

Sufficient Statistics and Split Idempotents in Discrete Probability Theory

Bart Jacobs^{1,2}

*Institute for Computing and Information Sciences (iCIS)
Radboud University
Nijmegen, The Netherlands.*

Abstract

A sufficient statistic is a deterministic function that captures an essential property of a probabilistic function (channel, kernel). Being a sufficient statistic can be expressed nicely in terms of string diagrams, as Tobias Fritz showed recently, in adjoint form. This reformulation highlights the role of split idempotents, in the Fisher-Neyman factorisation theorem. Examples of a sufficient statistic occur in the literature, but mostly in continuous probability. This paper demonstrates that there are also several fundamental examples of a sufficient statistic in discrete probability. They emerge after some combinatorial groundwork that reveals the relevant dagger split idempotents and shows that a sufficient statistic is a deterministic dagger epi.

Keywords: Discrete probability theory, sufficient statistic, disintegration, string diagram, split idempotent

1 Introduction

The notion of a *sufficient statistic* plays an important role in statistics, but a precise definition is hard to find. Informally, it involves a function $s: X \rightarrow Y$ that expresses a characteristic property about certain parameterised probability distributions $p(\theta)$ on X . The statistic provides enough information such that this distribution can be reconstructed from the push-forward of $p(\theta)$ along s . The original formulation has been introduced and studied in the 1920s by Ronald Fisher, see [10].

A precise formulation has been given recently by Tobias Fritz as part of his efforts to express basic concepts and results from probability theory and statistics categorically, see [11] and also [12,13]. This approach uses *Markov categories*, which are symmetric monoidal categories, with a final object as tensor unit, and with copiers (studied as ‘CD-categories’ in [5]). String diagrams provide a powerful and intuitive language for these Markov categories.

Within this setting, the notion of sufficient statistic is formalised as an equation of string diagrams, see [11, Defn. 14.3]. It can be read as a (symmetric) adjoint property, more in the style of adjoint matrices than adjoint functors. One of the highlights in [11, Thm. 14.5] is an abstract formalisation (and proof) of the Fisher-Neyman factorisation theorem (see also *e.g.* [2, Prop 4.10] or [31, §3.3]). It gives a necessary

¹ Thanks are due to Dario Stein and Dusko Pavlovic for lively discussions.

² Email: bart@cs.ru.nl

and sufficient condition for the existence of a sufficient statistic, in terms of a split idempotent. Thus, [11] provides a great clarification of what a sufficient statistic is all about, via abstraction. But it does not really deal with examples of sufficient statistics.

The existing descriptions of sufficient statistics in the literature have several disadvantages.

- The examples are mostly in continuous probability theory. There are hardly any illustrations in discrete probability theory, except the one with sums and Poisson distributions that we reproduce in Section 7. This is a pity, because there are several fundamental discrete instances.
- Typically in these examples, only the sufficiency condition of the Fisher-Neyman factorisation theorem is established — consisting of the disappearance of the parameter after conditioning. These descriptions stop short of giving the full picture, with the relevant split idempotent that does the real work and produces the adjoint situation.
- These examples do not match the abstract description of [11] — understandably so, given that [11] is a very recent publication.

Thus there is room for a fresh look at sufficiency of statistics, given the clear and abstract reformulation provided recently by [11]. That is what the current paper will do. The emphasis is on discrete probability theory, since it already offers ample material. This restriction has the (technical) benefit that we do not have to bother with “almost surely equal”, as in continuous probability. Thus, the paper only looks at one particular kind of Markov category, namely the Kleisli category of the discrete distribution monad.

This paper elaborates examples, but it does not offer new (category) theory. We like to compare examples of sufficient statistics with examples of adjunctions, like: the forgetful functor from compact Hausdorff spaces to sets has a left adjoint given by ultrafilters. Proving this is quite a bit of work. It offers valuable mathematical insight.

Similarly, we claim that the examples of sufficient statistics that we describe below offer valuable insight in the relevant mathematical structures. As is often the case in discrete probability theory, the nature of these structures is combinatorial. Indeed, this paper develops some new combinatorial results, especially about (multiset) partitions, see *e.g.* Proposition 2.4 below (extending [20]). The emphasis is on uncovering the relevant split idempotents.

We recall that a split idempotent is a map $f: A \rightarrow A$ that can be written as $f = s \circ r$, where $r \circ s = \text{id}$. In such a situation, the map r is called a retraction and s a section. It is easy to see that the section s is the equaliser of $f, \text{id}: A \rightrightarrows A$, and that the retraction r is their coequaliser. Such a splitting of f , if it exists is unique, up-to-isomorphism. Split idempotents are used in the Karoubi envelope, the idempotent-splitting completion of a category. Splittings of dagger (self-adjoint) idempotents form classical objects in a quantum setting [29]. Here we also have daggers, relative to a prior distribution. We show that a sufficient statistic (in discrete probability) is a deterministic dagger epi (see [14]), as retraction part of a split dagger idempotent, see Lemma 2.3.

The paper starts with a relatively long section on background material, covering multisets, discrete probability distributions, channels and their daggers, updating, and partitions. These partitions are special multisets over the natural numbers that have been studied in their own right [1], but also to capture mutations in population biology (see *e.g.* [8,9,25,26,27,32]; we offer some new result. Section 3 repeats the description of sufficient statistics from [11] and puts it in perspective, in relation to disintegration. The role of split idempotents is emphasised to obtain examples of sufficient statistics. The subsequent four sections 4 – 7 cover particular illustrations of the notion of sufficient statistic. Section 4 looks at accumulation (from sequences to multisets) as sufficient statistic for independent and identically distributed (iid) elements in sequences. This sufficiency situation, in Diagram 16, captures the very basic relations at the heart of discrete probability theory. Section 5 introduces a new sufficiency situation for partitions, building on the multiplicity count function (from [20]). Sections 6 and 7 review two examples from the literature in the current setting, with the adjoint description based on split idempotents.

2 Background

We briefly review the essence of multisets, distributions and partitions, and also of channels as probabilistic functions. The main goal is to fix notation.

2.1 Multisets

A multiset (or bag) is a finite ‘subset’ in which elements may occur multiple times. There are two equivalent representations for a multiset with elements from a set X .

- As formal finite sums $\sum_i n_i |x_i\rangle$ of elements $x_i \in X$ with multiplicity $n_i \in \mathbb{N}$.
- As functions $\varphi: X \rightarrow \mathbb{N}$ with finite support $\text{supp}(\varphi) := \{x \in X \mid \varphi(x) > 0\}$.

We switch between these representations whenever convenient.

We write $\mathcal{M}(X)$ for the set of multisets over X . It is the free commutative monoid on X . This \mathcal{M} forms a monad on the category of sets. We only mention how functoriality works: for a function $f: X \rightarrow Y$ there is a function $\mathcal{M}(f): \mathcal{M}(X) \rightarrow \mathcal{M}(Y)$ defined by:

$$\mathcal{M}(f)\left(\sum_i n_i |x_i\rangle\right) := \sum_i n_i |f(x_i)\rangle \quad \text{or, equivalently, } \mathcal{M}(f)(\varphi) = \sum_{y \in Y} \left(\sum_{x \in f^{-1}(y)} \varphi(x) \right) |y\rangle. \quad (1)$$

With a multiset φ we associate the following numbers.

- $\|\varphi\| := \sum_x \varphi(x)$ is the size of φ : the total number of elements in φ , including multiplicities;
- $\varphi\!|\!| := \prod_x \varphi(x)!$ is the product of the factorials of multiplicities;
- $\binom{\varphi}{\varphi} := \frac{\|\varphi\|!}{\varphi\!|\!|} = \frac{(\sum_x \varphi(x))!}{\prod_x \varphi(x)!}$ is the multinomial coefficient;
- $\binom{n}{\varphi} := \frac{n!}{\varphi\!|\!| \cdot (n - \|\varphi\|)!}$ when $n \geq \|\varphi\|$.

For a number K we define the subset $\mathcal{M}[K](X) \subseteq \mathcal{M}(X)$ of multisets of size K :

$$\mathcal{M}[K](X) := \{\varphi \in \mathcal{M}(X) \mid \|\varphi\| = K\}.$$

There is an ‘accumulation’ function $\text{acc}: X^K \rightarrow \mathcal{M}[K](X)$ that accumulates multiple occurrences in a sequence: $\text{acc}(x_1, \dots, x_K) = 1|x_1\rangle + \dots + 1|x_K\rangle$. For instance $\text{acc}(a, a, b, a, c, b) = 3|a\rangle + 2|b\rangle + 1|c\rangle$.

For a finite set A we write $|A| \in \mathbb{N}$ for its number of elements and $\text{Perm}(A)$ for the set of permutations $\pi: A \xrightarrow{\cong} A$. Recall that $|\text{Perm}(A)| = |A|!$.

Lemma 2.1 (i) For a multiset $\varphi \in \mathcal{M}[K](X)$ there are $\binom{\varphi}{\varphi}$ many sequences $\ell \in X^K$ with $\text{acc}(\ell) = \varphi$, that is, $|\text{acc}^{-1}(\varphi)| = \binom{\varphi}{\varphi}$.

(ii) If X is finite, then $\sum_{\varphi \in \mathcal{M}[K](X)} \binom{\varphi}{\varphi} = |X|^K$. □

2.2 Distributions

A (discrete finite probability) distribution over a set X is a formal sum $\sum_i r_i |x_i\rangle$, where $x_i \in X$ and $r_i \in [0, 1]$ with $\sum_i r_i = 1$. Alternatively, it is a function $\omega: X \rightarrow [0, 1]$ with finite support and $\sum_x \omega(x) = 1$. We write $\mathcal{D}(X)$ for the set of distributions over X . Such distributions may also be called states, see [16]. This \mathcal{D} is a monad too, like \mathcal{M} . Functoriality works as in (1).

For two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ one can form a product distribution $\omega \otimes \rho \in \mathcal{D}(X \times Y)$ via $(\omega \otimes \rho)(x, y) = \omega(x) \cdot \rho(y)$. Equivalently, $\omega \otimes \rho = \sum_{x, y} \omega(x) \cdot \rho(y) |x, y\rangle$. For a single distribution

$\omega \in \mathcal{D}(X)$ we write the K -fold (tensor) product as:

$$\text{iid}[K](\omega) := \omega^K = \omega \otimes \cdots \otimes \omega \in \mathcal{D}(X^K). \quad (2)$$

The abbreviation ‘iid’ stands for ‘independent and identically distributed’. Marginalisation of a ‘joint’ state $\tau \in \mathcal{D}(X_1 \times \cdots \times X_n)$ is the operation of projecting τ to a state in $\mathcal{D}(X_i)$, on one of the components, obtained via functoriality, as $\mathcal{D}(\pi_i)(\tau)$.

We shall make frequent use of multinomial distributions, which we describe via a function $mn[K]: \mathcal{D}(X) \rightarrow \mathcal{D}(\mathcal{M}[K](X))$. We think of X as a set of colours and of $\omega \in \mathcal{D}(X)$ as an urn with coloured balls, where $\omega(x) \in [0, 1]$ is the probability of drawing a ball of colour $x \in X$. Then, $mn[K](\omega) \in \mathcal{D}(\mathcal{M}[K](X))$ is a distribution on multisets of size K , corresponding to draws of K -many balls from the urn (with replacement). This distribution is given by:

$$mn[K](\omega) := \mathcal{D}(\text{acc})\left(\text{iid}[K](\omega)\right) = \sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} | \varphi \rangle. \quad (3)$$

See [18,23] for more details. As illustration, we consider draws of size 3 with a distribution $\frac{1}{8}|a\rangle + \frac{1}{2}|b\rangle + \frac{3}{8}|c\rangle$, corresponding to an urn with 8 balls, one of colour a , 4 of colour b , 3 of colour c . These draws (of size 3) then come with the following distribution.

$$\begin{aligned} & mn[3]\left(\frac{1}{8}|a\rangle + \frac{1}{2}|b\rangle + \frac{3}{8}|c\rangle\right) \\ &= \frac{1}{512} |3|a\rangle\rangle + \frac{3}{128} |2|a\rangle + 1|b\rangle\rangle + \frac{3}{32} |1|a\rangle + 2|b\rangle\rangle + \frac{1}{8} |3|b\rangle\rangle + \frac{9}{512} |2|a\rangle + 1|c\rangle\rangle + \\ & \quad \frac{9}{64} |1|a\rangle + 1|b\rangle + 1|c\rangle\rangle + \frac{9}{32} |2|b\rangle + 1|c\rangle\rangle + \frac{27}{512} |1|a\rangle + 2|c\rangle\rangle + \frac{27}{128} |1|b\rangle + 2|c\rangle\rangle + \frac{27}{512} |3|c\rangle\rangle \end{aligned}$$

In the previous subsection we have seen the accumulation *function* $\text{acc}: X^K \rightarrow \mathcal{M}[K](X)$. In the other direction there is a *channel* $\text{arr}: \mathcal{M}[K](X) \rightarrow X^K$, called arrangement (see [18]); it is given by:

$$\text{arr}(\varphi) := \sum_{\mathbf{x} \in \text{acc}^{-1}(\varphi)} \frac{1}{(\varphi)} | \mathbf{x} \rangle. \quad (4)$$

This uniform distribution is well-defined, by Lemma 2.1 (i). We shall see that accumulation and arrangement form a split idempotent, for a special ‘Kleisli’ composition, see Lemma 2.2.

We introduced distributions $\omega \in \mathcal{D}(X)$ with *finite* support. In Section 7 we drop this finiteness requirement and shall work with more general discrete probability distributions. We then write $\mathcal{D}_\infty(X)$ for the set of such distributions.

2.3 Channels

We have mentioned that taking distributions \mathcal{D} is a monad. The associated ‘Kleisli’ maps $X \rightarrow \mathcal{D}(Y)$ are called *channels* and will be written as $X \rightarrowtail Y$, with a circle on the shaft. These arrows \rightarrowtail are the morphisms in the (symmetric monoidal) Kleisli category $\mathcal{Kl}(\mathcal{D})$ of the distribution monad \mathcal{D} . We recall the basic operations, of Kleisli extension $\gg=$ and Kleisli composition \circ .

Given a channel $c: X \rightarrowtail Y$ and a distribution / state $\omega \in \mathcal{D}(X)$ we write $c \gg= \omega \in \mathcal{D}(Y)$ for the *state transformation* along c , obtained via Kleisli extension:

$$(c \gg= \omega)(y) := \sum_{x \in X} \omega(x) \cdot c(x)(y).$$

Via \gg we can define (Kleisli) composition \circ of channels as $(d \circ c)(x) = d \gg c(x)$. The unit $\eta: X \rightarrow \mathcal{D}(X)$ of the monad \mathcal{D} , with $\eta(x) = 1|x\rangle$, is the identity (channel) for this composition \circ .

For two channels $c: X \rightarrow A$ and $d: Y \rightarrow B$ there is a parallel product $c \otimes d: X \times Y \rightarrow A \times B$ defined as pointwise tensor: $(c \otimes d)(x, y) = c(x) \otimes d(y)$. When the domains X, Y are equal we can form a tuple channel $\langle c, d \rangle := (c \otimes d) \circ \Delta: X \rightarrow A \times B$, where $\Delta: X \rightarrow X \times X$ is the copy function $\Delta(x) = (x, x)$. We shall see special instances of this tupling of channels when c or d is the identity channel. We shall write such tuples as $\langle c, \text{id} \rangle$ or $\langle \text{id}, d \rangle$. We spell out the associated state transformation:

$$\langle c, \text{id} \rangle \gg \omega = \sum_{x \in X, y \in Y} \omega(x) \cdot c(x)(y) |x, y\rangle. \quad (5)$$

Each function $f: X \rightarrow Y$ gives rise to a *deterministic* channel $\langle f \rangle := \eta \circ f: X \rightarrow Y$, via the unit η of the monad, so that $\langle f \rangle(x) = 1|f(x)\rangle$. These deterministic channels, of the form $\langle f \rangle$, can be characterised as those channels that commute with copyiers Δ . They satisfy: $\langle g \rangle \circ \langle f \rangle = \langle g \circ f \rangle$ and $d \circ \langle f \rangle = d \circ f$ and $\langle f \rangle \circ c = \mathcal{D}(f) \circ c$. Often we omit the brackets $\langle - \rangle$ when it is implicitly clear that an ordinary function is promoted to a channel. We do so in particular for projection and copy functions π_i and Δ , and also for accumulation in the next result.

Lemma 2.2 *Accumulation and arrangement (4) form (the section and retraction of) a split idempotent in the Kleisli category $\mathcal{Kl}(\mathcal{D})$ of the distribution monad \mathcal{D} , since: $\text{acc} \circ \text{arr} = \text{id}$. The resulting split idempotent channel $\text{tp} := \text{arr} \circ \text{acc}: X^K \rightarrow X^K$ is tuple-permutation (“transposition”):*

$$\text{tp}(x_1, \dots, x_K) = \sum_{\pi \in \text{Perm}(\{1, \dots, K\})} \frac{1}{K!} |x_{\pi(1)}, \dots, x_{\pi(K)}\rangle. \quad \square$$

For convenience and clarity, we shall use the graphical notation of string diagrams for channels, as probabilistic computations, where information is flowing upwards. A channel is represented as a box, and wires are typed by the input and output sets of these channels. We assume a basic level of familiarity with these string diagrams and refer to [5,11,30] for more information about their use in probability theory. In this paper these string diagram are interpreted in the Kleisli category $\mathcal{Kl}(\mathcal{D})$, or $\mathcal{Kl}(\mathcal{D}_\infty)$, but not in general Markov categories.

2.4 Predicates, updating and daggers

A predicate on a set X is a function $p: X \rightarrow [0, 1]$. We shall write $\text{Pred}(X)$ for the set of all such (fuzzy) predicates on X . Updating (or conditioning) of a state/distribution is a basic operation in probabilistic reasoning, where $\omega \in \mathcal{D}(X)$ is updated in the light of evidence given by a predicate $p \in \text{Pred}(X)$. We first write $\omega \models p := \sum_x \omega(x) \cdot p(x)$ for the validity (expected value) of p in ω . If this validity is non-zero, we can define the updated distribution $\omega|_p \in \mathcal{D}(X)$ as the normalised product:

$$\omega|_p(x) := \frac{\omega(x) \cdot p(x)}{\omega \models p}.$$

One can then prove that updating makes the predicate p ‘more true’, in the sense that $\omega|_p \models p \geq \omega \models p$, see e.g. [17,18] for details.

Along a channel $c: X \rightarrow Y$ one can do backwards predicate transformation: for a predicate $q: Y \rightarrow [0, 1]$ on Y we can form a predicate $c \ll q$ on X via:

$$(c \ll q)(x) := \sum_{y \in Y} c(x)(y) \cdot q(y).$$

There is then a basic equality of validities: $c \gg \omega \models q = \omega \models c \ll q$.

Given a channel $c: X \rightarrow Y$ and a ‘prior’ state $\omega \in \mathcal{D}(Y)$ we can obtain a reversed channel $c_\omega^\dagger: Y \rightarrow X$, with ω as prior distribution, called the *dagger* of c , see [7,5,11]. Its formulation is:

$$c_\omega^\dagger(y) := \omega|_{c \ll \mathbf{1}_y} = \sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{(c \gg \omega)(y)} |x\rangle. \quad (6)$$

This definition uses the point predicate $\mathbf{1}_y: Y \rightarrow [0, 1]$ which is 1 on $y \in Y$ and 0 everywhere else. The definition only works when the pushforward distribution $c \gg \omega$ has full support. The dagger is also called Bayesian inversion; it corresponds to turning a conditional probability $p(y|x)$ into $p(x|y)$.

Daggers of *deterministic* channels play a special role; they give rise to dagger idempotents.

Lemma 2.3 *Let $f: X \rightarrow Y$ be a function, considered as a (deterministic) channel, and let $\omega \in \mathcal{D}(X)$ be a distribution such that $f \gg \omega = \mathcal{D}(f)(\omega)$ has full support. Then:*

$$f_\omega^\dagger \gg \mathcal{D}(f)(\omega) = \omega \quad f \circ f_\omega^\dagger = \text{id} \quad (f_\omega^\dagger \circ f)_\omega^\dagger = f_\omega^\dagger \circ f.$$

This makes f a deterministic dagger epi, with prior ω , and $f_\omega^\dagger \circ f$ a split dagger idempotent, in $\mathcal{Kl}(\mathcal{D})$.

Proof. We only do the middle equation and leave the other equations to the reader.

$$(f \circ f_\omega^\dagger)(y) = \sum_{x \in X} f_\omega^\dagger(y)(x) |f(x)\rangle \stackrel{(6)}{=} \sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)} |f(x)\rangle = \frac{\sum_{x \in f^{-1}(y)} \omega(x)}{\mathcal{D}(f)(\omega)(y)} |y\rangle = 1|y\rangle. \quad \square$$

2.5 Multiset partitions

A *multiset partition*, or simply a *partition*³, is a multiset over the non-negative natural numbers $\mathbb{N}_{>0}$. Such a partition $\sum_i n_i |i\rangle \in \mathcal{M}(\mathbb{N}_{>0})$ has a *sum* (or mean, or average) defined as:

$$\text{sum}\left(\sum_i n_i |i\rangle\right) := \sum_i n_i \cdot i \in \mathbb{N}_{>0} \quad \text{and used in: } MP(K) := \{\varphi \in \mathcal{M}(\mathbb{N}_{>0}) \mid \text{sum}(\varphi) = K\}.$$

Thus, $MP(K)$ is the set of partitions with sum K . The sizes $|MP(K)|$ of the sets of partitions with sum $K = 1, 2, 3, \dots$ are given by the so-called partition numbers [1]: 1, 2, 3, 5, 7, 11, 15, 22, 30, 42, ... for which no closed-form expression is known. For instance, $MP(4)$ contains the following five multisets.

$$4|1\rangle \quad 2|1\rangle + 1|2\rangle \quad 1|1\rangle + 1|3\rangle \quad 2|2\rangle \quad 1|4\rangle. \quad (7)$$

These partitions can be seen as the different ways to describe the sum 4 via coins: 4 coins with value 1, 2 coins of 1 and 1 of 2, *etc*, see [20]. Such partitions are studied for instance in population biology, see *e.g.* [8,9,25,26,27] and also Section 6. In economics they can be used for (un)fairness [28]: if you have 4 units of ‘wealth’ that you can distribute over 4 people, you can give each of them one unit (as on left in (7)) or you can give all units to one individual (as on the right in (7)). The partitions in the middle are intermediate forms of fairness. Notice that the number $\text{sum}(\varphi)$, for $\varphi \in \mathcal{M}(\mathbb{N}_{>0})$, is typically different from its size $\|\varphi\|$. We do have $\|\varphi\| \leq \text{sum}(\varphi)$. For instance, the partitions in (7) all have sum four, but their sizes are, in order, 4, 3, 2, 2, 1.

The next result involves a *multiplicity count* function mc , introduced in [20] (and used also in [19,24]), that plays a fundamental role in the sequel — as analogue of accumulation *acc*. It counts the multiplicities

³ Commonly, the word ‘partition’ is used for ‘cover’: a collection of pairwise disjoint subsets whose union is the whole set. This may be confusing. We use the phrases ‘multiset partition’ and ‘set partition’ (cover) to distinguish them in [19,24]. Here, only multiset partitions are used, so we often simply use ‘partition’ for ‘multiset partition’.

in a multiset. For instance,

$$mc\left(3|a\rangle + 1|b\rangle + 3|c\rangle + 2|d\rangle\right) = 1|1\rangle + 1|2\rangle + 2|3\rangle.$$

This expresses that in the above (argument) multiset over $\{a, b, c, d\}$, 1 element occurs 1 time (namely b), 1 element occurs 2 times (namely d) and 2 elements occur 3 times (namely a and c). Thus, the multiplicity count function abstracts away from the elements in a multiset and only looks at how many elements occur how many times. It is a dagger epi, see Section 5.

The next result collects some combinatorial results about the relation between multisets and partitions, via this new multiplicity count function mc . The focus is on describing mc as the retraction part of a split idempotent.

Proposition 2.4 *Let X be a finite set with at least $n \geq K$ elements, that is, with $n := |X| \geq K$.*

- (i) *Define for a multiset $\varphi \in \mathcal{M}(X)$, its multiplicity count as: $mc(\varphi) := \sum_{x \in \text{supp}(\varphi)} 1|\varphi(x)\rangle$. This function restricts to $mc: \mathcal{M}[K](X) \rightarrow MP(K)$.*

- (ii) *For each $\sigma \in MP(K)$ there are*

$$\binom{n}{\sigma} = \frac{n!}{\sigma_{\mathbb{0}} \cdot (n - \|\sigma\|)!} \quad \text{many multisets } \varphi \in \mathcal{M}[K](X) \text{ with } mc(\varphi) = \sigma.$$

This means: $|\mathcal{M}[K](X) \cap mc^{-1}(\sigma)| = \binom{n}{\sigma}$.

- (iii) *Thus, we can define a stack channel $stk: MP(K) \rightarrow \mathcal{M}[K](X)$ via uniform distributions as:*

$$stk(\sigma) := \sum_{\varphi \in mc^{-1}(\sigma)} \frac{1}{\binom{n}{\sigma}} |\varphi\rangle. \quad (8)$$

Then: $mc \circ stk = \text{id}: MP(K) \rightarrow MP(K)$.

- (iv) *Write $ep := stk \circ mc: \mathcal{M}[K](X) \rightarrow \mathcal{M}[K](X)$ for the resulting split idempotent in $\mathcal{Kl}(\mathcal{D})$, where ep stands for element permutation (“substitution”). It satisfies, for $\varphi \in \mathcal{M}[K](X)$,*

$$ep(\varphi) = \sum_{\psi \in mc^{-1}(mc(\varphi))} \frac{1}{\binom{n}{mc(\varphi)}} |\psi\rangle = \sum_{\pi \in \text{Perm}(X)} \frac{1}{n!} |\mathcal{M}(\pi)(\varphi)\rangle.$$

The first equation holds by definition; the second one is the statement.

Proof. (i) Straightforward.

- (ii) By assumption, the set X has n elements. The question is: how many ways are there to partition this set X into subsets with $\sigma(1), \sigma(2), \dots, \sigma(K)$ elements, and finally, with $n - \|\sigma\|$ many remaining elements? As we have seen in Lemma 2.1 (i) this number of partitions is given by the multinomial coefficient:

$$\binom{n}{\sigma(1), \dots, \sigma(K), n - \|\sigma\|} = \frac{n!}{\sigma(1)! \cdot \dots \cdot \sigma(K)! \cdot (n - \|\sigma\|)!} = \frac{n!}{\sigma_{\mathbb{0}} \cdot (n - \|\sigma\|)!} = \binom{n}{\sigma}.$$

- (iii) For a partition $\sigma \in MP(K)$ with sum K ,

$$(mc \circ stk)(\sigma) \stackrel{(8)}{=} \sum_{\varphi \in mc^{-1}(\sigma)} \frac{1}{\binom{n}{\sigma}} |mc(\varphi)\rangle = \left(\sum_{\varphi \in mc^{-1}(\sigma)} \frac{1}{\binom{n}{\sigma}} \right) |\sigma\rangle = 1|\sigma\rangle.$$

This last equation follows from item (ii).

(iv) For $\varphi \in \mathcal{M}[K](X)$ write $\sigma = mc(\varphi)$. It suffices to prove:

$$\sum_{\psi \in mc^{-1}(\sigma)} \sigma_{\mathbb{0}} \cdot (n - \|\sigma\|)! |\psi\rangle = \sum_{\pi \in \text{Perm}(X)} 1 \left| \mathcal{M}(\pi)(\varphi) \right\rangle. \tag{*}$$

For a permutation $\pi: X \cong X$ the multiset $\mathcal{M}(\pi)(\varphi)$ obtained by permuting the elements occurring in φ , according to π , also has multiplicity count $\sigma = mc(\varphi)$. Some of these permutations of φ are equal to φ . We have to count their number.

For each $i \in \mathbb{N}$ with $\sigma(i) > 0$ there are $\sigma(i)$ many elements of X in φ occurring i times. Permuting these $\sigma(i)$ many elements does not change the multiset φ . This can be done in $\sigma(i)!$ many ways. Using this argument for each i explains the occurrence of $\sigma_{\mathbb{0}} = \prod_i \sigma(i)!$ in (*). There are $n - \|\sigma\| = |\text{supp}(\varphi)|$ many elements of X that do not occur in φ . They can be permuted in $(n - \|\sigma\|)!$ many ways without changing φ . □

This last item (iv) is somewhat mysterious, so we like to illustrate what is going on.

Example 2.5 Consider a set $X = \{a, b, c, d\}$ with a multiset $\varphi = 2|a\rangle + 1|b\rangle + 1|c\rangle + 1|d\rangle$ of size 5, whose multiplicity count is $\sigma := mc(\varphi) = 3|1\rangle + 1|2\rangle$, with sum 5. The multisets $\psi \in \mathcal{M}[5](X)$ with $mc(\psi) = \sigma$ are:

$$\begin{aligned} 2|a\rangle + 1|b\rangle + 1|c\rangle + 1|d\rangle, & \quad 1|a\rangle + 2|b\rangle + 1|c\rangle + 1|d\rangle, \\ 1|a\rangle + 1|b\rangle + 2|c\rangle + 1|d\rangle, & \quad 1|a\rangle + 1|b\rangle + 1|c\rangle + 2|d\rangle. \end{aligned}$$

There are $4! = 24$ permutations π of X . The ‘swapped’ multiset $\mathcal{M}(\pi)(\varphi)$ is of one of these four forms. Each of these forms occurs 6 times in:

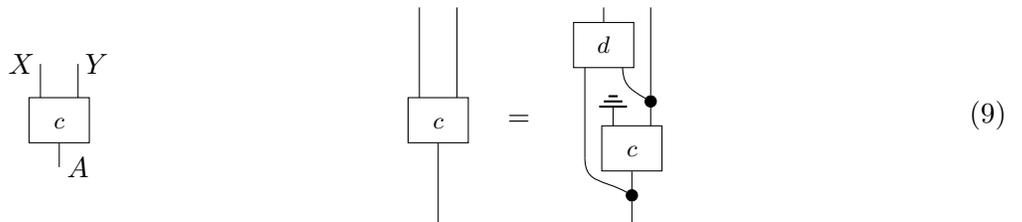
$$\begin{aligned} \sum_{\pi \in \sigma(X)} \frac{1}{24} \left| \mathcal{M}(\pi)(\varphi) \right\rangle &= \frac{1}{4} \left| 2|a\rangle + 1|b\rangle + 1|c\rangle + 1|d\rangle \right\rangle + \frac{1}{4} \left| 1|a\rangle + 2|b\rangle + 1|c\rangle + 1|d\rangle \right\rangle + \\ &\quad \frac{1}{4} \left| 1|a\rangle + 1|b\rangle + 2|c\rangle + 1|d\rangle \right\rangle + \frac{1}{4} \left| 1|a\rangle + 1|b\rangle + 1|c\rangle + 2|d\rangle \right\rangle = ep(\varphi). \end{aligned}$$

This probability $\frac{1}{4}$ equals $\frac{\sigma_{\mathbb{0}} \cdot (|X| - \|\sigma\|)!}{24} = \frac{3! \cdot 1! \cdot 0!}{24}$, as prescribed in Proposition 2.4 (iv).

3 Sufficient statistic

The diagrammatic formulation of the notion of sufficient statistic has been introduced recently in [11]. This formulation involves a simple equation between two diagrams, see (10) below. Implicitly, it relies on disintegration, which we introduce first (based on [11] and also [5]). It helps to understand sufficiency as a form of updating that trivialises, by making a parameter obsolete.

Definition 3.1 Let $c: A \rightarrow X \times Y$ be a channel, represented as a box on the left below. A *disintegration* of c is a channel $d: A \times Y \rightarrow X$ that satisfies the equation on the right.



In this diagram we use $\bar{\top}$ for discarding (marginalisation, projecting away) and \forall for copying. On the right-hand-side of the equation, the X -output of the c channel is discarded, but reconstructed via the d -channel.

We do not worry about uniqueness of disintegration in the current setting. It does exist, under additional requirements, see [11,5] for details.

We use surjective functions $s: X \rightarrow Y$ to identify elements in X , resulting in abstractions in Y . This Y may give numerical information, *e.g.* when Y is the set of natural or real numbers. Such a map is also called a *statistic*. A statistic s is called *sufficient* for a channel, as statistical model depending on a parameter, when the dependence on the parameter disappears via updating through s . Sufficiency refers to being adequate as a summary of essential aspects of the elements in X . Sufficiency involves the existence of a reversal of the *function* $s: X \rightarrow Y$ to a *channel* $d: Y \rightarrow X$. This channel d can be used to reconstruct a distribution X , from its summary given by s . We first give a string diagrammatic description, following [11, Defn. 14.3]. Subsequently we explain the relation to disintegration.

Definition 3.2 Let $c: A \rightarrow X$ be a channel, where we think of A as the space of parameters.

- (i) A *statistic* for the channel c is a function $s: X \rightarrow Y$.
- (ii) Such a statistic s is *sufficient* if there is a channel $d: Y \rightarrow X$ such that:

$$\text{Diagram (10)} \tag{10}$$

Remark 3.3 (i) The channel d on the right-hand-side in (10) results in a special way from disintegration, namely disintegration of the channel $(\text{id}, s) \circ c$ on the left-hand-side in (10). If we write $D: A \times Y \rightarrow X$ for its disintegration channel, as in (9), then, we should get an equation as given on the left below. The distinguishing property of a sufficient statistic is that the dashed line is absent.

$$\text{Diagram (11)} \tag{11}$$

The situation on the left can be expressed by the equation on the right.

- (ii) The fact that the dashed wire is missing in (11) is often expressed as: the conditional distribution $c(a)$, given s , does not depend on a . This can be made more precise. The box/channel D in (11) can be calculated — in discrete probability — as a distribution $D(a, y) \in \mathcal{D}(X)$, for $a \in A$ and $y \in Y$, via a dagger (see Subsection 2.4). Explicitly,

$$D(a, y) = s_{c(a)}^\dagger(y) \stackrel{(6)}{=} c(a)|_{s \ll \mathbf{1}_y} = \sum_{x \in s^{-1}(y)} \frac{c(a)(x)}{\mathcal{D}(s)(c(a))(y)} |x\rangle. \tag{12}$$

The absence of the dashed arrow in (11) corresponds to the non-dependence of the latter expressions on a . This is the essence of the Fisher-Neyman factorisation theorem, see [11, Thm. 14.5] and [2, Prop 4.10] or [31, §3.3].

- (iii) In the sequel we are interested in actually demonstrating sufficiency of certain statistics, in discrete probability. Proving Equation 10 amounts to showing an equality of distributions of the following

form. For each parameter $a \in A$,

$$\begin{aligned} \langle \text{id}, s \rangle \gg c(a) &= \sum_{x \in X} c(a)(x) \mid x, s(x) \rangle \\ &= \sum_{x, z \in X} c(a)(z) \cdot d(s(z))(z) \mid z, s(x) \rangle = \langle d, \text{id} \rangle \gg \mathcal{D}(s)(c(a)). \end{aligned} \tag{13}$$

(iv) Two distributions $\omega, \rho \in \mathcal{D}(X)$ are equal when $\omega \models p = \rho \models p$ for all predicates p on X . This easily follows by just looking at point predicates $\mathbf{1}_x$, where $x \in X$, for which $\omega \models \mathbf{1}_x = \omega(x)$.

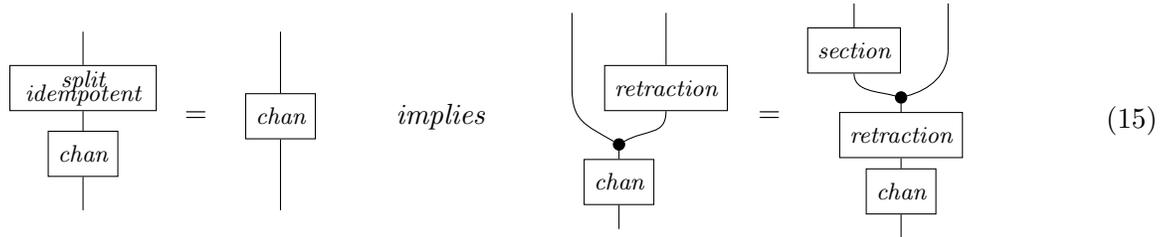
This validity formulation of equality of distributions can also be used for the sufficiency equation (10). It amounts to, for all predicates $p \in \text{Pred}(X)$, $q \in \text{Pred}(Y)$,

$$c(a) \models p \ \& \ (s \ll q) = \mathcal{D}(s)(c(a)) \models (d \ll p) \ \& \ q. \tag{14}$$

This formulation uses conjunction $\&$ for pointwise multiplication of predicates $(p_1 \ \& \ p_2)(x) = p_1(x) \cdot p_2(x)$. It has a clear adjointness flavour. However, there is no direction involved, or left-right distinction, since $\&$ is a commutative operation — in classical probability, not in quantum probability, see *e.g.* [6,15,33].

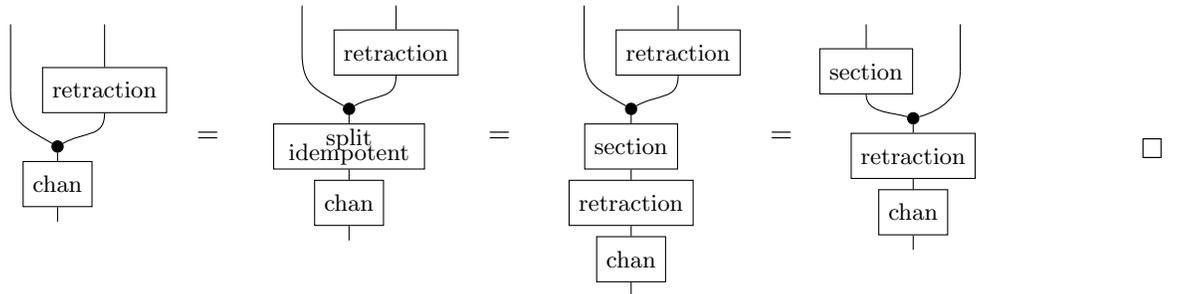
All our examples below follow the following pattern to obtain a sufficient statistic.

Lemma 3.4 *Consider a channel together with a split idempotent in $\mathcal{Kl}(\mathcal{D})$, consisting of a section followed by a retraction. Then:*



The latter equation says that the retraction is a sufficient statistic for the channel.

Proof. This is in essence a special case of the Fisher-Neyman factorisation theorem, as formulated in diagrammatic form in [11, Thm. 14.5]. It uses that $\mathcal{Kl}(\mathcal{D})$ is a positive Markov category, justifying the last equation below, see [11, Rem.11.23].



This result will be exploited in the next four sections.

4 Accumulation as sufficient statistic

We can now start harvesting from the previous two sections.

Theorem 4.1 Fix a set X and a number $K \in \mathbb{N}$. The accumulation function $\text{acc}: X^K \rightarrow \mathcal{M}[K](X)$ is a sufficient statistic for the identically and independent distribution channel $\text{id}[K]: \mathcal{D}(X) \rightarrow X^K$, via arrangement. This is expressed by the following equation between channels $\mathcal{D}(X) \rightarrow X^K \times \mathcal{M}[K](X)$.

$$(16)$$

Proof. Accumulation acc and arrangement arr are the retraction and section part of the split idempotent $tp := \text{arr} \circ \text{acc}$ described in Lemma 2.2. Lemma 3.4 applies since:

$$\begin{aligned} (tp \circ \text{id}[K])(\omega) &= \sum_{\pi \in \text{Perm}(\{1, \dots, K\})} \sum_{\mathbf{x} \in X^K} \frac{\omega^K(\mathbf{x})}{K!} \left| x_{\pi(1)}, \dots, x_{\pi(K)} \right\rangle \\ &= \sum_{\mathbf{x} \in X^K} \left(\sum_{\pi \in \text{Perm}(\{1, \dots, K\})} \frac{\omega^K(x_{\pi(1)}, \dots, x_{\pi(K)})}{K!} \right) \left| \mathbf{x} \right\rangle \\ &= \sum_{\mathbf{x} \in X^K} \omega^K(\mathbf{x}) \left| \mathbf{x} \right\rangle = \omega^K = \text{id}[K](\omega). \end{aligned}$$

The formulation in (16) now follows because, by definition, $\text{mn}[K] = \text{acc} \circ \text{id}[K]$, see (3). \square

We think that Equation (16) expresses a fundamental and elementary relationship between the combinatorial and probabilistic properties of sequences and multisets, in terms of the iid and multinomial distributions. Although these distributions appear frequently in the literature, this (diagrammatic) equation (16) has not been identified before. There is a single source — as far as we are aware — namely [3, §2.2], where it is mentioned that a particular sum — amounting to accumulation — is sufficient as statistic for iid. But the situation is not elaborated further and arrangement does not occur, like in (16).

The reference [18, Thm. 4] contains the following two equations.

$$\text{arr} \circ \text{mn}[K] = \text{id}[K] \quad \text{and} \quad \text{acc} \circ \text{id}[K] = \text{mn}[K].$$

Both can be obtained from the above theorem by discarding one of the outgoing wires in (16). However, Equation (16) expresses more than these two equations obtained by marginalisation, since it involves a joint state formulation.

We elaborate in the concrete situation of Theorem 4.1 on two aspects discussed in Remark 3.3. First, the sufficiency of accumulation can also be expressed (and proven directly) via arbitrary predicates $p \in \text{Pred}(X^K)$ and $q \in \text{Pred}(\mathcal{M}[K](X))$, in the style of (14), as an adjointness property:

$$\begin{aligned} \text{id}[K](\omega) \models p \ \& \ (\text{acc} \ll q) &= \sum_{\mathbf{x} \in X^K} \omega^K(\mathbf{x}) \cdot p(\mathbf{x}) \cdot q(\text{acc}(\mathbf{x})) \\ &= \sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) \cdot \prod_y \omega(x)^{\text{acc}(\mathbf{x})(y)} \cdot \left(\sum_{\mathbf{x} \in \text{acc}^{-1}(\varphi)} \frac{1}{(\varphi)} \cdot p(\mathbf{x}) \right) \cdot q(\varphi) \\ &= \sum_{\varphi \in \mathcal{M}[K](X)} \text{mn}[K](\omega)(\varphi) \cdot (\text{arr} \ll p)(\varphi) \cdot q(\varphi) \\ &= \text{mn}[K](\omega) \models (\text{arr} \ll p) \ \& \ q. \end{aligned}$$

Next, we illustrate how in this case the dependency on ω disappears in iid if we condition on accumulation

channel $mn[K]$ forms a natural transformation $\mathcal{D} \Rightarrow \mathcal{DM}[K]$, see [18].

$$\begin{aligned}
(ep \circ mn[K])(\omega) &= \sum_{\pi \in \text{Perm}(X)} \sum_{\varphi \in \mathcal{M}[K](X)} \frac{mn[K](\omega)(\varphi)}{|X|!} \left| \mathcal{M}(\pi)(\varphi) \right\rangle \quad \text{see Proposition 2.4 (iv)} \\
&= \sum_{\pi \in \text{Perm}(X)} \frac{\mathcal{DM}(\pi)(mn[K](\omega))}{|X|!} \\
&= \sum_{\pi \in \text{Perm}(X)} \frac{mn[K](\mathcal{D}(\pi)(\omega))}{|X|!} \stackrel{(17)}{=} swmn[K](\omega).
\end{aligned}$$

Lemma 3.4 now provides a sufficient statistic situation with channel:

$$mc \circ swmn[K] = mc \circ ep \circ mn[K] = mc \circ stk \circ mc \circ acc \circ iid[K] = mc \circ acc \circ iid[K].$$

This composite is what we have called the partition multinomial $pamn[K]$ in (19). □

We like to get a better handle on this partition multinomial $pamn[K]$.

Proposition 5.2 *Let X be a finite set and let K be a natural number with $1 \leq K \leq |X|$.*

(i) *For $\sigma \in MP(K)$ define a multinomial coefficient for partitions $(\sigma)_P \in \mathbb{N}$ as:*

$$(\sigma)_P := \frac{K!}{\prod_{1 \leq i \leq K} (i!)^{\sigma(i)}}.$$

Then, for each $\varphi \in \mathcal{M}[K](X)$,

$$\varphi_{\mathbb{I}} = \prod_{1 \leq i \leq K} (i!)^{mc(\varphi)(i)} \quad \text{and thus } (\varphi) = (mc(\varphi))_P.$$

(ii) *The partition multinomial can be described concretely, on $\omega \in \mathcal{D}(X)$, as:*

$$pamn[K](\omega) = \sum_{\sigma \in MP(K)} (\sigma)_P \cdot \sum_{\varphi \in mc^{-1}(\sigma)} \prod_{x \in X} \omega(x)^{\varphi(x)} \left| \sigma \right\rangle. \quad (20)$$

Proof. (i) We first prove the equation $\varphi_{\mathbb{I}} = \prod_{1 \leq i \leq K} (i!)^{mc(\varphi)(i)}$, for $\varphi \in \mathcal{M}[K](X)$, by induction on $K \geq 1$. The case $K = 1$ is trivial, since then $\varphi = 1|x\rangle$, for some $x \in X$, so that $mc(\varphi) = 1|1\rangle$. Then: $\varphi_{\mathbb{I}} = 1 = (1!)^1$. For the induction step, write $\varphi \in \mathcal{M}[K+1](X)$ as $\varphi = \psi + 1|x\rangle$, for some $x \in \text{supp}(\varphi)$. Then $\psi \in \mathcal{M}[K](X)$.

$$\begin{aligned}
\prod_{1 \leq i \leq K+1} (i!)^{mc(\varphi)(i)} &= ((K+1)!)^{mc(\psi+1|x)(K+1)} \cdot \prod_{1 \leq i \leq K} (i!)^{mc(\psi+1|x)(i)} \\
&= \begin{cases} (K+1)! & \text{if } \psi = K|x\rangle \\ \prod_{1 \leq i \leq K} (i!)^{mc(\psi)(i)} \cdot (\psi(x)+1) & \text{otherwise} \end{cases} \\
\stackrel{(IH)}{=} &\left\{ \begin{array}{l} (\psi + 1|x)_{\mathbb{I}} \quad \text{if } \psi = K|x\rangle \\ \psi_{\mathbb{I}} \cdot (\psi(x)+1) \quad \text{otherwise} \end{array} \right\} = (\psi + 1|x)_{\mathbb{I}} = \varphi_{\mathbb{I}}.
\end{aligned}$$

For the second equation fix $\varphi \in \mathcal{M}[K](X)$. Then, by what we have just shown:

$$(\varphi) = \frac{K!}{\varphi_{\parallel}} = \frac{K!}{\prod_{1 \leq i \leq K} (i!)^{mc(\varphi)(i)}} = (mc(\varphi))_P.$$

(ii) By the previous item:

$$\begin{aligned} pamn[K](\omega) &\stackrel{(19)}{=} \mathcal{D}(mc) \left(mn[K](\omega) \right) \stackrel{(3)}{=} \sum_{\varphi \in \mathcal{M}[K](X)} (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} |mc(\varphi)\rangle \\ &= \sum_{\sigma \in MP(K)} \left(\sum_{\varphi \in mc^{-1}(\sigma)} (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \right) |\sigma\rangle \\ &= \sum_{\sigma \in MP(K)} (\sigma)_P \cdot \left(\sum_{\varphi \in mc^{-1}(\sigma)} \prod_{x \in X} \omega(x)^{\varphi(x)} \right) |\sigma\rangle. \quad \square \end{aligned}$$

We conclude with an illustration:

$$pamn[3] \left(\frac{1}{8}|a\rangle + \frac{1}{2}|b\rangle + \frac{3}{8}|c\rangle \right) = \frac{9}{64} |3|1\rangle + \frac{87}{128} |1|1\rangle + 1|2\rangle + \frac{23}{128} |1|3\rangle.$$

This channel $pamn[K]$ acts like a (Kingman) paintbox, see [4].

6 Size of partitions as sufficient statistic

We now write $size(\sigma) := \|\sigma\|$, for $\sigma \in MP(K)$, giving a function $size: MP(K) \rightarrow \{1, \dots, K\}$. These sizes vary, for a fixed sum, see (7). We also introduce notation for the product of the numbers in a partition: $prod(\sigma) := \prod_i i^{\sigma(i)}$. For instance, the partitions in (7) have product, in order: 1, 2, 3, 4, 4.

It is known from [9] that the *size* function is a sufficient statistic, namely for the channel of Ewens distributions, see also [8, §2.2] or [25]. We elaborate this situation in terms of Lemma 2.3 — which leads to split dagger idempotents. Due to space constraints, we only look at the main lines here. The appendix contains more details.

The Ewens distribution, from [9], can be described as a channel $ew[K]: \mathbb{R}_{>0} \rightarrow \mathcal{D}(MP(K))$, given by:

$$ew[K](t) := \frac{K!}{\prod_{0 \leq i < K} t+i} \cdot \sum_{\sigma \in MP(K)} \frac{t^{size(\sigma)}}{\sigma_{\parallel} \cdot prod(\sigma)} |\sigma\rangle. \quad (21)$$

We show that by updating with a fixed size n , where $1 \leq n \leq K$, the dependence on the parameter t disappears, and gives us our dagger $size^\dagger$.

$$\begin{aligned} ew[K](t) \Big|_{size \ll 1_n} &= \sum_{\sigma \in size^{-1}(n)} \frac{\frac{t^{size(\sigma)}}{\sigma_{\parallel} \cdot prod(\sigma)}}{\sum_{\tau \in size^{-1}(n)} \frac{t^{size(\tau)}}{\tau_{\parallel} \cdot prod(\tau)}} |\sigma\rangle \\ &= \sum_{\sigma \in size^{-1}(n)} \frac{\frac{1}{\sigma_{\parallel} \cdot prod(\sigma)}}{\sum_{\tau \in size^{-1}(n)} \frac{1}{\tau_{\parallel} \cdot prod(\tau)}} |\sigma\rangle =: size^\dagger(n). \end{aligned}$$

Theorem 6.1 For $K \geq 1$, the size function $size: MP(K) \rightarrow \{1, \dots, K\}$ is a sufficient statistic for the

Ewens channel (21), via the equality of channels $\mathbb{R}_{>0} \rightarrow MP(K) \times \{1, \dots, K\}$ in:

where

$$\text{stir}[K]: \mathbb{R}_{>0} \rightarrow \mathcal{D}(\{1, \dots, K\}) \text{ is the 'Stirling' channel given by:}$$

$$\text{stir}[K](t) := \frac{1}{\prod_{0 \leq i < K} t+i} \cdot \sum_{1 \leq k \leq K} \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^k |k\rangle.$$

The latter Stirling distribution $\text{stir}[K]$ involves the Stirling numbers of the first kind $\begin{bmatrix} K \\ k \end{bmatrix}$, see [21] for more details. It is sometimes called the Chinese restaurant table distribution. \square

7 Sums of sequences as sufficient statistic

For a fixed number K we consider the addition function $\text{sum}: \mathbb{N}^K \rightarrow \mathbb{N}$. It is the retraction part of a split idempotent, with section $\text{sum}^\dagger: \mathbb{N} \rightarrow \mathcal{D}(\mathbb{N}^K)$ given by:

$$\text{sum}^\dagger(n) := \sum_{\mathbf{k} \in \text{sum}^{-1}(n)} \frac{\binom{n}{\mathbf{k}}}{K^n} |\mathbf{k}\rangle = \sum_{\mathbf{k} \in \text{sum}^{-1}(n)} \frac{n!}{K^n \cdot \prod_i k_i!} |\mathbf{k}\rangle.$$

This yields a distribution by Lemma 2.1 (ii) since a sequence $\mathbf{k} \in \mathbb{N}$ with $\text{sum}(\mathbf{k}) = n$ can be identified with a multiset in $\mathcal{M}[n](\{1, \dots, K\})$. Our final result is a well known instance of sufficiency, involving the Poisson channel $\text{pois}: \mathbb{R}_{>0} \rightarrow \mathcal{D}_\infty(\mathbb{N})$ given by $\text{pois}[\lambda] := \sum_n e^{-\lambda} \cdot \frac{\lambda^n}{n!} |n\rangle$, with infinite support. The appendix contains a proof.

Theorem 7.1 For $K \geq 1$, the addition function $\text{sum}: \mathbb{N}^K \rightarrow \mathbb{N}$ is a sufficient statistic for K -fold poisson product $\otimes_K \text{pois}: \mathbb{R}_{>0} \rightarrow \mathcal{D}(\mathbb{N}^K)$ in the following equality of channels $\mathbb{R}_{>0} \rightarrow \mathbb{N}^K \times \mathbb{N}$.

where

$$\begin{cases} (\otimes_K \text{pois})(\lambda) = \text{pois}[\lambda] \otimes \dots \otimes \text{pois}[\lambda] \\ \text{pois}[K-](\lambda) = \text{pois}[K\lambda]. \end{cases} \quad \square$$

An analogous example uses geometric instead of Poisson distributions.

8 Conclusions

This paper elaborates several examples of a sufficient statistic, in the recently introduced ‘adjoint’ formulation using string diagrams of [11]. It does so for discrete distributions, since sufficiency is not well-developed there — in contrast to the continuous case. Four examples are described, following the same pattern, by identifying the relevant split idempotents — that arise via the Fisher-Neyman factorisation theorem. The paper’s focus is on concrete mathematical (combinatorial) structure, and not so much on theory development. Ewens’ distributions capture probabilistic mutations for multiset partitions. In [19] this approach is extended to other data types — including set partitions, also studied in [24] — and leads to many more examples of sufficient statistics. There is room for further investigations of, for instance, the use of ‘divisions’ from [20] instead of the ‘partition’ distributions in Section 5.

References

[1] G. Andrews. *The Theory of Partitions*. Cambridge Univ. Press, 1998. ISBN 978-0521637664.

- [2] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, 2000. Available via <https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316870>. doi:10.1002/9780470316870.
- [3] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006, Publisher. ISBN 978-0387310732
- [4] T. Broderick, J. Pitman, and M. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8:801–836, 2013. doi:10.1214/13-BA823.
- [5] K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019. doi:10.1017/s0960129518000488.
- [6] K. Cho, B. Jacobs, A. Westerbaan, and B. Westerbaan. An introduction to effectus theory. see <https://arxiv.org/abs/1512.05813>, 2015.
- [7] F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017. doi:10.1007/978-3-662-54458-7_21.
- [8] H. Crane. The ubiquitous Ewens sampling formula. *Statistical Science*, 31(1):1–19, 2016. doi:10.1214/15-ST529.
- [9] W. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112, 1972. doi:10.1016/0040-5809(72)90035-4.
- [10] R. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Royal Soc.*, 222:594–604, 1922. doi:10.1098/rsta.1922.0009.
- [11] T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020. doi:10.1016/J.AIM.2020.107239.
- [12] T. Fritz, T. Gonda, P. Perrone, and E. Rischel. Representable Markov categories and comparison of statistical experiments in categorical probability. See <https://arxiv.org/abs/2010.07416>, 2020.
- [13] T. Fritz and E. Rischel. Infinite products and zero-one laws in categorical probability. *Compositionality*, 2(3), 2020. doi:10.32408/compositionality-2-3.
- [14] C. Heunen and B. Jacobs. Quantum logic in dagger kernel categories. *Order*, 27(2):177–212, 2010. doi:10.1007/s11083-010-9145-5.
- [15] B. Jacobs. New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Comp. Sci.*, 11(3), 2015. doi:10.2168/lmcs-11(3:24)2015.
- [16] B. Jacobs. A recipe for state and effect triangles. *Logical Methods in Comp. Sci.*, 13(2), 2017. See <https://lmcs.episciences.org/3660>.
- [17] B. Jacobs. The mathematics of changing one’s mind, via Jeffrey’s or via Pearl’s update rule. *Journ. of Artif. Intelligence Research*, 65:783–806, 2019. doi:10.1613/jair.1.11349.
- [18] B. Jacobs. From multisets over distributions to distributions over multisets. In *Logic in Computer Science*. IEEE, Computer Science Press, 2021. doi:10.1109/lics52264.2021.9470678.
- [19] B. Jacobs. Basic combinatorics and sufficient statistics for mutations on multiple datatypes. Researchers.One, <https://researchers.one/articles/22.11.00003>, 2022.
- [20] B. Jacobs. Partitions and Ewens distributions in element-free probability theory. In *Logic in Computer Science*. IEEE, Computer Science Press, 2022. Article No. 23. doi:10.1145/3531130.3532419.
- [21] B. Jacobs. A reconstruction of Ewens’ sampling formula via lists of coins. In N. Jansen, M. Stoelinga, and P. van den Bos, editors, *A Journey from Process Algebra via Timed Automata to Model Learning. Essays Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday*, number 13560 in Lect. Notes Comp. Sci., pages 339–357. Springer, Berlin, 2022. doi:10.1007/978-3-031-15629-8_18.
- [22] B. Jacobs. Stick breaking, in coalgebra and probability. In H. Hansen and F. Zanasi, editors, *Coalgebraic Methods in Computer Science (CMCS 2022)*, number 13225 in Lect. Notes Comp. Sci. Springer, Berlin, 2022. doi:10.1007/978-3-031-10736-8_9.
- [23] B. Jacobs. Urns & tubes. *Compositionality*, 4(4), 2022. doi:10.32408/compositionality-4-4.

- [24] B. Jacobs and D. Stein. Counting and matching. In B. Klin and E. Pimentel, editors, *Computer Science Logic*, volume 252 of *LIPICs*, pages 28:1–28:15. Schloss Dagstuhl, 2023. doi:10.4230/LIPICs.CSL.2023.28.
- [25] P. Joyce. Partition structures and sufficient statistics. *Journ. of Applied Probability*, 35(3):622–632, 1998. doi:10.1239/jap/1032265210.
- [26] J. Kingman. Random partitions in population genetics. *Proc. Royal Soc., Series A*, 361:1–20, 1978. doi:10.1098/rspa.1978.0089.
- [27] J. Kingman. The representation of partition structures. *Journ. London Math. Soc.*, 18(2):374–380, 1978. doi:10.1112/jlms/s2-18.2.374.
- [28] S. Kolm. The optimal production of social justice. In J. Margolis and H. Guitton, editors, *Public Economics: An Analysis of Public Production and Consumption and their Relations to the Private Sectors*, pages 145–200, London, 1969. Macmillan.
- [29] P. Selinger. Idempotents in dagger categories (extended abstract). In P. Selinger, editor, *Proceedings of the 4th International Workshop on Quantum Programming Languages (QPL 2006)*, number 210 in *Elect. Notes in Theor. Comp. Sci.*, pages 107–122. Elsevier, Amsterdam, 2008. doi:10.1016/j.entcs.2008.04.021.
- [30] D. Stein. *Structural Foundations for Probabilistic Programming Languages*. PhD thesis, Cambridge Univ., 2022. Available via <https://dario-stein.de/notes/thesis.pdf>.
- [31] Y. Suhov and M. Kelbert. *Probability and Statistics by Example: Volume 1, Basic Probability and Statistics*. Cambridge Univ. Press, 2005.
- [32] S. Tavaré. The magical Ewens sampling formula. *Bull. London Math. Soc.*, 53:1563–1582, 2021. doi:10.1112/blms.12537.
- [33] B. Westerbaan. *Dagger and Dilation in the Category of Von Neumann algebras*. PhD thesis, Radboud Univ. Nijmegen, 2019. Available online via <https://arxiv.org/abs/1803.01911v1>.

A Appendix

Proof of Theorem 6.1

For $K \geq 1$ we have the Ewens and Stirling channels:

$$\mathbb{R}_{>0} \xrightarrow{\text{ew}[K]} MP(K) \qquad \mathbb{R}_{>0} \xrightarrow{\text{stir}[K]} \{1, \dots, K\}$$

For convenience, we repeat the definitions from Section 6: for a reproduction parameter $t \in \mathbb{R}_{>0}$,

$$\begin{aligned} \text{ew}[K](t) &= \frac{K!}{\prod_{0 \leq i < K} t+i} \cdot \sum_{\sigma \in MP(K)} \frac{t^{\text{size}(\sigma)}}{\sigma \circ \text{prod}(\sigma)} |\sigma\rangle \\ \text{stir}[K](t) &= \frac{1}{\prod_{0 \leq i < K} t+i} \cdot \sum_{1 \leq k \leq K} \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^k |k\rangle. \end{aligned} \tag{A.1}$$

In the description of the Stirling distribution $\text{stir}[K](t)$ We write $\begin{bmatrix} K \\ k \end{bmatrix}$ for the Stirling number of the first kind. These numbers are determined by the equations:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = 1 \quad \begin{bmatrix} 0 \\ k \end{bmatrix} = \begin{bmatrix} k \\ 0 \end{bmatrix} = 0 \quad \begin{bmatrix} n+1 \\ k \end{bmatrix} = n \cdot \begin{bmatrix} n \\ k \end{bmatrix} + \begin{bmatrix} n \\ k-1 \end{bmatrix}. \tag{A.2}$$

One can then derive, by induction on $K \geq 1$, that for all $t > 0$,

$$\sum_{1 \leq k \leq K} \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^k = \prod_{0 \leq i < K} t+i \tag{A.3}$$

Therefore, the probabilities in the above Stirling distributions add up to one.

We use the following general result about the Ewens and Stirling distributions (A.1). The part about Ewens is copied from [22].

Lemma A.1 Fix $K \in \mathbb{N}$ and $t \in \mathbb{R}_{>0}$.

(i) There is a commuting rectangle of ‘draw-add’ channels:

$$\begin{array}{ccc} MP(K) & \xrightarrow{PDA[K](t)} & \mathcal{D}(MP(K+1)) \\ \text{size} \downarrow & & \downarrow \mathcal{D}(\text{size}) \\ \{1, \dots, K\} & \xrightarrow{SDA[K](t)} & \mathcal{D}(\{1, \dots, K+1\}) \end{array}$$

In which the partition and set draw-add channels PDA and SDA are defined by:

$$\begin{aligned} PDA(t)(\sigma) &:= \frac{t}{K+t} |\sigma + 1|1\rangle + \sum_{1 \leq k \leq K} \frac{\sigma(k) \cdot k}{K+t} |\sigma - 1|k\rangle + 1|k+1\rangle \\ SDA(t)(k) &:= \frac{t}{K+t} |k+1\rangle + \frac{K}{K+t} |k\rangle. \end{aligned}$$

(ii) Both the Ewens and the Stirling distributions can be obtained inductively via: We follow [22] in the inductive description of Ewens distributions via state transformation as:

$$\left\{ \begin{array}{l} \text{ew}[1](t) = 1 |1|1\rangle \\ \text{ew}[K+1](t) = PDA[K](t) \gg \text{ew}[K](t) \end{array} \right\} \quad \left\{ \begin{array}{l} \text{stir}[1](t) = 1|1\rangle \\ \text{stir}[K+1](t) = SDA[K](t) \gg \text{stir}[K](t). \end{array} \right.$$

(iii) $\mathcal{D}(\text{size}) \circ \text{ew}[K] = \text{stir}[K]$.

In Section 6 we have already determined the dagger to the size function. Hence this equation in item (iii) suffices to prove Theorem 6.1.

Proof. (i) For $\sigma \in MP(K)$,

$$\begin{aligned} (\mathcal{D}(\text{size}) \circ PDA[K](t))(\sigma) &= \frac{t}{K+t} |\text{size}(\sigma + 1|1)\rangle + \sum_{1 \leq k \leq K} \frac{\sigma(k) \cdot k}{K+t} |\text{size}(\sigma - 1|k) + 1|k+1\rangle \\ &= \frac{t}{K+t} |\text{size}(\sigma) + 1\rangle + \frac{\sum_{1 \leq k \leq K} \sigma(k) \cdot k}{K+t} |\text{size}(\sigma)\rangle \\ &= \frac{t}{K+t} |\text{size}(\sigma) + 1\rangle + \frac{K}{K+t} |\text{size}(\sigma)\rangle \\ &= SDA[K](t)(\text{size}(\sigma)). \end{aligned}$$

(ii) We leave the Ewens case to the reader and concentrate on Stirling. First, for $K = 1$,

$$\text{stir}[1](t) = \frac{1}{\prod_{0 \leq i < 1} t+i} \cdot \sum_{1 \leq k \leq 1} \begin{bmatrix} 1 \\ k \end{bmatrix} \cdot t^k |k\rangle = \frac{1}{t} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot t^1 |1\rangle = 1|1\rangle.$$

Next,

$$\begin{aligned}
& SDA[K](t) \gg stir[K](t) \\
&= \sum_{1 \leq \ell \leq K+1} \sum_{1 \leq k \leq K} SDA[K](t)(k)(\ell) \cdot stir[K](t)(k) | \ell \rangle \\
&= \sum_{1 \leq k \leq K} \frac{t}{K+t} \cdot stir[K](t)(k) | k+1 \rangle + \sum_{1 \leq k \leq K} \frac{K}{K+t} \cdot stir[K](t)(k) | k \rangle \\
&\stackrel{(IH)}{=} \sum_{1 \leq k \leq K} \frac{t}{K+t} \cdot \frac{1}{\prod_{0 \leq i < K} t+i} \cdot \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^k | k+1 \rangle + \sum_{1 \leq k \leq K} \frac{K}{K+t} \cdot \frac{1}{\prod_{0 \leq i < K} t+i} \cdot \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^k | k \rangle \\
&= \frac{1}{\prod_{i < K+1} t+i} \cdot \left(\sum_{1 \leq k \leq K} \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^{k+1} | k+1 \rangle + \sum_{1 \leq k \leq K} K \cdot \begin{bmatrix} K \\ k \end{bmatrix} \cdot t^k | k \rangle \right) \\
&= \frac{1}{\prod_{i < K+1} t+i} \cdot \left(K \cdot \begin{bmatrix} K \\ 1 \end{bmatrix} \cdot t^1 | 1 \rangle + \left(K \cdot \begin{bmatrix} K \\ 2 \end{bmatrix} + \begin{bmatrix} K \\ 1 \end{bmatrix} \right) \cdot t^2 | 2 \rangle + \dots + \right. \\
&\quad \left. \left(K \cdot \begin{bmatrix} K \\ K \end{bmatrix} + \begin{bmatrix} K \\ K-1 \end{bmatrix} \right) \cdot t^K | K \rangle + \begin{bmatrix} K \\ K \end{bmatrix} \cdot t^{K+1} | K+1 \rangle \right) \\
&\stackrel{(A.2)}{=} \frac{1}{\prod_{i < K+1} t+i} \cdot \left(\begin{bmatrix} K+1 \\ 1 \end{bmatrix} \cdot t^1 | 1 \rangle + \begin{bmatrix} K+1 \\ 2 \end{bmatrix} \cdot t^2 | 2 \rangle + \dots + \right. \\
&\quad \left. \begin{bmatrix} K+1 \\ K \end{bmatrix} \cdot t^K | K \rangle + \begin{bmatrix} K+1 \\ K+1 \end{bmatrix} \cdot t^{K+1} | K+1 \rangle \right) \\
&= \frac{1}{\prod_{i < K+1} t+i} \cdot \sum_{1 \leq k \leq K+1} \begin{bmatrix} K+1 \\ k \end{bmatrix} \cdot t^k | k \rangle \\
&= stir[K+1](t).
\end{aligned}$$

(iii) By induction on K , using items (i) and (ii). \square

As an aside, there is a no draw-delete channel $\{1, \dots, K+1\} \rightarrow \{1, \dots, K\}$ that commutes with size, as in Lemma A.1 (i), see [22] for more information about draw-add and draw-delete, for multisets and partitions.

Proof of Theorem 7.1

For $K \geq 1$ we have two \mathcal{D}_∞ -channels:

$$\mathbb{R}_{>0} \xrightarrow{\otimes_K \text{pois}} \mathbb{N}^K \quad \mathbb{R}_{>0} \xrightarrow{\text{pois}[K-]} \mathbb{N} \quad \text{with} \quad \begin{cases} (\otimes_K \text{pois})(\lambda) = \text{pois}[\lambda] \otimes \dots \otimes \text{pois}[\lambda] \\ (\text{pois}[K-])(\lambda) = \text{pois}[K\lambda]. \end{cases}$$

In addition, there is the *sum* sufficient statistic with its dagger sum^\dagger in:

$$\mathbb{N}^K \xrightarrow{sum} \mathbb{N} \quad \mathbb{N} \xrightarrow{sum^\dagger} \mathbb{N}^K \quad \text{with} \quad \begin{cases} sum(\mathbf{k}) = \sum_i k_i \\ sum^\dagger(n) = \sum_{\mathbf{k} \in sum^{-1}(n)} \frac{\binom{n}{\mathbf{k}}}{K^n} | \mathbf{k} \rangle \end{cases}$$

We choose to prove Theorem 7.1 via the predicate formulation of (14). For predicates $p \in \text{Pred}(\mathbb{N}^K)$ and

$q \in \text{Pred}(\mathbb{N})$, there is an adjoint correspondence:

$$\begin{aligned}
(\otimes_K \text{pois})(\lambda) \models p \ \& \ (\text{sum} \approx\!\!\ll q) &= \sum_{\mathbf{k} \in \mathbb{N}^K} \prod_i \text{pois}[\lambda](k_i) \cdot p(\mathbf{k}) \cdot q(\text{sum}(\mathbf{k})) \\
&= \sum_{\mathbf{k} \in \mathbb{N}^K} \prod_i e^{-\lambda} \cdot \frac{\lambda^{k_i}}{k_i!} \cdot p(\mathbf{k}) \cdot q(\text{sum}(\mathbf{k})) \\
&= \sum_{\mathbf{k} \in \mathbb{N}^K} e^{-K\lambda} \cdot \frac{\lambda^{\text{sum}(\mathbf{k})}}{\prod_i k_i!} \cdot p(\mathbf{k}) \cdot q(\text{sum}(\mathbf{k})) \\
&= \sum_{n \in \mathbb{N}} \sum_{\mathbf{k} \in \text{sum}^{-1}(n)} e^{-K\lambda} \cdot \frac{\lambda^n}{n!} \cdot \binom{n}{\mathbf{k}} \cdot p(\mathbf{k}) \cdot q(n) \\
&= \sum_{n \in \mathbb{N}} e^{-K\lambda} \cdot \frac{(K\lambda)^n}{n!} \cdot \left(\sum_{\mathbf{k} \in \text{sum}^{-1}(n)} \frac{\binom{n}{\mathbf{k}}}{K^n} \cdot p(\mathbf{k}) \right) \cdot q(n) \\
&= \sum_{n \in \mathbb{N}} \text{pois}[K\lambda](n) \cdot (\text{sum}^\dagger \approx\!\!\ll p)(n) \cdot q(n) \\
&= (\text{pois}[K-])(\lambda) \models (\text{sum}^\dagger \approx\!\!\ll p) \ \& \ q.
\end{aligned}$$