or other characteristics of precise probabilistic models.

Table of contents

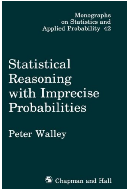
1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

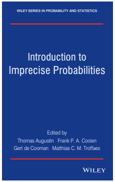
Imprecise Probability for Statistics?!

- Unfortunate misnomer: actually IP claims to provide more precise (better) models

- uncertainty as a multidimensional concept

- In general, often somewhat reserved reactions in the statistical community, although many researchers shaping the theory (Walley, Weichselberger, Seidenfeld, Dempster (and others)) are genuine statisticians





Imprecise Probability for Statistics! Fundamental Concepts

- Here: build simply on a very intuitive understanding
- *sets* of traditional probability models (credal sets) "$\Longleftrightarrow$" interval-valued probability $P(A) = [L(A), U(A)]$ of events A (, or more generally expectations)[5]

[5] $L(\cdot)$ and $U(\cdot)$ are non-additive set-functions, often called *capacities*.

Imprecise Probability for Statistics! Fundamental Concepts

- Here: build simply on a very intuitive understanding
- *sets* of traditional probability models (credal sets) "$\Longleftrightarrow$" interval-valued probability $P(A) = [L(A), U(A)]$ of events A (, or more generally expectations)[5]
- Take the set / the intervals as a basic entity! (No mixing, higher order distributions!)

[5] $L(\cdot)$ and $U(\cdot)$ are non-additive set-functions, often called *capacities*.

Imprecise Probability for Statistics! Fundamental Concepts

- Here: build simply on a very intuitive understanding
- *sets* of traditional probability models (credal sets) "$\Longleftrightarrow$" interval-valued probability $P(A) = [L(A), U(A)]$ of events A (, or more generally expectations)[5]

- Take the set / the intervals as a basic entity! (No mixing, higher order distributions!)

- quality of information: "size of set", width of interval
  - traditional probability as the extreme case of perfect probabilistic information, real number, set with a single element
  - P(A)=[0;1] for all nontrivial events – set of all probability measures: complete ignorance, full ambiguity

[5] $L(\cdot)$ and $U(\cdot)$ are non-additive set-functions, often called *capacities*.

First Ideas I: Directly Based on Precise Probabilitic Models

- different experts (with different precise probabilities)
- assigning probability only to certain events (de Finitti's fundamental theorem, cp. yesterday)
- handling of different granularities: unique extensions from any set-system to IP on the underlying measurable space
- indivisible evidence: high probability for $A \cup B$ can not be split between disjoint events A and B (Ellsberg, medical expert systems, coarsened data)

First Ideas II: Natural Applications

- direct modelling of partial knowledge: intervals of probabilities or expectations
- ordinal probabilities: $p(A) \leq p(B) \leq p(C)$...
- approximately true models $\longrightarrow$ neighborhood models, see below
- unobserved heterogeneity (slightly changing distribution for different individuals due to unobservable individual characteristics (e.g. genetic disposition)

First Ideas III: Interval Ordering

$$P_1(A) \supseteq P_2(A)\,, \quad \text{for all } A$$

- $P_1(\cdot)$ is more cautious than $P_2(\cdot)$.
- learning under homogenous information
- description of conflicting information: intervals get wider
- continuum of uniform distributions: $P(A) = P(B) = P(C) \ldots$
- distinction between negative symmetry (do not know any asymmetry) and positive symmetry (knowledge that symmetry is produced)

- modelling complete ignorance $P(A) = [0,1]$ for all nontrivial events A

Quality of Information

"Let's Be Imprecise in Order to Be Precise (About What We Don't Know)"

Title of Gong & Meng (2021, StatSc (Rejoinder), p. 210)

Ruobin Gong

Xiao-Li Meng[6]

[6] taken from https://ruobingong.github.io and https://statistics.fas.harvard.edu/people/xiao-li-meng [Aug 16th, 2002]

Several Updating / Conditioning Rules

- In general quite an complex issue (see also yesterday, Blackwell)
- standard way to proceed in IP: **generalized Bayes rule**, conditioning element by element (robust Bayes, justified by generalized coherence axioms: Walley (1991, Chapter 6))
- recent discussion in the light of typical statistical settings: Gong & Meng (2021, StatSc), Augustin & Schollmeyer (ibid.)

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Do small differences in models matter at all?

- Naturally, every abstraction yields some kind of imprecision.
- Do small differences in models matter at all?

The mantra of statistical modelling

Box & Draper (1987, Empirical Model Building and Response Surfaces, p. 424)

- "Essentially, all models are wrong,

The mantra of statistical modelling

Box & Draper (1987, Empirical Model Building and Response Surfaces, p. 424)

- "Essentially, all models are wrong,
- but some of them are useful",

Do small differences in models matter at all?

- Naturally, every abstraction yields some kind of imprecision.

- Do small differences in models matter at all?
- Are there probability models with

Model 1 "very similar" Model 2

BUT

Conclusions(Model 1) "quite different" Conclusions(Model 2)?

Assumptions may matter!

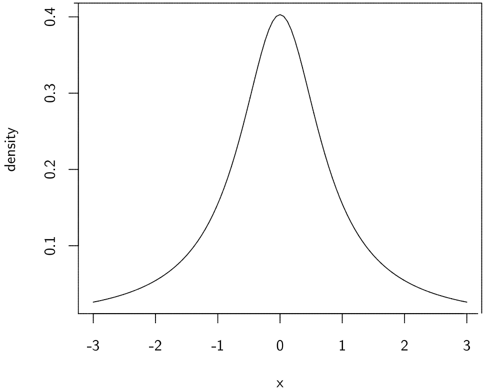


Figure: A “regular, bell-shaped” density

Assumptions may matter!

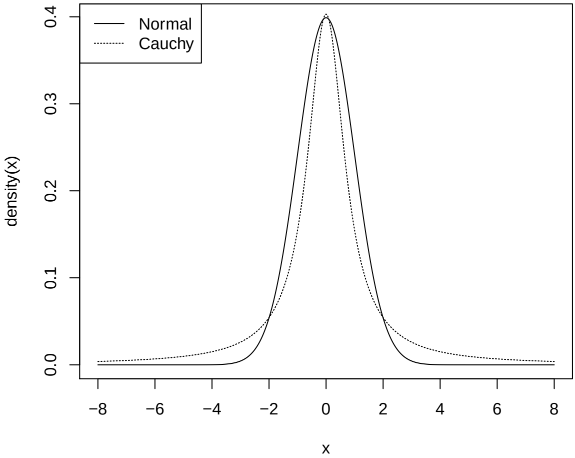


Figure: Densities of the Normal$(0, 1)$ and the Cauchy$(0, 0.79)$ distribution.

Assumptions may matter!

Consider sample mean $\overline{X}$.

- if $X_1, \ldots, X_n \sim N(\mu, 1)$ (normally distributed), then

$$\bar{X} \sim N(\mu, \frac{1}{n})$$

  Learning from the sample, with increasing sample size variance of $\overline{X}$ decreases.

- if $X_1, \ldots, X_n \sim \mathcal{C}(\mu, 1)$ (Cauchy-distributed), then

$$\overline{X} \sim C(\mu, 1)$$

  Distribution does not depend on n, no learning via sample mean possible

Robustness in testing: a motivating example

Consider the simplest testing situation:

- $X_1, \ldots, X_n$ i.i.d. sample, underlying normal distribution $\mathcal{N}(\mu, \sigma_0)$ with σ_0 known and fixed in advance.
- Test the hypotheses

$$H_0 : \mu = 0 \qquad \text{versus} \qquad H_1 : \mu \neq 0$$

at a given level of significance α (Here $\alpha = 0.05$.)

Robustness in testing: a motivating example

- standard test (indeed uniformly most powerful under all unbiased tests respecting the level of significance)
- test statistic

$$T = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\sigma_0}\sqrt{n}$$

- reject H_0 iff

$$\mid T \mid > z_{1-\frac{\alpha}{2}}$$

Robustness in testing: a motivating example

Simple simulation to study this test:

a) Simulate samples of size n from $\mathcal{N}(0, \sigma_0^2)$ (with $\sigma_0^2 = 1$). (Corresponds perfectly to H_0.)

b) Inner loop with say hundred repetitions: Calculate $\mid T \mid$ and count how often H_0 is rejected. Yields counter `C`.

c) Outer loop with say again hundred repetitions: Look at the empirical distribution of `C` and corresponding summary statistics.

What changes if $\mathcal{N}(0, 1)$ is replaced by $\mathcal{C}(0, 0.79)$?

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Insure yourself against non-robustness: neighborhood models

- General issue: many optimal procedures may show very bad behavior under minimal deviations from the ideal model.
- Give up some efficiency in the ideal model for being protected (compare buying an insurance policy).
- formalization via neigborhood models[7]
  instead of $p_\vartheta(\cdot)$ use a model expressing "approximately $p_\vartheta(\cdot)$," i.e. consider the credal set of all distributions "close to $p_\vartheta(\cdot)$"

---

[7]Huber, P.J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. Ann. Statist. 1:251–263
Montes, I., Miranda, E. and Destercke, S. (2020a). Unifying neighbourhood and distortion models: Part I: new results on old models. Int. J. Gen. Syst. 49:602–635.

Neighborhood models via distortion models

- Instead of $p_\vartheta(\cdot)$ use a model expressing "approximately $p_\vartheta(\cdot)$," i.e. consider the credal set of all distributions "close to $p_\vartheta(\cdot)$"
- Formalization via various probability metrics
- Many models can be expressed as an F-probability[8] $P_\vartheta(\cdot) = [L_\vartheta(\cdot), U_\vartheta(\cdot)]$ where for a suitable function $g : [0,1] \to [0,1]$ and arbitrary events A the lower interval limit $L_\vartheta(A)$ takes the form

$$L_\vartheta(A) = g(p_\vartheta(A)) \,. \tag{1}$$

Then $g(\cdot)$ is called *distortion function* and $p_\vartheta(\cdot)$ *central distribution.*[9]

[8] Small exercise: Show that the fact that $P(\cdot)$ from (1) is an F-probability implies $g(t) \le t$, for all $t \in [0,1]$.

[9] For the ϵ–contamination model take $g(t) = (1-\epsilon)\cdot t$.

Brief excursus: ideas based on neighborhood models in machine learning

- neighbourhood models help to avoid overfitting: lower entropy (Abellan & Moral (2003, IJUFKBS), Strobl (2005, ISIPTA))
- extended, for instance, in Fink (2018, Diss LMU), Fink (2018, Imptree:CRAN)
- abstain from predictions when the uncertainty is too high
- better interpretability without loosing much predictive power?

- summarize complex ensemble by easy to interpret tree with soft boundaries? Nalenz & Augustin (2021, AIStat)

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Bayes postulate (not decision theoretic)

After having observed the sample, the posterior distribution contains the full information, i.e., it describes the knowledge about the unknown parameter completely.
All statistical analyzes must rely exclusively on the posterior; in particular, the construction of

- Bayesian point estimates: *MPD estimators (Maximum Posterior Density estimators)*
- Bayesian interval estimates: *HPD intervals (Highest posterior density intervals)*
- Bayes tests.
- Furthermore, the following *Updating Principle* is used: When drawing a further sample, the posterior distribution is used as the new prior distribution. In this sense, conditional Bayes inference is often referred to as "updating the prior".

Bayes learning: fundamental scheme

Prior knowledge: $\pi(\vartheta)$
sampling distribution $f(x|\vartheta)$
$+$observation x

$\Longrightarrow$ current knowledge $\pi(\vartheta|x)$: posterior

Bayes learning: fundamental scheme

Think of the data coming in sequentially in batches at times $t_1, t_2, \ldots$. ("Online learning")

$$\text{prior}_{t_1} \xrightarrow{\text{data}_1} \text{posterior}_{t_1} = \text{prior}_{t_2} \xrightarrow{\text{data}_2} \text{posterior}_{t_2} = \text{prior}_{t_3} \ldots$$

Bayes learning: fundamental scheme

Think of the data coming in sequentially in batches at times $t_1, t_2, \ldots$. ("Online learning")

$$\text{prior}_{t_1} \stackrel{\text{data}_1}{\longrightarrow} \text{posterior}_{t_1} = \text{prior}_{t_2} \stackrel{\text{data}_2}{\longrightarrow} \text{posterior}_{t_2} = \text{prior}_{t_3} \ldots$$

That can be done in a particularly convenient way when prior and posterior are guaranteed to be from the same parametric family of distributions. The distributions describing the sampling model and the prior are then called *conjugated* to each other.

Bayes learning: fundamental scheme

Think of the data coming in sequentially in batches at times $t_1, t_2, \ldots$. ("Online learning")

$$\text{prior}_{t_1} \xrightarrow{\text{data}_1} \text{posterior}_{t_1} = \text{prior}_{t_2} \xrightarrow{\text{data}_2} \text{posterior}_{t_2} = \text{prior}_{t_3} \ldots$$

That can be done in a particularly convenient way when prior and posterior are guaranteed to be from the same parametric family of distributions. The distributions describing the sampling model and the prior are then called *conjugated* to each other.
Then, with γ the parameter describing the prior,

$$\gamma_{t_1} \xrightarrow{\text{data}_1} \gamma_{t_2} \xrightarrow{\text{data}_2} \gamma_3 \ldots$$

Examples for conjugacy between prior and sampling distribution

- normal-normal for the inference on the mean of a normal distribution
- beta-binomial model for inference on binary samples
- Dirichlet-multinomial model for inference on categorical data
- gamma-Poisson model for inference on count data
- ...

Conjugacy in canonical exponential families

See, e.g., Bernardo & Smith (2000, pp. 202 and 272f) and Quaeghebeur & de Cooman (2005, ISIPTA) for the first extension to IP.

- For the moment only special case: real-valued, canonical parameter ϑ
- n i.i.d. observations: sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$
- sampling model canonical exponential family, sufficient statistic $\tau(\mathbf{x})$, density/probability function

$$f(\mathbf{x}|\vartheta) \propto \exp\Big(\vartheta\tau(\mathbf{x}) - nb(\vartheta)\Big), \tag{2}$$

- conjugacy whenever prior has the form

$$\pi(\vartheta|n^{(0)}, y^{(0)}) \propto \exp\Big(n^{(0)}\Big[y^{(0)} \cdot \vartheta - b(\vartheta)\Big]\Big) \tag{3}$$

Conjugacy in canonical exponential families (continued)

- prior with parameter (to be chosen by the researcher!)

$$\Big(\underbrace{y^{(0)}}_{\text{prior guess}}, \underbrace{n^{(0)}}_{\substack{\text{prior strength} \\ \text{virtual sample size}}}\Big)$$

- posterior with parameter $\left(y^{(n)}, n^{(n)}\right)$ where

$$y^{(n)} = \underbrace{\frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\mathbf{x})}{n}}_{\text{weighted mean}}, \quad n^{(n)} = \underbrace{n^{(0)} + n\,.}_{\text{overall sample size}} \qquad (4)$$

Conjugacy in canonical exponential families (continued)

- prior with parameter (to be chosen by the researcher!)

$$\Big(\underbrace{y^{(0)}}_{\text{prior guess}}, \quad \underbrace{n^{(0)}}_{\substack{\text{prior strength} \\ \text{virtual sample size}}}\Big)$$

- posterior with parameter $\left(y^{(n)}, n^{(n)}\right)$ where

$$y^{(n)} = \underbrace{\frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\mathbf{x})}{n}}_{\text{weighted mean}}, \quad n^{(n)} = \underbrace{n^{(0)} + n}_{\text{overall sample size}}. \tag{4}$$

- $n^{(n)}$ is independent of the concrete observation of the sample
- If $n^{(0)}$ had been larger, $y^{(0)}$ would have received more weight.
- If n had been larger, the observed value $\frac{\tau(\mathbf{x})}{n}$ would have received more weight.

Example: normal-normal model

Inference on μ, with known variance σ_0^2

$$f(x|\mu,\sigma_0^2) \propto \exp\left\{\frac{\mu}{\sigma_0^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma_0^2}\right\}.$$

Thus, $\vartheta = \frac{\mu}{\sigma_0^2}$, $b(\vartheta) = \frac{\mu^2}{2\sigma_0^2}$, $\tau(x) = \sum_{i=1}^n x_i$, and for the conjugate prior

Example: normal-normal model

Inference on μ, with known variance σ_0^2

$$f(x|\mu,\sigma_0^2) \propto \exp\left\{\frac{\mu}{\sigma_0^2}\sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma_0^2}\right\}.$$

Thus, $\vartheta = \frac{\mu}{\sigma_0^2}$, $b(\vartheta) = \frac{\mu^2}{2\sigma_0^2}$, $\tau(x) = \sum_{i=1}^n x_i$, and for the conjugate prior

$$\pi\left(\frac{\mu}{\sigma_0^2} \,\middle|\, n^{(0)}, y^{(0)}\right) \propto \exp\left\{n^{(0)}\left(\langle y^{(0)}, \frac{\mu}{\sigma_0^2}\rangle - \frac{\mu^2}{2\sigma_0^2}\right)\right\},$$

and, transformed to the parameter of interest μ,

$$\pi\left(\mu|n^{(0)}, y^{(0)}\right) \propto \frac{1}{\sigma_0^2}\exp\left\{-\frac{n^{(0)}}{2\sigma_0^2}(\mu - y^{(0)})^2\right\} d\mu\,, \text{i.e.,}$$

$$\mu \sim \mathcal{N}(y^{(0)}, \frac{\sigma_0^2}{n^{(0)}})$$

The parameters of the posterior distribution are

$$y^{(n)} = \mathbb{E}[\mu|x] = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \bar{x} \tag{5}$$

$$\frac{\sigma_0^2}{n^{(n)}} = \mathbb{V}(\mu|x) = \frac{\sigma_0^2}{n^{(0)} + n} \, . \tag{6}$$

The parameters of the posterior distribution are

$$y^{(n)} = \mathbb{E}[\mu|x] = \frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \bar{x} \tag{5}$$

$$\frac{\sigma_0^2}{n^{(n)}} = \mathbb{V}(\mu|x) = \frac{\sigma_0^2}{n^{(0)}+n} \,. \tag{6}$$

- Indeed, the posterior expectation of μ is a weighted average of the prior expectation $y^{(0)}$ and the sample mean $\bar{x}$.
- The updating decreases the variance by the factor $n^{(0)}/(n^{(0)}+n)$.
- The variance is the larger, the larger σ_0^2, i.e. the larger the variability of the sample.

General aspects and some caveats

- For $n \longrightarrow \infty$ full weight on the sample, irrespective of prior: "asymptotic objectivity", holds more general under mild regularity conditions. Posterior asymptotically concentrates around the true parameter value.

- For finite (not very large n), the parameters of the prior have to be specified by the researcher, and this choice substantially influences the result

General aspects and some caveats (Continued)

Recall the discussion above

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

  "Bayesian methods are increasingly used in proof-of-concept studies. An important benefit of these methods is the potential to use informative priors, thereby reducing sample size. This is particularly relevant for treatment arms where there is a substantial amount of historical information such as placebo and active comparators." (Mutsvar, Tytgat & [Ros.] Walley, 2016, PharmaceutStatist, p. 28)

General aspects and some caveats (Continued)

Recall the discussion above

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

- But is the knowledge truly precise enough?
- What to do under ignorance?
- How to express valuable *partial knowledge*?

General aspects and some caveats (Continued)

Recall the discussion above

- Promises explicit incorporation of knowledge, e.g. "borrowing strength" to discover effects more quickly.

- But is the knowledge truly precise enough?
- What to do under ignorance?
- How to express valuable *partial knowledge*?

- insensitivity towards prior-data conflict (see below)

Prior-data conflict

Recall

$$y^{(n)} = \underbrace{\frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\mathbf{x})}{n}}_{\text{weighted mean}}, \qquad n^{(n)} = \underbrace{n^{(0)} + n}_{\text{overall sample size}}.$$

Prior-data conflict

Recall

$$y^{(n)} = \underbrace{\frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\mathbf{x})}{n}}_{\text{weighted mean}}, \qquad n^{(n)} = \underbrace{n^{(0)} + n}_{\text{overall sample size}}.$$

Example: Let in the normal-normal model $n^{(0)} = 5$ and $n = 20$, and consider the following three situations:

	$y^{(0)} quad$	$\frac{\tau(\mathbf{x})}{n}$
$a)$	-0.1	0.025
$b)$	-1	0.25
$c)$	-10	2.5

In a), the prior guess for the mean and sample mean are very close to each other, while in c) there is a big discrepancy between what was anticipated to occur and what was de facto observed. Assuming that no outliers have occurred, there is a severe *prior-data conflict*. (b) is somewhat in between.)

Prior-data conflict

Recall

$$y^{(n)} = \underbrace{\frac{n^{(0)}}{n^{(0)}+n} \cdot y^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{\tau(\mathbf{x})}{n}}_{\text{weighted mean}}, \qquad n^{(n)} = \underbrace{n^{(0)} + n}_{\text{overall sample size}}.$$

Example: Let in the normal-normal model $n^{(0)} = 5$ and $n = 20$, and consider the following three situations:

	$y^{(0)} quad$	$\frac{\tau(\mathbf{x})}{n}$
$a)$	-0.1	0.025
$b)$	-1	0.25
$c)$	-10	2.5

In a), the prior guess for the mean and sample mean are very close to each other, while in c) there is a big discrepancy between what was anticipated to occur and what was de facto observed. Assuming that no outliers have occurred, there is a severe *prior-data conflict*. (b) is somewhat in between.)

Prior-data conflict

What to do if prior information and (outlier-free) sample information are conflicting (and the sample is too small to rule out the effect of the prior) ?

> *"[...] if we can show that the observed data is surprising in light of the sampling model and the prior, then we must be at least suspicious about the validity of the inferences drawn [...]." (Evans & Moshonov, 2006, BayesianAnal, p. 893)*

Prior-data conflict

In all three situations of the example, one obtains the same posterior mean

$$y^{(n)} = 0$$

Prior-data conflict

In all three situations of the example, one obtains the same posterior mean

$$y^{(n)} = 0$$

and also the same variance, since

$$n^{(n)} \equiv n^{(0)} + n = 25\,,$$

and thus the same distribution.

Prior-data conflict

In all three situations of the example, one obtains the same posterior mean

$$y^{(n)} = 0$$

and also the same variance, since

$$n^{(n)} \equiv n^{(0)} + n = 25\,,$$

and thus the same distribution.
Conflicting information goes beyond variability, and thus can not be captured by the variance or other characteristics of precise probabilistic models.

Table of contents

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Powerful models

credal prior: set of prior distributions representing partial knowledge

- set of expert opinions as credal prior
- ordinal probabilities
- indivisible evidence
- parametrically constructed: utilize the parametric models just discussed with interval-valued parameter components (set of means, set of variances etc.) bounds on densities or distribution functions
- neighborhood models, e.g. distorted probabilities

Posterior loss versus prior risk

In the general framework developed later, it turns out that there is typically no counterpart to the main theorem of Bayesian decision theory. One has to decide whether to take

- the conditional perspective based on (some notion of) generalized posterior loss optimality
- the strategic perspective looking for decision functions minimizing (some notion of) generalized expected prior risk

For the moment: conditional perspective

- credal prior: F-probability Π or credal set $\mathcal{M}$,
- after having observed x update it to obtain the credal posterior Π_x or credal set $\mathcal{M}_x$
- take the credal posterior as the basis of all inferences and decision procedures (generalized Bayes postulate, compare with Remark 2.52)
- decision theoretic criteria (E-Admissibility, MaxEMin, ldots) directly applicable

Inference with credal posterior, some properties

- natural ordering with respect to "$\subseteq$":

$$\mathcal{M}^{(1)} \subseteq \mathcal{M}^{(2)} \iff \mathcal{M}_x^{(1)} \subseteq \mathcal{M}_x^{(2)}$$

Inference with credal posterior, some properties

- natural ordering with respect to "$\subseteq$":

$$\mathcal{M}^{(1)} \subseteq \mathcal{M}^{(2)} \Longleftrightarrow \mathcal{M}_x^{(1)} \subseteq \mathcal{M}_x^{(2)}$$

- "asymptotic objectivity" remains: (By general theory, it is, under regularity conditions, valid for all prior probabilities, and thus, in particular, for all elements of *calM*.)

Inference with credal posterior, some properties

- natural ordering with respect to "$\subseteq$":

$$\mathcal{M}^{(1)} \subseteq \mathcal{M}^{(2)} \Longleftrightarrow \mathcal{M}_x^{(1)} \subseteq \mathcal{M}_x^{(2)}$$

- "asymptotic objectivity" remains: (By general theory, it is, under regularity conditions, valid for all prior probabilities, and thus, in particular, for all elements of *calM*.)

- a closer look at extensions of the conjugated models in exponential families

Extensions of the conjugated models in exponential families

- precise sampling distribution from canonical exponential family in the form (2)
- credal prior described by parameter set $\Pi^{(0)} \subseteq \mathcal{Y}^{(0)} \times \mathcal{N}^{(0)}$, with $\mathcal{Y}^{(0)}$ and $\mathcal{N}^{(0)}$ sets of $y^{(0)}-$ and $n^{(0)}-$values in the sense of (3) (called *conjugated credal priors* here)
- applying GBR yields the credal posterior as a set of conjugated distributions described by[10]

$$\Pi^{(n)} := \left\{ (y^{(n)}, n^{(n)}) \,\middle|\, \exists (y^{(0)}, n^{(0)}) : y^{(n)}, n^{(n)} \text{ obey to (4)} \right\}$$

[10] (4) was:

$$y^{(n)} = \underbrace{\frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{\tau(\mathbf{x})}{n}}_{\text{weighted mean}}, \qquad n^{(n)} = \underbrace{n^{(0)} + n}_{\text{overall sample size}}.$$

Near-ignorance models

These models allow for the formulation of *near-ignorance models* on the parameter space, i.e. the specification of a prior credal sets $\ddot{\mathcal{M}}$ of probabilities on $\Theta, \sigma(\Theta)$ with

$$\inf_{\pi \in \ddot{\mathcal{M}}} \pi(Q) = 0 \quad \sup_{\pi \in \ddot{\mathcal{M}}} \pi(Q) = 1\,, \qquad Q \in \mathcal{Q}\,,$$

with $\mathcal{Q}$ containing the "standard events of interest"[11].

[11] Taking $\mathcal{Q} = \sigma(\Theta) \setminus \{\varnothing\}$ would lead to a entirely vacuous posterior $P(Q|x) = [0,1]$ for all $Q \in \mathcal{Q}$.

Work on near-ignorance models

- most prominent is the *imprecise Dirichlet model (IDM)* Walley, 1996, JRSSB for categorical inference under prior-near ignorance
- for general exponential families, one-parametric: Benavoli & Zaffalon (2012, JStatPlanInf), multivariate form Benavoli & Zaffalon (2014, Statistics)
- Gaussian processes: Mangili (2015, ISIPTA; 2017, IntJApproxReason)
- for recent machine learning applications, see, in the case of the IDM, Utkin (2019, Neurocomputing), Utkin (2020, ExpSysAppl), Moral-Garcia et al (2020 ExpSysAppl), for the multivariate normal model, Carranza Alarcon & Destecke (2021, Pattern Recognition), and for the imprecise Gaussian processes, Rodemann (2021, MSc LMU), Rodemann & Augustin (2021, IUKM)

Convenient special case: interval-valued parameters

- interval-valued prior location parameter

$$\left[\underline{y}^{(0)}, \overline{y}^{(0)}\right]$$

and/or

- interval-valued prior strength / number of virtual observations

$$\left[\underline{n}^{(0)}, \overline{n}^{(0)}\right]$$

Convenient special case: interval-valued parameters

- interval-valued prior location parameter

$$\left[\underline{y}^{(0)}, \bar{y}^{(0)}\right]$$

  and/or

- interval-valued prior strength / number of virtual observations

$$\left[\underline{n}^{(0)}, \bar{n}^{(0)}\right]$$

Shows also attractive behavior under prior data conflict

Prior-data conflict

Consider an i.i.d. sample from a normal distribution and conjugated credal priors based on $\Pi^{(0)} = \mathcal{Y}^{(0)} \times \mathcal{N}^{(0)}$ with $\mathcal{Y}^{(0)} = [\underline{y}^{(0)}, \bar{y}^{(0)}]$ and $\mathcal{N}^{(0)} = [\underline{n}^{(0)}, \bar{n}^{(0)}]$. For the credal posterior based on $\Pi^{(n)}$ and with

$$\underline{y}^{(n)} := \inf_{(y^{(n)}, n^{(n)}) \in \Pi^{(0)}} y^{(n)} \qquad \text{and} \qquad \overline{y}^{(n)} := \sup_{(y^{(n)}, n^{(n)}) \in \Pi^{(0)}} y^{(n)}$$

it holds that

$$\underline{y}^{(n)} = \begin{cases} \dfrac{\overline{n}^{(0)}\underline{y}^{(0)} + n\bar{x}}{\overline{n}^{(0)} + n} & \bar{x} \geq \underline{y}^{(0)} \\[2ex] \dfrac{\underline{n}^{(0)}\underline{y}^{(0)} + n\bar{x}}{\underline{n}^{(0)} + n} & \bar{x} < \underline{y}^{(0)} \end{cases}, \quad \overline{y}^{(n)} = \begin{cases} \dfrac{\overline{n}^{(0)}\overline{y}^{(0)} + n\bar{x}}{\overline{n}^{(0)} + n} & \bar{x} \leq \overline{y}^{(0)} \\[2ex] \dfrac{\underline{n}^{(0)}\overline{y}^{(0)} + n\bar{x}}{\underline{n}^{(0)} + n} & \bar{x} > \overline{y}^{(0)} \end{cases}.$$

Prior-data conflict

In particular, for the "posterior imprecision in the means"

$$\overline{y}^{(n)} - \underline{y}^{(n)} = \frac{\overline{n}^{(0)}(\overline{y}^{(0)} - \underline{y}^{(0)})}{\overline{n}^{(0)} + n} + \underbrace{\inf_{y^{(0)} \in \mathcal{Y}^{(0)}} \left| \bar{x} - y^{(0)} \right|}_{\text{prior-data conflict}} \frac{n\,(\overline{n}^{(0)} - \underline{n}^{(0)})}{(\underline{n}^{(0)} + n)(\overline{n}^{(0)} + n)}\,.$$

Prior-data conflict

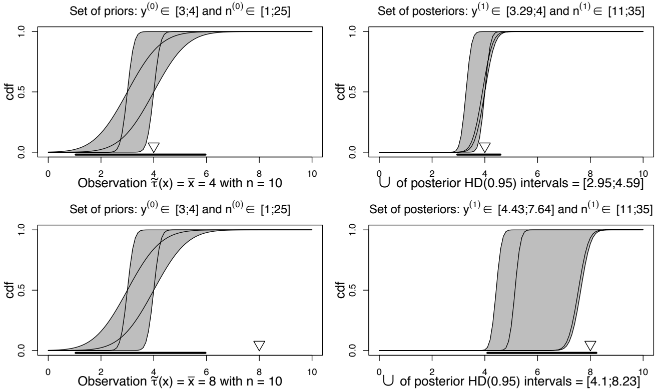


Figure: Taken from Walter & Augustin (2009, JStatThPrac p. 268)

Prior-data conflict

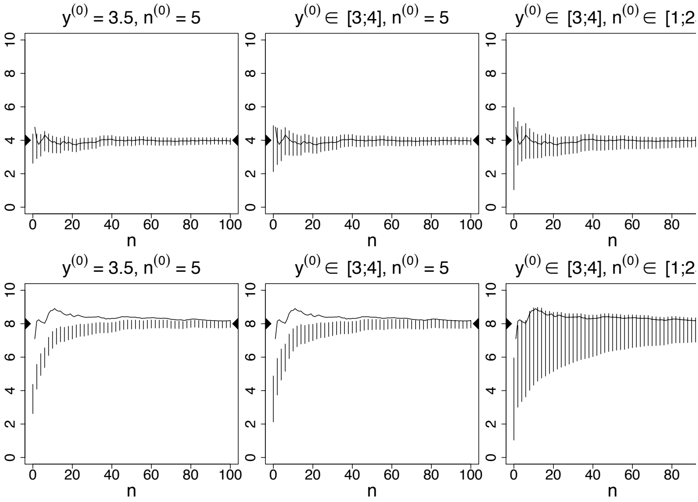


Figure: Taken from Walter & Augustin (2009, JStatThPrac p. 268)

Table of contents

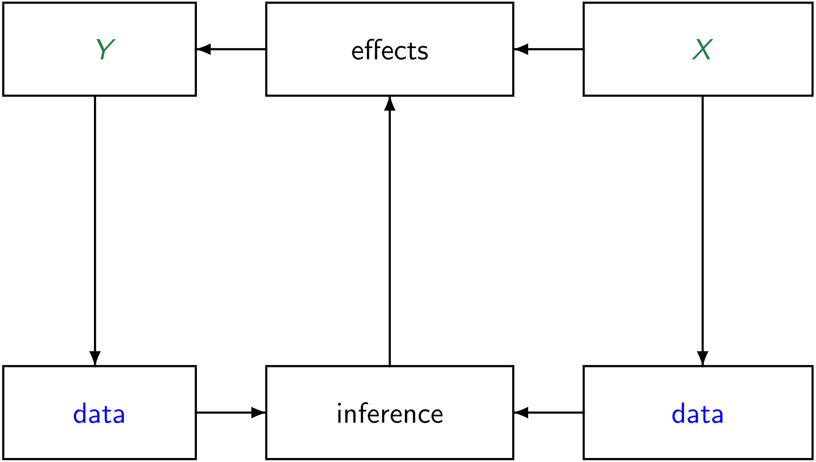
1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Classical View Point: Sampling Uncertainty

- use of probability theory
- quantifies the error made by certain inference procedures
  - tests
  - point and interval estimators

- decreases with increasing sample size n
- goes to zero for $n \to \infty$

Complex Relationships between Variables



Typical examples: Measurement Error

Quite often the relationship between theoretically formulated variables and the observed data is rather complex, too.

- Error-prone measurements of true quantities
  - ◇ error in technical devices
  - ◇ indirect measurement
  - ◇ response effects
  - ◇ use of aggregated quantities, averaged values, imputation, rough estimates etc.
  - ◇ anonymization of data by deliberate contamination
- Measured indicators of complex constructs; latent variables
  - ◇ long term quantities: long term protein intake, long term blood pressure
  - ◇ permanent income
  - ◇ importance of a patent
  - ◇ extent of motivation, degree of costumer satisfaction
  - ◇ severity of malnutrition
  - ◇ ...

Big Data Uncertainty

Quite often the relationship between theoretically formulated variables and the observed data is rather complex, too.

Big Data Uncertainty

Quite often the relationship between theoretically formulated variables and the observed data is rather complex, too.

- measurement error and misclassification (including operationalization of complex constructs, anonymized data)
- rounding and heaping
- omitted variables
- coarsening
- censoring
- missing data (including missingness by design: treatment evaluation, statistical matching)

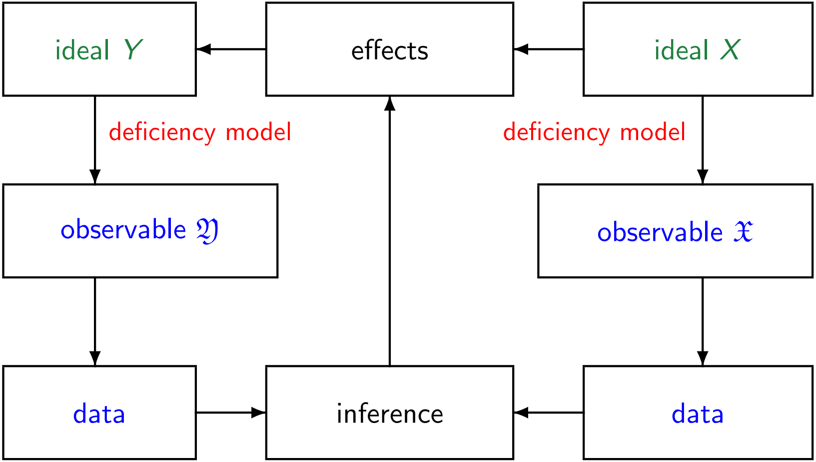
Big Data Uncertainty

Quite often the relationship between theoretically formulated variables and the observed data is rather complex, too.

- measurement error and misclassification (including operationalization of complex constructs, anonymized data)
- rounding and heaping
- omitted variables
- coarsening
- censoring
- missing data (including missingness by design: treatment evaluation, statistical matching)

big data uncertainty: such uncertainty does **not** diminish with increasing sample size

The two-layers perspective



What to do?

What to do? Make **assumptions** to be able

- to justify ignoring this uncertainty or
- to correct for it, etc. by integrating its effects out
- for instance:
  - missing / coarsening at random (MAR / CAR), noninformative censorship,
  - measurement error models

What to do? Make **assumptions** to be able

- to justify ignoring this uncertainty or
- to correct for it, etc. by integrating its effects out
- for instance:
  - missing / coarsening at random (MAR / CAR), noninformative censorship,
  - measurement error models
- "classical model of testing theory": Measurement error model must be known precisely
  - type of error, especially assumptions on (conditional) independence
    - independence of true value
    - independence of other covariates
    - independence of other measurements
  - type of error distribution
  - moments of error distribution

  validation studies typically not available

Assumptions as information

> *"There is always a trade-off between assumptions and data – both bring information. With better data, fewer assumptions are needed."*

Rubin (2005, JASA, here p. 324); compare also the talk by Elisabeth Stuart in the last Institutskolloquium

Quote taken from Rubin in more detail

"Nothing is wrong with making assumptions; causal inference is impossible without making assumptions, and they are the strands that link statistics to science. It is the scientific quality of those assumptions, not their existence, that is critical. There is always a trade-off between assumptions and data – both bring information. With better data, fewer assumptions are needed. But in the causal inference setting, assumptions are always needed, and it is imperative that they be explicated and justified. One reason for providing this detail is so that readers can understand the basis of conclusions. A related reason is that such understanding should lead to scrutiny of the assumptions, investigation of them, and, ideally, improvements. Sadly, this stating of assumptions is typically absent in many analyses purporting to be causal and replaced by a statement of what computer programs were run, which I regard as entirely inadequate scientifically."

Rubin (2005, JASA, here p. 324)

Missing data

- response $Y_1, Y_2, \ldots, Y_n$, covariates $X_1, X_2, \ldots, X_n$
- for the moment, missingness in Y variable only
- missingness/observability indicator $R \in \{0, 1\}$

Missing data

- response $Y_1, Y_2, \ldots, Y_n$, covariates $X_1, X_2, \ldots, X_n$
- for the moment, missingness in Y variable only
- missingness/observability indicator $R \in \{0, 1\}$

- missingness complete at random (MCAR): R independent of X and Y
- missingness at random (MAR): R may dependent on X, but is independent of Y
- missingness not at random (NMAR): else

Missing data

- response $Y_1, Y_2, \ldots, Y_n$, covariates $X_1, X_2, \ldots, X_n$
- for the moment, missingness in Y variable only
- missingness/observability indicator $R \in \{0, 1\}$

- missingness complete at random (MCAR): R independent of X and Y
- missingness at random (MAR): R may dependent on X, but is independent of Y
- missingness not at random (NMAR): else

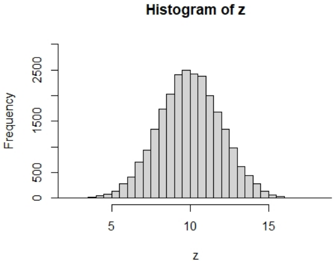
- many statistical results and techniques rely on MAR (or MCAR)
- for instance multiple imputation or the EM-algorithm

How to test between MAR and MNAR?

Motivating simulation example

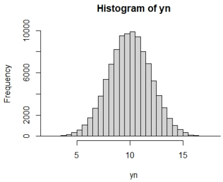
Simplification for illustration: no covariates, thus MAR = MCAR

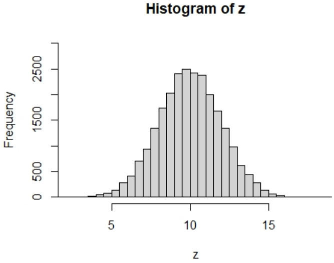
Motivating simulation example



Motivating simulation example

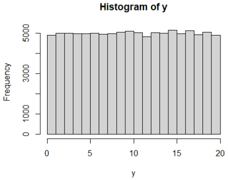
normal distribution + MCAR



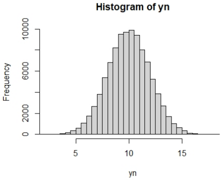


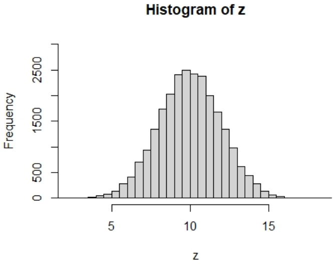
Motivating simulation example

uniform distribution + NMAR



normal distribution + MCAR





How to test between MAR and MNAR?

No chance to distinguish between MAR and MNAR on empirical grounds only.
To every MAR situation there are infinitely many models that lead to the same observable distribution.

$$P(Y = y|R = 1) = \frac{P(R = 1|Y = y) \cdot P(Y = y)}{P(R = 1)}$$

Recap: traditional handling of big data uncertainty

What to do? Make **assumptions** to be able

- to ignore this uncertainty
- to correct for it, etc. by integrating its effects out
- for instance:
  - missing / coarsening at random (MAR / CAR), noninformative censorship,
  - measurement error models

But these assumptions

- are assumptions on the relationships of unobservable quantities,
- are thus by themselves not testable, different models lead to the same data,
- and thus need indispensably external justification by background domain knowledge.

Manski's Law of Decreasing Credibility

Credibility ?

"The credibility of inference decreases with the strength of the assumptions maintained." (Manski (2003, p. 1))

Charles Manski[12]

[12] http://faculty.wcas.northwestern.edu/~cfm754/; [August 16th, 2022]

Manski's Law of Decreasing Credibility

Credibility ?
"The credibility of inference decreases with the strength of the assumptions maintained." (Manski (2003, p. 1))

partial identification: Set of all models compatible with the data and tenable assumptions.

Charles Manski[13]

[13] http://faculty.wcas.northwestern.edu/~cfm754/; [August 16th, 2022]

Reliable inference instead of overprecision!!

Consequences to be drawn from the Law of Decreasing Credibility:

- adding untenable assumptions to produce precise solution may destroy credibility of statistical analysis, and therefore its relevance for the subject matter questions.
- make *realistic* assumptions and consider the *set* of *all* models that are compatible with the data and these assumptions (and then add successively additional assumptions, if desirable)
- the results may be imprecise, but are more reliable
- the extent of imprecision is related to data quality!
- as a welcome by-product: clarification of the implication of certain assumptions
- often still sufficient to answer subjective matter question
- "weak information" may be powerful in refining results

Partial identification

- Classical setting:

$$\text{Big data uncertainty} \qquad \longrightarrow \qquad \text{no identification}$$

OR

$$\text{Big data uncertainty} \qquad \xrightarrow{\text{strong assumptions}} \qquad \text{single model}$$

Partial identification

- Classical setting:

$$\text{Big data uncertainty} \quad \longrightarrow \quad \text{no identification}$$

OR

$$\text{Big data uncertainty} \quad \xrightarrow{\text{strong assumptions}} \quad \text{single model}$$

- Now

$$\text{Big data uncertainty} \quad \longrightarrow \quad \textbf{partial} \text{ identification}$$

set of models

Election Forecasting with Yet Undecided Voters

- Project with the polling institute Civey, together with Dominik Kreiss
- pre-election polling data for the 2021 German federal election
- new questionnaire design: explicit collection of the consideration sets (Oscarsson & Rosema (2019, Elect.Stud)) of yet undecided voters ("Between which parties are you undecided?")
- valuable information far beyond "don't know":
  - typically indecisiveness only between (very) few parties
  - precise vote for all coalitions containing parties in the voter's consideration set
- Kreiss & Augustin (2021, ArXiv) and the work cited therein

- S set of parties standing for election
- two levels of (generic) response variables
  - $\mathfrak{Y}$: consideration set, set $\mathfrak{l}$ of preferred parties, observable
  - Y: final choice, party $\ell \in \mathfrak{l}$, not observable
  - covariates X, realizations x

- S set of parties standing for election
- two levels of (generic) response variables
  - $\mathfrak{Y}$: consideration set, set $\mathfrak{l}$ of preferred parties, observable
  - Y: final choice, party $\ell \in \mathfrak{l}$, not observable
  - covariates X, realizations x
- point estimator for percentage of votes a set A of parties achieves

$$\widehat{p}(Y \in A) = \sum_{\substack{(\ell, \mathfrak{l}, x) \in \\ A \times \mathcal{P}(S) \times \mathcal{X}}} \underbrace{p(Y = \ell \mid \mathfrak{Y} = \mathfrak{l}, X = x)}_{\text{latent transition model}} \cdot \underbrace{\widehat{p}(\mathfrak{Y} = \mathfrak{l} \mid X = x)}_{\text{from data}} \cdot \underbrace{\widehat{p}(X = x)}_{\text{from data, sampling weights}}$$

- S set of parties standing for election
- two levels of (generic) response variables
  - $\mathfrak{Y}$: consideration set, set $\mathfrak{l}$ of preferred parties, observable
  - Y: final choice, party $\ell \in \mathfrak{l}$, not observable
  - covariates X, realizations x
- point estimator for percentage of votes a set A of parties achieves

$$\widehat{p}(Y \in A) = \sum_{\substack{(\ell, \mathfrak{l}, x) \in \\ A \times \mathcal{P}(S) \times \mathcal{X}}} \underbrace{p(Y = \ell \mid \mathfrak{Y} = \mathfrak{l}, X = x)}_{\text{latent transition model}} \cdot \underbrace{\widehat{p}(\mathfrak{Y} = \mathfrak{l} \mid X = x)}_{\text{from data}} \cdot \underbrace{\widehat{p}(X = x)}_{\text{from data, sampling weights}}$$

Structure of the equation

- $p(Y = \ell) = \sum_x p(Y \in \ell | X = x) \cdot p(X = x)$
- $p(Y \in A) = \sum_{x,\ell} p(Y \in \ell | X = x) \cdot p(X = x)$
- Now condition on a further variable, Z with values z say (later set $Z = \mathfrak{Y}$ with values $z = \mathfrak{l}$)
  $p(Y \in A) = \sum_{x,\ell,z} p(Y \in \ell | Z = z, X = x) \cdot p(Z = z | X = x) \cdot p(X = x)$
- go over to "hats" to express estimation

- S set of parties standing for election
- two levels of (generic) response variables
  - $\mathfrak{Y}$: consideration set, set $\mathfrak{l}$ of preferred parties, observable
  - Y: final choice, party $\ell \in \mathfrak{l}$, not observable
  - covariates X, realizations x
- point estimator for percentage of votes a set A of parties achieves

$$\widehat{p}(Y \in A) = \sum_{\substack{(\ell, \mathfrak{l}, x) \in \\ A \times \mathcal{P}(S) \times \mathcal{X}}} \underbrace{p(Y = \ell \mid \mathfrak{Y} = \mathfrak{l}, X = x)}_{\text{latent transition model}} \cdot \underbrace{\widehat{p}(\mathfrak{Y} = \mathfrak{l} \mid X = x)}_{\text{from data}} \cdot \underbrace{\widehat{p}(X = x)}_{\text{from data, sampling weights}}$$

"Modelling"

- For the moment let's argue without the covariates: $p_{(\mathfrak{l},x)} \hookrightarrow p_{(\mathfrak{l})}$
- Thinking of a concrete example may be helpful; consider, e.g., $\mathfrak{l} = \{SPD, Left, Green\}$.
- See above: results depend strongly on the unknown transition model.
- Therefore, think of the forecast as a function of the transition model underlying, i.e. consider

$$\widehat{p}(Y \in A)\left[(p_{\mathfrak{l}})_{(\mathfrak{l} \in \mathcal{P}(S))}.\right]$$

"Precise modelling"

Potential ideas to specify the latent transition model precisely:

- prophetic: give exact numbers for $(p_{\mathfrak{l}})_{(\mathfrak{l}\in\mathcal{P}(S))}$
- transfer knowledge from polls of older elections
- uniform (max ent)

$$p(Y=\ell \mid \mathfrak{Y}=\mathfrak{l}) := \frac{1}{|\mathfrak{l}|}$$

- homogeneous with respect to the decided

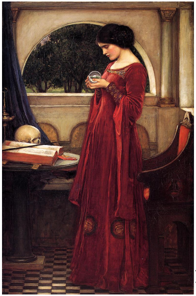
$$p(Y=\ell \mid \mathfrak{Y}=\mathfrak{l}) := \frac{p(\mathfrak{Y}=\{\ell\})}{\sum_{\ell'\in\mathfrak{l}} p(\mathfrak{Y}=\{\ell'\})}$$

- noninformativeness of coarsening (CAR: coarsening at random) (indirect)

$$\forall \mathfrak{l}\in\mathcal{P}(S) : \forall \ell_1,\ell_2\in\mathfrak{l} : \frac{p(Y=\ell_1|\mathfrak{Y}=\mathfrak{l})}{p(Y=\ell_2|\mathfrak{Y}=\mathfrak{l})} = \frac{p(Y=\ell_1)}{p(Y=\ell_2)}$$

Justification of these Assumptions

Justification of these assumptions

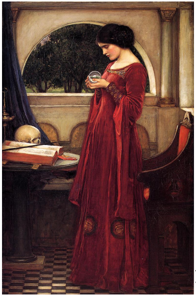
 [14]

[14] John William Waterhouse: The Crystal Ball (1902)

http://www.wikiart.org/en/john-william-waterhouse/the-crystal-ball-1902,pulicdomain, [Aug 16th, 2022

Justification of these assumptions

- Assumptions specifying the transition model have to be well-grounded in good subject-matter arguments, derived from the domain knowledge.
- All the assumptions just stated (and many more) are indistinguishable by relying on the data only.
- There CANNOT be any meaningful statistical test to support/reject any of these assumptions.


[14]

- Relying on such assumptions just for the sake of receiving (seemingly) precise solutions is questionable.

[14] John William Waterhouse: The Crystal Ball (1902)
http://www.wikiart.org/en/john-william-waterhouse/the-crystal-ball-1902,pulicdomain, [Aug 16th, 2022

What has the Theory of Partial Identification to Offer here?

- Enveloping all scenarios: worst- and best case estimates
- When weak, but well- supported information is available, utilize it to increase precision!

Enveloping all Possible Specifications of the Transition Model

- What do we know "for sure"?
- Consider all possible specifications for

$$\big(p(Y = \ell \mid \mathfrak{Y} = \mathfrak{l}, X = x)\big)_{\ell \in S, \mathfrak{l} \in \mathcal{P}(S)}$$

- That is, consider for each $\mathfrak{l}$, the set of all probabilities on $(\mathfrak{l}, \mathcal{P}(\mathfrak{l}))$.

Enveloping all Possible Specifications of the Transition Model

- What do we know "for sure"?
- Consider all possible specifications for

$$\Big(p(Y = \ell \mid \mathfrak{Y} = \mathfrak{l}, X = x)\Big)_{\ell \in S, \mathfrak{l} \in \mathcal{P}(S)}$$

- That is, consider for each $\mathfrak{l}$, the set of all probabilities on $(\mathfrak{l}, \mathcal{P}(\mathfrak{l}))$.
- By assuming error-freeness of coarsening

$$p(Y \in A) \mid \mathfrak{Y} = \mathfrak{l}, X = x)\Big) = \begin{cases} 0 & \mathfrak{l} \subseteq A^C \\ 1 \quad \text{if} & \mathfrak{l} \subseteq A \\ [0;1] & \mathfrak{l} \cap A \neq \varnothing \wedge \mathfrak{l} \cap A^C \neq \varnothing \end{cases}$$

Enveloping all Possible Specifications of the Transition Model (continued)

- $$p(Y \in A) \mid \mathfrak{Y} = \mathfrak{l}, X = x)) = \begin{cases} 0 & \text{if } \mathfrak{l} \subseteq A^C \\ 1 & \text{if } \mathfrak{l} \subseteq A \\ [0;1] & \mathfrak{l} \cap A \neq \varnothing \wedge \mathfrak{l} \cap A^C \neq \varnothing \end{cases}$$
- move probability mass around where not fixed
- **lower bound ("guarantee"):**

$$\underline{P}(Y \in A) = \sum_{\mathfrak{l} \subseteq A} p(\mathfrak{Y} = \mathfrak{l})$$

$$\underline{P}(SPD, Gr, FDP) = p(SPD) + p(Gr) + p(FDP) + p(SPD, Gr) + p(SPD, FDP) + p(Gr, FDP) + p(SPD, Gr, FDP)$$

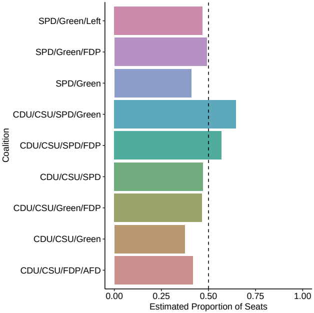
- **upper bound ("potential"):**

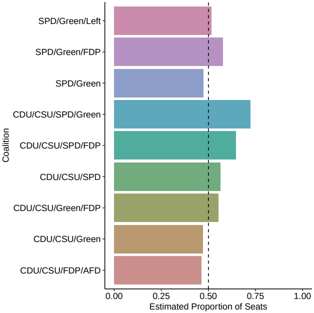
$$\overline{P}(Y \in A) = \sum_{\mathfrak{l} \cap A \neq \varnothing} p(\mathfrak{Y} = \mathfrak{l}) .$$

- Construction goes back to Dempster (1967, Ann.Math.Stat) and Shafer (1976, Princeton UP) in the context of fiducial inference and modelling uncertain knowledge, respectively.

Dempster Bounds

- most cautious analysis:[15] appropriate communication of full uncertainty about transitions
- Considerable increase in precision when coalitions are considered! For instance, being undecided between SPD and Green is a precise vote for any coalition containing these parties!

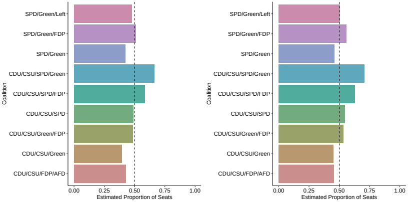




[15] Figure is taken from Kreiss & Augustin (2021, Arxiv; p. 10)

Exploit Weak Knowledge about Transition Probabilities

- weigh precision and credibility
- communication of the uncertainty present
- work with plausible weak assumptions not exploitable in traditional statistics
- expert opinions, like: “the undecided between Party I and Party II tend as least as much to Party I than to Party II”
- weaken “precise conditions” by considering neighborhood models
- generalized uniform probability: between 50-c% and 50+c% for all parties[16]
- easy technical handling via linear optimization



[16] Figure for $c = 30$ is taken from Kreiss & Augustin (2021, Arxiv; p. 10)

Looking Back

1. Introduction and Background
2. Statistics in a Nutshell
3. First Inquires on the Classical Approaches
4. Imprecise Probability for Statistics! First Ideas
5. Imprecise Sampling Models
   - Robustness Issues in Frequentist Estimation and Testing
   - Neighborhood Models
6. Bayesian Inference under Credal Sets
   - Classical Bayes Learning: Brief Repetition and the Concept of Conjugacy
   - Generalized Bayesian Inference
7. Selected Aspects of Data Imprecision
   - Big Data Uncertainty and Non-/Partial Identfiability
   - An Ongoing Case Study: Yet Undecided Voters
8. Concluding Remarks

Concluding Remarks

- IP (including partial identification) offers substantially new opportunities for sound statistical inference and modelling!
- so much yet to develop, explore and apply
- Bring in your enthusiasm and expertise!

- looking forward to vivid discussions here or later.

- thomas.augustin@stat.uni-muenchen.de