Open Science Framework DOI:10.31219/osf.io/***

Foundations of inference under symmetry:
A derivation of algorithms for non-parametric density inference

P.G.L. Porta Mana
<pgl@portamana.org>

15 January 2023; updated 5 February 2023 [draft]

***

0 A warning on notation

The probability calculus, being a generalization of the propositional truth-calculus, concerns propositions or statements. In the expression "P($A|B$)", A and B thus stand for statements or compositions of statements. In scientific applications most statements of interest are of the form "Measurement of the quantity X yields outcome x" or "The quantity X is set to the value x", implicitly or explicitly accompanied by other statements expressing contextual information, for instance "The measurement is made at time … in laboratory … under conditions …".

Writing full statements such as these within probability formulae would take an impractical amount of space. This impracticality is solved by an abuse of notation: outside a probability formula, a symbol like X may denote a quantity; but within a probability formula, say $\mathrm{p}(X \mid \dots)$, it instead denotes a full statement such as "Measurement of the quantity X yields outcome x"; sometimes the latter is written "$X = x$" when the value x is not generic or needs to be explicit.

To limit the confusion that can arise from this notation abuse, I use a convention similar to Jaynes's (2003 § 2.5 p. 43): the probability symbol "P" is only used when the symbols in its arguments univocally denote statements; the symbol "p" is used instead to warn that some of the symbols in its arguments are abused. For example, if the quantity N denotes the number of people satisfying some condition, A := "Measurement of N yields 101 people", and B denotes some contextual statement, then these three expressions have the same meaning:

$$\mathrm{P}(A \mid B) = 0.3\,, \qquad \mathrm{p}(N \mid B) = 0.3\,, \qquad \mathrm{p}(N = 101 \mid B) = 0.3\,.$$

We take the logical connectives ¬ (not), ∨ (or), ∧ (and) as a functionally complete (redundant) base set. A comma "," is interchangeably used for

$\wedge$ in probability formulae. The probability-calculus rules for the base connectives are

$$\begin{aligned} \mathrm{P}(\neg A \mid H) + \mathrm{P}(A \mid H) &= 1 \\ \mathrm{P}(A \wedge B \mid H) - \mathrm{P}(A \mid B \wedge H)\,\mathrm{P}(B \mid H) &= 0 \\ \mathrm{P}(A \vee B \mid H) + \mathrm{P}(A \wedge B \mid H) - \mathrm{P}(A \mid H) - \mathrm{P}(B \mid H) &= 0 \end{aligned} \tag{0}$$

which can be used in different guises, for example

$$\frac{\mathrm{P}(A|H)=x}{\mathrm{P}(\neg A|H)=1-x} \qquad \frac{\mathrm{P}(\neg A|H)=x}{\mathrm{P}(A|H)=1-x}$$

$$\frac{\mathrm{P}(A|B\wedge H)=x \quad \mathrm{P}(B|H)=y}{\mathrm{P}(A\wedge B|H)=xy} \qquad \frac{\mathrm{P}(A|B\wedge H)=x \quad \mathrm{P}(A\wedge B|H)=y}{\mathrm{P}(B|H)=y/x} \qquad \frac{\mathrm{P}(B|H)=x \quad \mathrm{P}(A\wedge B|H)=y}{\mathrm{P}(A|B\wedge H)=y/x}$$

$$\frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(B|H)=y \quad \mathrm{P}(A\wedge B|H)=z}{\mathrm{P}(A\vee B|H)=x+y-z} \qquad \frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(B|H)=y \quad \mathrm{P}(A\vee B|H)=z}{\mathrm{P}(A\wedge B|H)=x+y-z} \qquad \frac{\mathrm{P}(A|H)=x \quad \mathrm{P}(A\wedge B|H)=y \quad \mathrm{P}(A\vee B|H)=z}{\mathrm{P}(B|H)=y+z-x}$$

Compare with the inference rules for sequent calculi.[1]

1 Inference problem

The general context of our inference problem is the following. There is a collection of "units", each of which has an associated set of measurable quantities $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, which we call "variates" according to statistics terminology. The collection of variates $X_1, X_2, \ldots$ is denoted X, and similarly for Y. Examples: These "units" could be mechanical gadgets coming out of a production line; the X variates, some of their mechanical properties; the Y variates, other hidden properties or future events such as mechanical failure. Or the units could be clinical patients; the X variates, outcomes of clinical tests performed on these patients; the Y variates, medical conditions or future events such as life length or occurrence of some disease. Our inference context is thus quite general indeed. The units can be labelled with some index i, but the values of this index and their ordering have no informational relevance. The variate X_1 specific to unit i is denoted by X_1^i, and similarly for the other variates.

Our general goal is to make inferences about some of the variates for some of the units, given knowledge about other variates units.

One specific and frequently occurring inference is about the values of the variates Y^0 for some unit, call it $i = 0$, given:

[1] e.g. Huth & Ryan 2004 ch. 1; Prawitz 1965 App. A; see also Boričić 2020.

- the values of the variates X^0 for the same unit;
- the values of the X, Y variates for other units $i = 1, 2, \ldots$;
- some auxiliary information I^0 for the unit 0;
- all remaining relevant information, denoted I.

The most general inference gives probabilities to the possible values of Y^0. Completely certain inference (deduction) is included as a special case with 0 or 1 probabilities.

Thus in mathematical notation we want to determine the distribution of probability

$$\mathrm{p}(Y^0 \mid X^0,\ Y^1, X^1, Y^2, X^2, \ldots,\ I^0, I) \tag{1}$$

over the possible values of Y^0 (more precisely, over statements such as "Measurement of Y for unit 0 yields y").

Variations of this specific inference are also of interest. For instance, some of the X or Y variate values maybe be unknown for some units; or we want to infer about more than one unit; and similar variations. In the following we show the steps to specifically calculate the probabilities (1); variations of this calculation will be discussed later.

2 Initial probabilities

The propositional truth-calculus allows us to calculate the truth-values of non-tautological statements only if we give the truth-values of some other, related, non-tautological statements, usually called "premises". Likewise, the probability calculus allows us to calculate probability-values of non-tautological statements only if we give the probability-values of some other, related, non-tautological statements. These are usually called "prior" or "initial" probabilities (these adjectives do *not* imply that these probabilities concern statements about chronologically earlier events).

Therefore we must first assign the probability values of some statements which can lead us to the probabilities (1) through the rules (0).

There are several prior-probability assignments appropriate to our inference problem. All of them express observations of informational

symmetry. Their analysis goes under the name of "exchangeability"[2]. Let us start with the most important.

Consider the set of units $i = 1, 2, \ldots$, excluding $i = 0$ According to our information I, we have no reason a priori to assign special probabilities to particular units, before we know their variate values $Y^1, X^1, Y^2, X^2, \ldots$. This is reflected in the irrelevance of any possible permutation of their indices. This informational symmetry is expressed by the equality of all joint probabilities for sets of values that can be obtained from one another by permutations of the index i:

$$\begin{aligned}&\mathrm{p}(Y^1 = y^1, X^1 = x^1,\ Y^2 = y^2, X^2 = x^2,\ Y^3 = y^3, X^3 = x^3, \ldots \mid I) =\\ &\mathrm{p}(Y^1 = y^{r(1)}, X^1 = x^{r(1)},\ Y^2 = y^{r(2)}, X^2 = x^{r(2)},\ Y^3 = y^{r(3)}, X^3 = x^{r(3)}, \ldots \mid I)\\ &\qquad\qquad \text{for all permutations } r \text{ and all allowed values } y, x. \qquad (2)\end{aligned}$$

If the number of units is large enough, de Finetti's theorem[2] proves that such a joint probability distribution must then have the following mathematical form:

$$\begin{aligned}&\mathrm{p}(Y^1 = y^1, X^1 = x^1,\ Y^2 = y^2, X^2 = x^2, \ldots \mid I) =\\ &\qquad\qquad \int F(y^1, x^1)\, F(y^2, x^2) \cdots \mathrm{p}(\mathrm{d}F \mid I) \qquad (3)\end{aligned}$$

the integral being over all normalized, positive weight distributions $F(y, x)$ over the variables y, x. The term $\mathrm{p}(\mathrm{d}F \mid I)$ is a positive measure over such distributions.

This mathematical representation has an intuitive interpretation. The weight distribution $F(y, x)$ can be interpreted as the frequency distribution of the values y, x among all units. If we knew this frequency distribution, then by symmetry we would assign the probability $F(y, x)$ to measuring y, x on an unsystematically chosen unit, exactly as in sampling from an urn[3]. Observations on further units would have probabilities according to "sampling without replacement"; however, if the total number of units is large and the observed units are few, these probabilities

[2] For a detailed account with references see Bernardo & Smith 2000; for a shorter summary, Dawid 2013; important references and variations are in Johnson 1924 (Appendix on eduction); 1932 §§ 4.2–4.3; de Finetti 1930; 1937; Hewitt & Savage 1955; Ericson 1969a; Heath & Sudderth 1976; Diaconis 1977; Diaconis & Freedman 1980a,b; Lindley & Novick 1981; Lindley 2014 especially around §§ 7.3, 8.6. [3] Jaynes 2003 ch. 3.

would approximately equal those for "sampling with replacement"[4]. These probabilities correspond to the term $F(y^1, x^1)\, F(y^2, x^2) \cdots$ in the integral. If the frequency distribution F is unknown and we assign a probability density $\mathrm{p}(\mathrm{d}F \mid I)$ to it, then the resulting probability for the observations is a convex mixture of the sampling-with-replacement probabilities, weighted by this probability density. This last step follows from the law of total probability, a consequence of the rules (0). The integral above corresponds to this convex mixture.

The symmetry or "exchangeability" requirement does not determine specific numeric values for the joint probability (3), but greatly restrict our freedom in assigning such values. The only freedom we have left is in the specification of the density $\mathrm{p}(\mathrm{d}F \mid I)$. This specification will depend on the particular inference problem and on the nature of the units and variates. The intuitive interpretation of $\mathrm{p}(\mathrm{d}F \mid I)$ as our uncertainty about the "long-run frequencies" can greatly help in this specification. De Finetti's theorem has moreover an important consequence in this respect: as the values of more and more units become available, the initial numerical choice of density $\mathrm{p}(\mathrm{d}F \mid I)$ affects further inferences less and less. Let us discuss these further inferences.

Once the joint probability distribution (3) has been numerically assigned, through specification of $\mathrm{p}(\mathrm{d}F \mid I)$, all other marginal and conditional distributions concerning units and variates are also numerically fully determined by repeated application of the rules (0). Some examples:

$$\mathrm{p}(Y^2 = y^2 \mid I) = \int \sum_x F(y^2, x)\, \mathrm{p}(\mathrm{d}F \mid I)\,, \tag{4a}$$

$$\mathrm{p}(Y^2 = y^2 \mid X^2 = x^2, I) = \frac{\int F(y^2, x^2)\, \mathrm{p}(\mathrm{d}F \mid I)}{\int \sum_x F(y^2, x)\, \mathrm{p}(\mathrm{d}F \mid I)}\,, \tag{4b}$$

$$\mathrm{p}(Y^2 = y^2 \mid X^1 = x^1, I) = \frac{\int \sum_x F(y^2, x) \sum_y F(y, x^1)\, \mathrm{p}(\mathrm{d}F \mid I)}{\int \sum_y F(y, x^1)\, \mathrm{p}(\mathrm{d}F \mid I)}\,. \tag{4c}$$

3 Inferences about unit 0

Any inferences about any variates of any units for which the joint probability distribution (3) is defined can now be probabilistically answered.

[4] Heath & Sudderth 1976; for error bounds see Diaconis & Freedman 1980a.

Let us consider inferences about a unit, with index $i = 0$, about which we have some additional information I^0.

If I^0 amounts to saying that the symmetry underlying the probabilities (3) also includes unit 0, then any probabilities are uniquely determined, as in the examples (4). For instance, our main probability (1) is given by

$$\mathrm{p}(Y^0 = y^0 \mid X^0 = x^0,\ Y^1 = y^1, X^1 = x^1, \dots,\ I^0, I) = \frac{\int F(y^0, x^0)\, F(y^1, x^1) \cdots\ \mathrm{p}(\mathrm{d}F \mid I)}{\int \sum_y F(y, x^0)\, F(y^1, x^1) \cdots\ \mathrm{p}(\mathrm{d}F \mid I)}\,. \tag{5}$$

A more interesting kind of additional information I^0_{c} may amount to saying that there is an informational symmetry involving unit 0 an the other units, but only with regard to the Y variates, and only with units sharing the same values x^* of the X variates. Using a different terminology, this would be a symmetry restricted to a subpopulation of units which is defined by particular X values. Mathematically,

$$\begin{aligned}&\mathrm{p}(Y^0 = y^0, Y^1 = y^1, Y^2 = y^2 \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{c}}, I) = \\ &\mathrm{p}(Y^0 = y^{r(0)}, Y^1 = y^{r(1)}, Y^2 = y^{r(2)} \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{c}}, I) \\ &\qquad\qquad \text{for all permutations } r \text{ and all allowed values } y, x.\end{aligned} \tag{6}$$

If the number of units is large enough, a variant of de Finetti's theorem[5] proves that such a joint probability distribution must then have the following mathematical form:

$$\mathrm{p}(Y^0 = y^0, Y^1 = y^1, Y^2 = y^2 \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{c}}, I) = \int \frac{F(y^0, x^*)}{\sum_y F(y, x^*)}\, \frac{F(y^1, x^*)}{\sum_y F(y, x^*)}\, \frac{F(y^2, x^*)}{\sum_y F(y, x^*)} \cdots\ \mathrm{p}(\mathrm{d}F \mid I)\,. \tag{7a}$$

This representation has an intuitive interpretation similar to (3); the difference is that a term like $\frac{F(y^0, x^*)}{\sum_y F(y, x^*)}$ now corresponds to the long-run *conditional* frequency $F(y^0 \mid x^*)$. Our conditional probabilities would be equal to these conditional frequencies if they were known. The uncertainty about the latter leads again to a convex mixture of these

[5] Bernardo & Smith 2000 § 4.6; their presentation is in terms of the integral (7b).

conditional probabilities, from the rules (0). With a change of variables we could represent the joint frequency $F(y, x)$ as the pair of conditional and marginal frequencies $\big(F_{Y|X}(y \mid x), F_X(x)\big)$. The formula above then simplifies, by marginalization over $\mathrm{d}F_X$, to

$$\mathrm{p}(Y^0 = y^0, Y^1 = y^1, Y^2 = y^2 \mid X^0 = x^*, X^1 = x^*, X^2 = x^*, \dots,\ I^0_{\mathrm{c}}, I) = \int F(y^0 \mid x^*)\, F(y^1 \mid x^*)\, F(y^2 \mid x^*) \cdots \mathrm{p}(\mathrm{d}F_{Y|X} \mid I)\,. \quad (7b)$$

Under this kind of auxiliary information I^0_{c}, our main probability (1) is given by

$$\mathrm{p}(Y^0 = y^0 \mid X^0 = x^0,\ Y^1 = y^1, X^1 = x^1, \dots,\ I^0_{\mathrm{c}}, I) = \frac{\int \frac{F(y^0,x^0)}{\sum_y F(y,x^0)}\, F(y^1, x^1) \cdots \mathrm{p}(\mathrm{d}F \mid I)}{\int F(y^1, x^1) \cdots \mathrm{p}(\mathrm{d}F \mid I)}\,. \quad (8)$$

Bibliography

("de X" is listed under D, "van X" under V, and so on, regardless of national conventions.)

Alvarez-Melis, D., Broderick, T. (2015): *A translation of "The characteristic function of a random phenomenon" by Bruno de Finetti*. arXiv DOI:10.48550/arXiv.1512.01229. Transl. of de Finetti (1929).

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). DOI:10.1002/9780470316870. First publ. 1994.

Boričić, M. (2020): *Probabilized sequent calculus and natural deduction system for classical logic*. In: Ognjanović (2020): ch. 7:197–213. DOI:10.1007/978-3-030-52954-3_7.

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford). DOI:10.1093/acprof:oso/9780199695607.001.0001.

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013): ch. 2:19–29. DOI:10.1093/acprof:oso/9780199695607.003.0002.

de Finetti, B. (1929): *Funzione caratteristica di un fenomeno aleatorio*. In: *atti del congresso internazionale dei matematici:* ed. by S. Pincherle (Zanichelli, Bologna): 179–190. https://www.mathunion.org/icm/proceedings, http://www.brunodefinetti.it/Opere.htm. Transl. in Alvarez-Melis, Broderick (2015). See also de Finetti (1930).

— (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**^{5}, 86–133. http://www.brunodefinetti.it/Opere.htm. Summary in de Finetti (1929).

— (1937): *La prévision: ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré **7**^{1}, 1–68. http://www.numdam.org/item/AIHP_1937__7_1_1_0. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.

Diaconis, P. (1977): *Finite forms of de Finetti's theorem on exchangeability*. Synthese **36**^{2}, 271–281. doi:10.1007/BF00486116, http://statweb.stanford.edu/~cgates/PERSI/year.html.

Diaconis, P., Freedman, D. (1980a): *Finite exchangeable sequences*. Ann. Probab. **8**^{4}, 745–764. doi:10.1214/aop/1176994663.

— (1980b): *De Finetti's generalizations of exchangeability*. In: Jeffrey (1980): 233–249.

Ericson, W. A. (1969a): *Subjective Bayesian models in sampling finite populations*. J. R. Stat. Soc. B **31**^{2}, 195–224. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See also discussion in Sampford, Scott, Stone, Lindley, Smith, Kerridge, Godambe, Kish, et al. (1969).

— (1969b): *A note on the posterior mean of a population mean*. J. R. Stat. Soc. B **31**^{2}, 332–334.

Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. Am. Stat. **30**^{4}, 188–189.

Hewitt, E., Savage, L. J. (1955): *Symmetric measures on Cartesian products*. Trans. Am. Math. Soc. **80**^{2}, 470–501. doi:10.1090/S0002-9947-1955-0076206-8.

Huth, M. R. A., Ryan, M. D. (2004): *Logic in Computer Science: Modelling and reasoning about systems*, 2nd ed. (Cambridge University Press, Cambridge). First publ. 2000.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. doi:10.1017/CBO9780511790423, https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffrey, R. C., ed. (1980): *Studies in inductive logic and probability. Vol. II*. (University of California Press, Berkeley).

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03johnn.

— (1932): *Probability: the deductive and inductive problems*. Mind **41**^{164}, 409–423. With some notes and an appendix by R. B. Braithwaite. doi:10.1093/mind/XLI.164.409.

Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.

Lindley, D. V. (2014): *Understanding Uncertainty*, rev. ed. (Wiley, Hoboken, USA). First publ. 2006.

Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. Ann. Stat. **9**^{1}, 45–58. doi:10.1214/aos/1176345331.

Ognjanović, Z., ed. (2020): *Probabilistic Extensions of Various Logical Systems*. (Springer, Cham). doi:10.1007/978-3-030-52954-3.

Prawitz, D. (1965): *Natural Deduction: A Proof-Theoretical Study*. (Almqvist & Wiksell, Stockholm).

Sampford, M. R., Scott, A., Stone, M., Lindley, D. V., Smith, T. M. F., Kerridge, D. F., Godambe, V. P., Kish, L., et al. (1969): *Discussion on professor Ericson's paper*. J. R. Stat. Soc. B **31**^{2}, 224–233. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See Ericson (1969b).