LOVEGROVE MATHEMATICALS

THE FUNDAMENTALS OF LIKELINESSES

RESEARCH REPORT 2015-01

Roger Lovegrove

Likelinesses can be viewed in at least two ways. The routine way is as best-estimates of probabilities. Alternatively, they can be viewed as being a fundamental entity in their own right. If the latter view is taken -as it is here- then a rich theoretical background is opened up.

Research Reports is an occasional series of papers on various subjects of interest. They are supplied for information only on an informal, non-contractual basis; no responsibility is accepted for any consequences of the use of their contents. Although copyright is retained by the Author reasonable extracts may be quoted freely, with acknowledgment.

This is the first reprint of 2013-10, which is now withdrawn.

The recommended citation is

Lovegrove R., (2015), The Fundamentals of Likelinesses, Lovegrove Mathematicals Research Report 2015-01. April 2015.

LONDON
UNITED KINGDOM
April 2015

www.lovegrovemaths.co.uk
roger@lovegrovemaths.co.uk

Contents

1 Introduction

An end-user problem requiring the estimation of a probability distribution often requires that distribution to be of a particular geometric shape (ranked, unimodal, bell-shaped etc) but not necessarily of any particular parametrically-defined form. The problem might, for example, require a bell-shaped distribution but not necessarily any particular bell-shaped distribution such as the Normal.

Under such circumstances, to use a particular parametrically-defined form would be to consider only a subset of the possible solutions. This could affect both the precision and the accuracy of the solution and thus potentially cause the end-user to place greater, or lesser, reliance upon results than is justified by any data.

Consider what is perhaps the simplest case; that of ranked distributions. These have long been a source of interest because of their mixture of simplicity and complexity. Because they cover such a disparate range of data, the emphasis by various authors has historically been on particular types of ranked distribution which can be represented parametrically in various ways.

Zipf [13] famously saw in the distribution of the commoner words in the English language the pattern called Zipf's Law, now often called the Zipf/Pareto Law in recognition of Pareto's earlier work [9] on economics. Here, the governing equation is $Z(i) = K.(1/i)$, for some K the role of which is to normalise the distribution by bringing the sum of the terms to 1; on the finite set $i \in \{1, \ldots, N\}$ this becomes

$$Z(i) = \frac{1}{\sum\limits_{j=1}^{N}(1/j)}(\frac{1}{i}). \tag{1}$$

Mandelbrot [7] extended this (with different notation) to

$$Z(i) = \frac{(i+V)^{-1/D}}{\Sigma_{j=1}^{N}(j+V)^{-1/D}}$$

for some D,V.

Regardless of which form is used, however, it is clear that neither representation could be expected to cover all ranked distributions, and that -indeed- there is no parametric form which could do so.

A similar situation arises with other types of distributions, such as unimodal distributions. The basic concept is clearly of genuine potential use, but taking a parametric approach virtually forces us to consider specifically bell-shaped distributions of various forms. This happens to such an extent that it is difficult to find any description of unimodal distributions which does not assume their fundamental shape to be bell-shaped. This is despite that this clearly cannot be the case in general since it is all too easy to draw an unimodal distribution which is not bell-shaped.

If we want to analyse ranked distributions generally, or unimodal distributions rather than specifically bell-shaped distributions, then we need to take a set-oriented approach, rather than a parameter-oriented one, by considering the set of all ranked, unimodal etc distributions (on a finite domain). Such sets of distributions form the natural environment for best-estimates of probabilities: what will here be called"likelinesses".

This change in terminology from "best-estimate of probability" to "likeliness" is not simply to reduce the wordage. It also heralds a change in philosophical emphasis away from probabilities to likelinesses: a change in what Whittle called [10] the "entity of prime interest". Whittle was interested in making expectation the entity of prime interest; here, we shall be following a different line by making likelinesses the centre of our investigations.

The consequences of this change in entity of prime interest include:

- We shall not be interested in how good an estimate of probability our best-estimates are. Such questions are appropriate when the entity of prime interest is probability, but not when it is likeliness.
- We shall not be concerned with whether or not a sample is "large". Large samples are of importance when using frequentist probability because of the definition in terms of limit. Likelinesses are defined without using limits, so the size of a sample will not be of concern. Consequently, likelinesses are very much, but not only, a "theory of small samples".
- We do not need the concept of probability to define likeliness. Instead, we shall be defining probability in terms of likeliness.
- We shall be able to distinguish between a probability and a best-estimate of a probability, and so say things such as "Probabilities may be substituted into the Multinomial Theorem, but best-estimates may not". Such statements are meaningless in the Kolmogorov approach -which is so wide-ranging that it captures best-estimates.

There is, unavoidably, the problem of how to calculate likelinesses. Although there are simple cases for which formulae have been developed to enable us to find likelinesses, most real-life problems are so complicated that only a numerical analysis can be used, involving sampling from an appropriate set of distributions, called the underlying set; this might, for example, be the set of all ranked distributions or the set of all unimodal distributions, etc. Until recently, the scale of the number-crunching involved has placed such analyses beyond everyday computers; this has led to a lack of interest in the subject. Recent improvements in computer technology, namely improvements in FORTRAN and the change to 64-bit architecture for PCs, have brought the analyses back into the realm of the practical.

2 Notation and Terminology

Basics

Let $\mathbb{R}^+$ be the non-negative reals, and $\mathbb{N}^+$ be the non-negative integers. For $N \in \mathbb{N}$ let $X_N = \{1, \ldots, N\}$. N is called the *degree.*

Let $f : X_N \to]0, 1]$ be such that $\Sigma_{i=1}^{N} f(i) = 1$. Then f is called a *distribution on* X_N . $S(N)$ is the set of all such distributions.

Histograms and Integrams

Let $G(N) = \{g|g : X_N \to \mathbb{N}^+\}$, $H(N) = \{h|h : X_N \to \mathbb{R}^+\}$, so $G(N) \subset H(N)$. The elements of H(N) are called *histograms* on X_N and those of G(N) *integer-valued histograms*, shortened to *integrams*, on X_N. The histogram h is identified with the point $(h(1), \ldots, h(N))$.

For $h \in H(N)$, the *sample size* of h is $\omega(h) = \Sigma_{i=1}^N h(i)$.
For $h \in H(N)$, $f \in S(N)$ we define $f^h = f(1)^{h(1)} ... f(N)^{h(N)}$.
For $g \in G(N)$, the *Multinomial coefficient associated with g* is

$$M(g) = \frac{\omega(g)!}{\Pi_{i=1}^N g(i)!}.$$

The integram of degree N and sample size 0 is $(0, \ldots, 0)$, which is denoted by $\underline{0}$, or –if greater clarity is needed– by $\underline{0}_N$.

For $h \in H(N) \backslash \{\underline{0}\}, i \in X_N$, the *Relative Frequency of i given h* is $RF(i|h) = \dfrac{h(i)}{\omega(h)}$.

If we roll a die and throw the number 2 then we have not only thrown a 2 once but we have also thrown 1, 3, 4, 5 and 6 zero times each. So we can think of ourselves as having thrown the integram (0,1,0,0,0,0). Also, we have not actually thrown the number 2 but have, rather, thrown the face labelled “2”. It will be very convenient to adopt notation which associates the formal symbol “2” with (0,1,0,0,0,0).

We define $''i''_N$ to be that integram $(x_1, \ldots, x_N)$ for which $x_i = 1$ but $x_n = 0$ otherwise; for example, $''2''_6 = (0, 1, 0, 0, 0, 0)$. It is usually possible to write $''i''$ rather than $''i''_N$ without introducing ambiguity. Importantly, $f^{''i''} = f(i)$ and $M(''i'') = 1$.

Likeliness of an integram

For g∈G(N), h∈H(N), P a non-empty subset of S(N) we define

$$L_P(g|h) = M(g) \frac{\Sigma_{f \in P} f^g f^h}{\Sigma_{f \in P} f^h}$$

where Σ is the Daniell integral.

$L_P(g|h)$ is called the *likeliness, over P, of g given h.* Since P, g or h will usually be clear from the context, this terminology can normally be shortened by omitting appropriate terms.

h is called the *given histogram*, g the *required integram* and P the *underlying set.* More generally, any non-empty subset of S(N) is called an *underlying set* in S(N).

We have $L_P(\underline{0}|h) = 1$ for all (h,P). $L_P(g|\underline{0})$ is written as $L_P(g)$.

The distribution ($L_P(''1''|h), ..., L_P(''N''|h)$) is called the *L-point.* This is a member of P if P is convex: in general, it belongs to Core(P).

If P is a singleton set, $P = \{f\}$, then $L_P(g|h) = M(g)f^g$, which is denoted by $Pr(g|f, h)$: since this is independent of h the notation may be simplified to $Pr(g|f)$; however, the presence of the h, although technically unnecessary, can sometimes add clarity.

If the context allows, which it usually does, we can further simplify the notation by writing $L_P(i|h)$ rather than $L_P(''i''|h)$ and $Pr(i|f)$ rather than $Pr(''i''|f)$.

Likeliness of a set of distributions

Let $V \subset S(N)$. Then the *likeliness of V, over P and given h*, is

$$L_P(V|h) = \frac{\Sigma_{f \in V \cap P} f^h}{\Sigma_{f \in P} f^h}.$$

For $x \in [0, 1]$ let $V_x = \{f \in S(N)|Pr(g|f) < x\}$. Then $L_P(V_x|h)$ is the likeliness, over P and given h, of the set of those $f \in P$ for which $Pr(g|f) < x$. We denote this by $L_P(Pr(g|f) < x|h)$. That is, in the text string '$L_P(V|h)$' we replace the name 'V' by the definition of V.

The function $\bullet : [0, 1] \to [0, 1] : x \mapsto L_P(Pr(g|f) \leq x|h)$ is the *expected CDF of $Pr(g|f)$*.

Likewise, if $0 \leq x_0 \leq x_1 \leq 1$ then we define $L_P(Pr(g|f) \in [x_0, x_1]|h)$ to be $L_P(V|h)$ where $V = \{f \in P|Pr(g|f) \in [x_0, x_1]\}$: that is, by again replacing the name 'V' by the definition of the set V. By covering [0,1] by cells in this way, we obtain an *expected frequency distribution for $Pr(g|f)$*.

3 General Results

3.1 Chain Rule

Theorem 1. *Chain Rule*
$(\forall g_2, g_1 \in G(N))\ (\forall h \in H(N))\ L_P(g_2 + g_1|h) = \dfrac{M(g_2 + g_1)}{M(g_2)M(g_1)} L_P(g_2|g_1 + h)L_P(g_1|h)$

Proof. The proof is by substitution of the definitions of the likelinesses. □

This does generalise in the expected way:

$$L_P(g_m + \cdots + g_1|h) = \frac{M(g_m + \cdots + g_1)}{M(g_m) \ldots M(g_1)} L_P(g_m|g_{m-1} + \cdots + g_1 + h) \ldots L_P(g_1|h). \quad (2)$$

For example,

$$L_P(g_3 + g_2 + g_1|h) = \frac{M(g_3 + g_2 + g_1)}{M(g_3)M(g_2)M(g_1)} L_P(g_3|g_2 + g_1 + h)L_P(g_2|g_1 + h)L_P(g_1|h).$$

3.2 Bayes's Rule (Likeliness form)

Theorem 2. *Bayes's Rule (Likeliness form)*

$$L_P(g_2|g_1+h) = L_P(g_1|g_2+h).\frac{L_P(g_2|h)}{L_P(g_1|h)},$$

Proof. Since addition of integrams is commutative, we have $L_P(g_2+g_1|h) = L_P(g_1+g_2|h)$ and therefore, from the Chain Rule (Theorem (1)),

$$L_P(g_2|g_1+h).L_P(g_1|h) = L_P(g_1|g_2+h).L_P(g_2|h). \qquad \square$$

In particular, putting $h = \underline{0}$ gives

$$L_P(g_2|g_1) = L_P(g_1|g_2).\frac{L_P(g_2)}{L_P(g_1)}.$$

3.3 Multinomial Theorem

3.3.1 Multinomial Consistent

Let $g \in G(N), h \in h(N)$ and P be an underlying set in S(N). Then we shall say that (g, h, P) is *Multinomial Consistent* (MC) if

$$L_P(g|h) \;=\; M(g)F_h^g$$

where $F_h \in S(N)$ is given by $F_h(i) = L_P(i|h)$ for all $i \in X_N$.

In general, (g, h, P) is not MC. Theorem 3 gives two circumstances under which it is.

Theorem 3. *If $\omega(g) \leqslant 1$ or P is singleton then (g, h, P) is MC.*

Proof. (i) If $\omega(g) = 0$ then g=$\underline{0}$ so $L_P(g|h) = 1$, $M(g) = 1$ and $F_h^g = 1$. So $L_P(g|h) = 1 = 1.1 = M(g)F_h^g$.

So (g, h, P) is MC.

(ii) If $\omega(g) = 1$ then $g ='' i''$ for some $i \in X_N$ and we may wlg take $i = 1$. Then $M(g)F_h^g = 1.L_P(1|h)^1 L_P(2|h)^0 \dots L_P(N|h)^0 = L_P(1|h) = L_P(g|h)$.

So (g, h, P) is MC.

(iii) If P is singleton, $P = \{f\}$, then $(\forall i \in X_N)\ L_P(i|h) = f(i)$, so $F_h = f$. So $L_P(g|h) = M(g)f^g = M(g)F_h^g$.

So (g, h, P) is MC.

□

3.3.2 Multinomial Consistency

With the same notation, we define the *Multinomial Consistency* of (g, h, P) by

$$C_M(g, h, P) = \frac{L_P(g|h)}{M(g)F_h^g} \quad .$$

$C_M(g, h, P) = 1$ iff (g, h, P) is MC.

3.3.3 Hypothesis

If $\alpha \geqslant 0$ then, for any given (g, h, P), $C_M(g, \alpha h, P)$ is a function of α. It is hypothesised, but has not been proved, that, unless P is singleton, or $\omega(g) \leqslant 1$ or $h = \underline{0}$, that function is always either strictly increasing or strictly decreasing. If this is the case then, with the exceptions noted, that function takes the value of 1 at most once, so the set of α for which $(g, \alpha h, P)$ is MC is of measure zero.

4 Symmetry

4.1 (i,j)-symmetry

Let $i, j \in X_N$. For $f \in S(N)$ we define $f_{(i,j)}$ to be that distribution obtained from f by interchanging f(i) and f(j). We then say that the underlying set P is *(i,j)-symmetric* if P contains $f_{(i,j)}$ whenever P contains f. We say that P is *symmetric* if P is (i,j)-symmetric for all $i, j \in X_N$.

We say that $h \in H(N)$ is (i,j)-symmetric if h(i)=h(j), and that it is symmetric if it is (i,j)-symmetric for all i,j, ie. if it is a constant histogram $h = \underline{c} = (c, c, ..., c)$ for some c.

4.2 Indifference

Theorem 4. *Indifference Theorem*

Let P be an underlying set in S(N), $i, j \in X_N$ and $h \in H(N)$.
If P and h are both (i,j)-symmetric then $L_P(j|h) = L_P(i|h)$.

Proof. The proof is a simple matter of label-interchange. Interchange the labels ‘i’ and ‘j’; P is unaffected because it is (i,j)-symmetric.

Likewise, h is also unaffected since it, too, is (i,j)-symmetric.

So if we write the expression for $L_P(i|h)$, namely

$$L_P(i|h) = \frac{\Sigma_{f \in P} f(i) f^h}{\Sigma_{f \in P} f^h}$$

then all that changes is that the ‘i’ in ‘$L_P(i|h)$’ becomes a ‘j’, so that ‘$L_P(i|h)$’ becomes ‘$L_P(j|h)$’. In particular, the RHS does not alter and so retains its original value of $L_P(i|h)$.

Hence $L_P(j|h) = L_P(i|h)$ $\square$

4.2.1 Probabilistic versions

Historically, the Principle of Indifference (aka the "Law of Insufficient Reason") has always been incorrectly worded as giving probabilities with the singleton underlying set {f} where f is the L-point. The result is various so-called paradoxes, of which the most famous is the Perfect Cube Factory.

Keynes [5] wrote it as

If there is no *known* reason for predicating of our subject one rather than another of several alternatives, then relative to such knowledge the assertations of each of these alternatives have an *equal* **probability**. Thus *equal* **probabilities** must be assigned to each of several arguments, if there is an absence of positive ground for assigning *unequal* ones

Hájek [2] gives it as

"*whenever there is no evidence favoring one possibility over another, they have the same* ***probability***."

Wikipedia [11] introduces the concept of the names (labels) of the outputs:-

"*Suppose that there are* $n > 1$ *mutually exclusive and collectively exhaustive possibilities. The principle of indifference states that if the n possibilities are indistinguishable except for their names, then each possibility should be assigned a* ***probability*** *equal to 1/n.*"

(In each case, the bolding of the word '**probability**' is mine.)

4.2.2 Perfect Cube Factory

There are several versions, all essentially the same, of the Perfect Cube Factory. This version is taken from Hájek (*ibid*).

A factory produces cubes with side-length between 0 and 1 foot; what is the probability that a randomly chosen cube has side-length between 0 and 1/2 a foot? The tempting answer is 1/2, as we imagine a process of production that is uniformly distributed over side-length. But the question could have been given an equivalent restatement: A factory produces cubes with face-area between 0 and 1 square-feet; what is the probability that a randomly chosen cube has face-area between 0 and 1/4 square-feet? Now the tempting answer is 1/4, as we imagine a process of production that is uniformly distributed over face-area. This is already disastrous, as we cannot allow the same event to have two different probabilities (especially if this interpretation is to be admissible!). But there is worse to come, for the problem could have been restated equivalently again: A factory *produces cubes with volume between 0 and 1 cubic feet; what is the probability that a randomly chosen cube has volume between 0 and 1/8 cubic-feet? Now the tempting answer is 1/8, as we imagine a process of production that is uniformly distributed over volume. And so on for all of the infinitely many equivalent reformulations of the problem (in terms of the fourth, fifth, power of the length, and indeed in terms of every non-zero real-valued exponent of the length). What, then, is the probability of the event in question?*

The paradox arises because the principle of indifference can be used in incompatible ways. We have no evidence that favors the side-length lying in the interval [0, 1/2] over its lying in [1/2, 1], or vice versa, so the principle requires us to give probability 1/2 to each. Unfortunately, we also have no evidence that favors the face-area lying in any of the four intervals [0, 1/4], [1/4, 1/2], [1/2, 3/4], and [3/4, 1] over any of the others, so we must give probability 1/4 to each. The event the side-length lies in [0, 1/2], receives a different probability when merely redescribed. And so it goes, for all the other reformulations of the problem. We cannot meet any pair of these constraints simultaneously, let alone all of them.

Theorem 4 makes it clear that Indifference applies to the Likeliness, that is the mean probability. In the case of the Perfect Cube Factory, the underlying set is S(N), where N is the number of subintervals which we divide]0.1[into, and the given histogram is $\underline{0}$. The problem as stated incorrectly takes the underlying set as {f} where f is the L-point.

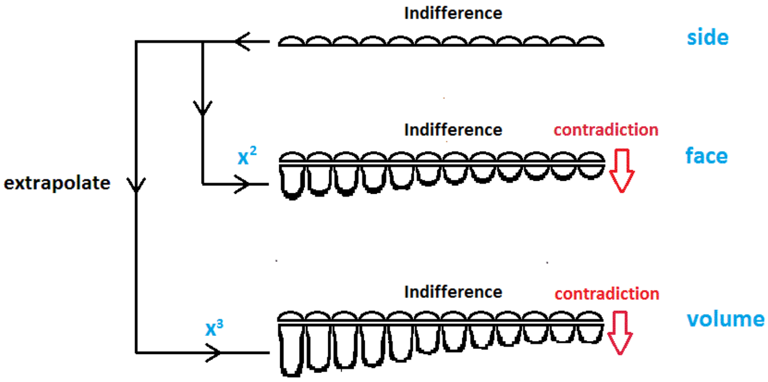


Figure 1: Contradictions in the Perfect Cube Factory

Figure 1 shows how the problem views contradictions as arising.

The falsely-worded Indifference Principle claims to give the 'probability'. Does this mean the actual probability, which we would obtain if we were to visit the Factory and measure a large number of cubes? It clearly does not; the whole point of the problem, is that we do not know what that distribution is.

The uniform distribution, being over S(N), is an approximation to that distribution, not the distribution itself.

When we extrapolate, by squaring or cubing as appropriate, we obtain another approximation.

So we have two different approximations, obtained in different ways.

There is nothing wrong per se with having two different approximations: there is no logical contradiction. The paradox arises only if we forget that we are dealing with approximations and start thinking that Indifference and its extrapolation are both supposed to be giving the actual distribution and therefore should be giving the same result as one-another.

5 Underlying set= S(N)

5.1 Law of Succession

For $P = S(N), h \in G(N), i \in X_N$ we have

$$L_{S(N)}(''i''|h) = \frac{\Sigma_{f \in S(N)} f^{''i''} f^h}{\Sigma_{f \in S(N)} f^h}$$

Evaluation of the RHS is a standard problem, the solution to which is the multinomial version of Laplace's Law of Succession: -

Theorem 5. *Law of Succession*

$$L_{S(N)}(''i''|h) = \frac{1 \ + \ h(i)}{N \ + \ \omega(h)}$$

□

This implies that (for $\omega(h) > 0$) $L_{S(N)}(''i''|h)$ is between $\dfrac{1}{N}$ and $\dfrac{h(i)}{\omega(h)}$: the former is $L_{S(N)}(i)$ and the latter is $RF(i|h)$.

The proof of Theorem 5 is surprisingly complex and lengthy, and is also perhaps-not-surprisingly difficult to track down in the literature. For the case $h \in G(N)$, which is sufficient for most of our purposes, the procedure is to use integration by parts to set up two reduction formulae -one to reduce the power of f(N) by 1, and another to reduce the degree by 1 if the power of f(N) is 0. These are used, turn-and-turn-about until the degree has been reduced to 1, at which point the integrals have become a single-variable Riemann integrals over an interval, which can then be integrated directly.

5.2 Combination Theorem

Introduction

This Theorem is here named after the Combination Postulate, which was proposed by Johnson [3] but not proved by him.

William Ernest Johnson was a late 19th-early 20th century logician, based in Cambridge, working on probability theory and economics. His work is important in the history of the development of probability theory since it was linked to, and a close forerunner of, de Finetti's work on exchangeability.

At the time of his death in 1931, Johnson was working [8] on a 4-volume work called *Logic*, the first three volumes of which were published posthumously; the fourth volume was not completed.

In Volume 3 [3] he wrote:-

> ...the calculus of probability does not enable us to infer any probability-value unless we have some probabilities or probability relations given.
>
> The following two postulates in the Theory of Eduction are concerned with the possible occurrences of the determinates $p_1, \dots, p_n$ under the determinable P.
>
> (1) Combination Postulate
>
> In a total of M instances, any proportion, say $m_1 : m_2 : \dots : m_\alpha$ where $m_1 + m_2 + \cdots + m_\alpha = M$, is as likely as any other, prior to any knowledge of the occurrences in question.
>
> (2) Permutation Postulate
>
> Each of the different orders in which a given proportion $m_1 : m_2 : \alpha$ for M instances may be presented is as likely as any other, whatever may have been the previously known orders.
>
> In what follows certitude will be represented by unity.
>
> By (1), the probability of any one proportion in M instances
>
> $$= \frac{M!}{\alpha(\alpha+1)(\alpha+2)\dots(\alpha+M-1)}$$

[The word 'Eduction' is correct. It is not, as seems to have been assumed by Good [1] (or his proof-reader), a mis-spelling of 'Education'.]

The final expression is the reciprocal of

$$^{M+\alpha-1}C_{\alpha-1}$$

Using different notation, Good [*ibid*], apparently in the belief that he was quoting Johnson, gave this as the number of ordered α-partitions of M. Johnson, himself, described it [*ibid, page 178*] as the number of integral solutions of the equation $m_1 + m_2 + \cdots + m_\alpha = M$.

Johnson did not prove the Combination Postulate, and reportedly [1, 12] abandoned it in favour of another postulate [4] because he was not entirely satisfied with it. Good called that other postulate the "Sufficientness Postulate".

Since its wording might be slightly obscure, it may help to explain the Combination Postulate at this stage. Imagine rolling a die. In any one roll, there are six possibilities, or 'determinates', namely "1", "2",..., "6"; so $\alpha = 6$. If we were to roll the die 10 times then there would be 10 'instances' of those determinates; that is $M = 10$. Say the number of rolls of each face were (1, 3, 0, 2, 2, 2) respectively: these are the values of the m_i. Of course, $1 + 3 + 0 + 2 + 2 + 2 = 10$: that is, we have an ordered 6-tuple of non-negative integers summing to 10: this is an ordered 6-partition of 10.

What is confusing to the modern eye is Johnson's use of the word 'proportion' to refer to something which we would not usually think of as a proportion. He is using it to refer to an ordered 6-tuple such as (1, 3, 0, 2, 2, 2), ie. what we are here calling an integram. The Combination Postulate, when it says that any proportion is as likely as any other, is saying that any integram is as likely as any other of the same sample size.

There are two questions remaining about Johnson's wording, concerning the circumstances under which they are equally likely, and the meaning of the word 'likely',

Johnson uses the expression "prior to any knowledge of the occurences in question". That knowledge can come from two places: theory and observation, so there must be no knowledge from either source. No knowledge from theory suggests that the underlying set should be S(N); no knowledge from observation suggests that the given histogram should be $\underline{0}$. So, for example, if we consider the tossing of a coin then all we know about the probability-pair (Pr(“H”),Pr(“T”)) is that it is –as all probability-pairs must be– somewhere on the line segment from (0,1) to (1,0).

So far as the meaning of the word ‘likely’ is concerned, there are two possible contenders: probability and best-estimate of probability, ie. likeliness. It has to be remembered that we are specifically maintaining the distinction between the two.

In the formal wording of the Combination Postulate, Johnson uses the word ‘likely’ but does not actually refer to probabilities. He does use the word ‘probability’, but only outside of that formal wording. This admits the possibility that, when drafting the formal wording, Johnson may have been thinking (albeit at an intuitive level) of a wider concept than ‘probability’ but subsequently interpreted it as meaning specifically probability. Whether or not this was the case must, of course, be a matter of speculation but the condition he states does suggest that he may have been thinking about the expected value of the probability, ie. what happens on average, rather than the probability itself.

So could it be that Johnson’s wording of the Combination Postulate was correct but that his stated interpretation of it in terms of probabilities, rather than expected values of probabilities, was not? This would certainly cause him difficulties, as we know happened.

It will be shown that the Combination Postulate does apply to likelinesses if the underlying set is S(N) and the given histogram is $\underline{0}$.

Lemma 1. *Let $g \in \Omega_N(n)$ and $i \in X_N$ be such that $g(i) > 0$. Then*

$$L_{S(N)}(g) = K_n \cdot L_{S(N)}(g -'' i'')$$

where

$$K_n = \frac{n}{N+n-1}.$$

Proof. Write g in a form which allows the Chain Rule to be applied:

$$L_{S(N)}(g) = L_{S(N)}(''i'' + (g -'' i''))$$

$$= \frac{M(g)}{M(''i'')M(g -'' i'')} \cdot L_{S(N)}(''i''|g -'' i'') \cdot L_{S(N)}(g -'' i'').$$

Expanding the leading coefficient and from the Law of Succession, this is

$$\frac{\omega(g)}{g(i)} \cdot \frac{1 + (g(i) - 1)}{N + \omega(g -'' i'')} \cdot L_{S(N)}(g -'' i''),$$

which simplifies to the required result. □

Combination Theorem

Theorem 6. *Combination Theorem*

$$(\forall g \in \Omega_N(n), n > 0) L_{S(N)}(g) = \frac{n!}{N(N+1)\ldots(N+n-1)}.$$

Proof. Given any g for which $n = \omega(g) > 0$, it is always possible to find an i for which $g(i) > 0$. Since K_n is independent of that i we may use Lemma as a reduction formula to repeatedly reduce the sample size in steps of 1 –without needing to worry about which i is being used at any step– until it reaches 0, at which point we have

$$L_{S(N)}(g) = \Pi_{m=1}^{n} K_m \cdot L_{S(N)}(\underline{0}) = \Pi_{m=1}^{n} K_m = \Pi_{m=1}^{n} \frac{m}{N+m-1}$$

$$= \frac{n!}{N(N+1)\ldots(N+n-1)},$$

□

After allowing for the difference in notation compared to that used by Johnson, this agrees with the Combination Postulate as stated by him but not as interpreted by him.

It follows that $(\forall g_1, g_2 \in \Omega_N(n))\ L_{S(N)}(g_1) = L_{S(N)}(g_2)$.

5.3 Order of sample space

Likelinesses sum to 1 across the sample space so, from the Combination Theorem, we have

$$(\forall g \in \Omega_N(n)) L_{S(N)}(g) = \frac{1}{|\Omega_N(n)|}$$

and so

$$|\Omega_N(n)| = \frac{N(N+1)\ldots(N+n-1)}{n!} = {}^{n+N-1}C_{N-1}.$$

5.4 Integram Theorem

Theorem 7. *Integram Theorem*

$$(\forall g_1, g_2 \in G(N)) \quad L_{S(N)}(g_1|g_2) = \frac{M(g_1)M(g_2)}{M(g_1+g_2)} \frac{|\Omega_N(\omega(g_2))|}{|\Omega_N(\omega(g_1+g_2))|}$$

Proof. By the Chain Rule,

$$L_{S(N)}(g_1+g_2) = \frac{M(g_1+g_2)}{M(g_1)M(g_2)} L_{S(N)}(g_1|g_2) L_{S(N)}(g_2),$$

so

$$\frac{1}{|\Omega_N(\omega(g_1+g_2))|} = \frac{M(g_1+g_2)}{M(g_1)M(g_2)} L_{S(N)}(g_1|g_2) \frac{1}{|\Omega_N(\omega(g_2))|},$$

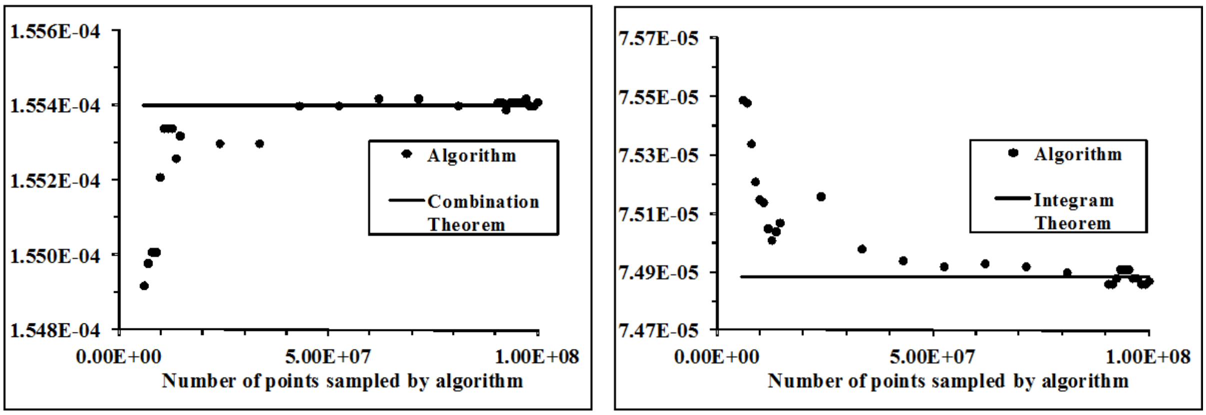
which rearranges to give the Integram Theorem: □

This includes both the Law of Succession and the Combination Theorem as special cases. The Law of Succession is recovered when $\omega(g_1) = 1$, that is $g_1 ='' i''$ for some i; the Combination Theorem when $g_2 = \underline{0}$.

5.5 Algorithm for S(N)

The algorithm used to generate an element, f, of S(N) is to use the RAND function to select N-1 distinct points in]0,1[, label them in ascending order as $P(1), \ldots, P(N-1)$, define P(0)=0 and P(N)=1, and then for $i = 1, \ldots, N$ take $f(i) = P(i) - P(i-1)$.

Take $N = 9$ and $g = (2,3,0,0,0,0,1,0,1)$. Then $\omega(g) = 7$, so, by the Combination Theorem, $L_{S(9)}(g) = 1/6435 = 1.554 * 10^{-4}$. See Figure 2a for a comparison of results produced by the algorithm with this value.



(a) Example 1: $L_{S(N)}(g)$

(b) Example 2: $L_{S(N)}(g|h)$

Figure 2: Convergence of algorithm for $L_{S(N)}(g)$ and $L_{S(N)}(g|h)$

With the same N and g, taking $h = (1,1,0,3,2,0,0,1,2)$ gives $\omega(h) = 10, \omega(g+h) = 17, M(g) = 420, M(h) = 151200, M(g+h) = 3.431 * 10^{10}$ and hence, by the Integram Theorem, $E_{S(9)}(g|h) = 7.489 * 10^{-5}$. See Figure 2b.

5.6 Coins

Define a *coin* to be any element of S(2). A physical disc used for the purposes of trade will be called a *minted coin.* Write $''H''$ rather than $''1''$ and $''T''$ rather than $''2''$. Interpret a result of tossing a coin (possibly minted) n times as an element of $\Omega_2(n)$.

That $f \in S(N)$ for which $(\forall i, j \in X_N) f(i) = f(j)$ will be called the *fair* element of S(N). The fair coin is the coin $(\frac{1}{2}, \frac{1}{2})$.

The likelinesses over $\{(\frac{1}{2}, \frac{1}{2})\}$ of the results of tossing **the fair coin** n times are given by Pascal's Triangle, derived from the Binomial Theorem (ie. the Multinomial Theorem), in Table 1a. The likelinesses over S(2) of the results of tossing **a coin** n times are also given by a triangle, as shown in Table 1b, derived from the Combination Theorem. In both tables, the rows $n = 0, 1$ correspond to the cases $\omega(g) = 0, 1$ in Theorem 3.

Table 2a shows the likelinesses over S(2) as the number of successive Heads increases. The likeliness of obtaining all $''H''$, ie. $L_{S(2)}(n''H'')$, where n is the number of tosses, comes from the Combination Theorem. The likeliness of $''H''$ once n Heads have been obtained, ie $L_{S(2)}(''H''|n''H'')$, is from the Law of Succession.

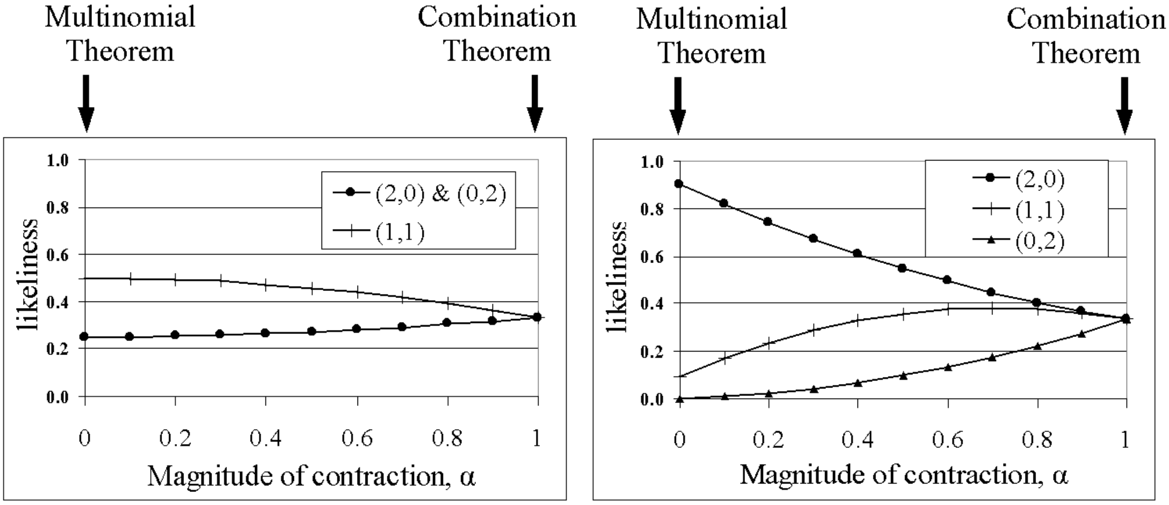
Table 2b shows the equivalent for the fair coin; because P is singleton, Pr has been written rather than L_P. $Pr(n''H''|f)$ comes from the Multinomial Theorem; since $Pr(''H''|f, n''H'')$ is independent of $n''H''$ it is always $Pr(''H''|f)$, ie. 0.50.

If a minted coin were being tossed then it would be unrealistic to expect it to be a precise concretisation of the fair coin: doing so would be equivalent to selecting a set of measure zero. On the other hand, since minted coins are manufactured under some form of quality control, it would also be unrealistic to expect them to have frequentist probabilities which were randomly distributed over S(2). Intuitively, we might in some sense anticipate a minted coin to be 'close to', but not necessarily coincident with, the fair coin.

We can investigate this type of situation by using an underlying set which is a contraction, centred on the fair coin, of S(2). We do this by introducing the mapping

$$S(2) \to S(2) : f \mapsto q + \alpha \cdot (f - q) \tag{3}$$

where q is the centre of the contraction, in this case $(\frac{1}{2}, \frac{1}{2})$, and α is its magnitude. If $\alpha = 0$ then f $\mapsto$ q, so the underlying set becomes the singleton set {q} and the Multinomial Theorem applies. If $\alpha = 1$ then f $\mapsto$ f, so nothing changes: the underlying set remains S(2) and the Combination Theorem applies.



(a) Centre=$(\frac{1}{2}, \frac{1}{2})$

(b) Centre=(0.95,0.05)

Figure 3: Contractions of S(2)

By varying α from 0 to 1 we can construct a diagram showing the transition of the likeliness of any given integram from the Multinomial Theorem to the Combination Theorem. This is shown in Figure 3a for the integrams (2,0), (1,1), (0,2); that is, for all possible results of two tosses.

We do not have to choose the fair coin as the centre of the contraction. Figure 3b shows the transitions with q = (0.95, 0.05).

n	Row sum				
0	1	1			
1	2	1	1		
2	4	1	2	1	
3	8	1	3	3	1

(a) Binomial Theorem

n	Row sum				
0	1	1			
1	2	1	1		
2	3	1	1	1	
3	4	1	1	1	1

(b) Combination Theorem

Table 1: Representations of likelinesses over (a)$\{(\frac{1}{2},\frac{1}{2})\}$, (b) S(2)

No. of tosses, n=	0	1	2	3	4	5	6	7
$L_P(n''H'')$	1	0.50	0.33	0.25	0.20	0.17	0.14	0.12
$L_P(''H''\|n''H'')$	0.50	0.67	0.75	0.80	0.83	0.86	0.88	0.89

(a) P=S(2)

No. of tosses, n=	0	1	2	3	4	5	6	7
$Pr(n''H''\|f)$	1	0.50	0.25	0.12	0.06	0.03	0.01	
$Pr(''H''\|f,n''H'')$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

(b) $P=\{f\}=\{(\frac{1}{2},\frac{1}{2})\}$

Table 2: Repeated tosses of $''H''$

6 Discussion

The difference in meaning between 'I estimate my next-door-neighbour's height to be $5'6''$ ' and 'My next-door-neighbour's height is $5'6''$ ' is clear. One is an estimate and the other a statement of fact.

The difference between the value of something and an estimate of its value is important and worth preserving. Yet probabilities and likelinesses (which are best-estimates (means) of probabilities) are easily confused. Although this confusion is certainly not helped by the terminology associated with the Kolmogorov approach, it in fact goes back to the earliest days of probability theory, where Laplace [6], for example, specifically referred to the Law of Succession as giving the probability.

It is not too difficult to see why this confusion happens:-

- Historically-important examples of probabilities such as fair coins, unbiased dice, well-shuffled packs of cards, thoroughly-mixed urns of balls, are all of a case for which the concepts of probability and expected value of probability (likeliness) are identical: that of a singleton underlying set.

- Both probabilities and expected values of probabilities can be -and are- used as measures of the informal concept of 'how likely something is to happen'.

- Kolmogorov defined 'probability' in such a wide way that it covers Likelinesses. This makes it impossible to say things such as 'Probabilities are a special case of Likelinesses'. His mathematics was fine, but -from the terminological point of view- it would have been better if he had used another word, rather than the already-existing 'probability'. Retaining the word 'probability' has made it virtually impossible to have a sensible discussion about the differences between a value and an estimate of that value.

- If we were presented with the ordered pair $(0.5, 0.5)$ then how could we tell whether it gave the probabilities of the two faces of a coin or the best-estimates of their probabilities?

One example of such confusion is the use of Uniformity by firstly taking the mean over S(N) and then substituting it into the Multinomial Theorem. Here, Theorem 3 tells us that the use of the Multinomial Theorem would be justified if the underlying set were singleton, whereas the underlying set is S(N) - which is not singleton.

A similar example is the 'Perfect Cube Factory' where the nature of uniformity as the mean over a non-singleton set is ignored, and it is incorrectly treated as if it were the actual distribution rather than an estimate of it.

The question of when Likelinesses 'obey' the Multinomial Theorem, formalised here as Multinomial Consistency, is intrinsically about when mean values (Likelinesses) are mapped to mean values by the taking of powers, ie. is concerned with asking when **the mean value of** $[\boldsymbol{x^n}]$ is equal to **[the mean value of x]**$^{\boldsymbol{n}}$, that is, when $\boldsymbol{\Sigma_P\ x^n}$ is equal to $[\boldsymbol{\Sigma_P\ x}]^{\boldsymbol{n}}$. It is not difficult to see that we have equality when n=0,1 or P is singleton.

The most important of these sufficient conditions for MC is that P be singleton. Outside of academic examples which specify the actual distribution to be used (such as saying that a coin is *fair*), the circumstances under which we have such a singleton underlying set are arguably non-existent. The closest that we can come to such a situation is perhaps when sampling from a known, finite population, although the question may still arise as to randomness in the form of uniformity.

This is where Johnson appears to have encountered the difficulties which eventually led him to abandon the Combination Postulate. In his interpretation of this Postulate, he refers explicitly to it as applying to probabilities, whereas it applies when the underlying set is S(N). Having made that error oof interpretation, he would then have had irresolvable difficulties with the Multinomial Theorem -as did iin fact happen.

At first glance, Table 1 may seem to be saying that uniformity can 'survive' mapping by the Multinomial Theorem. This is not the case. With the exception of the identical two topmost rows of Tables 1(a) and 1(b) (and their equivalents for other degrees), the Combination Theorem and the Multinomial Theorem are applicable under different conditions.

We cannot escape lightly from the question "Does or does not the distribution $(\frac{1}{2}, \frac{1}{2})$ give probabilities?". Because if it does then it may be used in conjunction with the Multinomial Theorem but if it does not then it may not.

The answer to this is "It all depends".

Tempting though it may be, the question is badly posed. "Giving" or "not giving" a probability is not a property of the distribution itself: it is a property of how that distribution was obtained.

- If the distribution was obtained by some form of averaging -for example- by an appeal to symmetry- then it is an average and so does not give probabilities and may not be used with the Multinomial Theorem.

- If the distribution was deduced from data then it is a likeliness but is not a probability -at least it is not a $Pr(g|f)$- since it is not independent of data, as all $Pr(g|f)$ must be. It is here that we encounter the linguistic weakness which has been forced upon us by the retention of the word 'probability' when the subject was being axiomatised, because we do use 'probability' to describe such things.

- If the distribution arose as a *given* then it is a probability, at least it is a $Pr(g|f)$. Consider the question *"A fair coin is tossed 20 times, and comes down Heads 15 times and Tails 5 times. What is the probability of Heads?"*. The answer is $\frac{1}{2}$ because the coin is fair. The given histogram (15,5) has the status of being a fluke, nothing more: it does not affect anything; the coin is fair, and that is an end to the matter. The underlying set is the singleton $\{(\frac{1}{2}, \frac{1}{2})\}$, which is the defining property of a probability, and the result is independent of data, which is a necessary property.

In practice, even when the Multinomial Theorem should not be used it often is. When this is done, then it is an approximation, and the question arises as to how good an approximation it is. A minted coin is a good approximation to the fair coin, but a minted coin which has had a lump of lead fixed to one face will be less good.

The need for such approximations is because of simplicity. Except in a few standard cases, the practical calculation of likelinesses has to be numerical, involving sampling from the underlying set.

To summarise the standard cases considered here:-

- The Multinomial Theorem gives the likeliness of g given h when the underlying set is singleton.
- The Law of Succession gives the likeliness, over S(N), of g given h when $\omega(g) = 1$.
- The Combination Theorem gives the likeliness, over S(N), of g given h when $h = \underline{0}$.
- The Integram Theorem gives the likeliness, over S(N), of g given h when g & h are both integrams.

These represent simple situations which are very useful in theoretical calculations but of limited (which is not a polite way of saying 'no') use in practice -for example, because the underlying set will commonly be neither singleton nor S(N).

Although some of the simpler likelinesses, such as these, are easily calculated, analysis using more complicated underlying sets is not simple, and usually requires numerical techniques.

Until fairly recently, the scale of those numerical techniques had usually placed such analyses beyond practical use, so the subject has tended to be out of favour. Computer technology has, however, been improving so that it is now at a level where practical implementation is possible. For example, using the Author's own program (freely downloadable from his website), each point in Figure 3 involved the taking of a sample of 200,000 distributions from the underlying set, with a run-time of less than 10 seconds.

Nonetheless, the limits on double-precision storage have placed severe limitations on the size of problems which could be addressed: with the exception of the simpler examples, data sets have been limited to the order of 100 or so observations. Recent changes to the way FORTAN handles double-precision numbers, combined with the move to 64-bit PCs, has changed this to such an extent that most practical problems could (in principle) now be addressed, with data sets of thousands of observations.

References

[1] Good,I.J., 1965, 'The Estimation of Probabilities: An essay on modern Bayesian Methods', Research Monograph 30, The M.I.T. Press, Cambridge,Massachusetts

[2] Hájek, Alan, 2012, 'Interpretation of Probability', The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalter (ed.), URL=http://plato.stanford.edu/archives/win2012/entries/probability-interpret/

[3] Johnson,W.E., 1924, 'LOGIC Part III The Logical Foundations of Science', Cambridge University Press//

readable online at http://www.archive.org/stream/logic03john#page/182/mode/2up, accessible via http://tinyurl.com/logicWEJ

[4] Johnson, W.E., 1932, 'Probability: The Deductive and Inductive Problems', Mind, New Series, 41, No.164, October 1932, pp 409-423, Editor R.B.Braithwaite.

[5] Keynes,John Maynard, 1920, 'A Treatise on Probability', Wildside Press LLC

[6] Laplace, Pierre-Simon, 1814 'Essai philosophique sur les probabilits'. Paris

[7] Mandelbrot, B.B., 1982 'The fractal geometry of nature', W.H.Freeman and Company

[8] Moscati, Ivan, 2005, 'W.E.Johnson's 1913 Paper and the Question of his Knowledge of Pareto', J.History of Economic Thought,27(03), September 2005, pp 283-304.

[9] Pareto,V., 1986, 'Cours deconomie politique', Droz, Geneva Switzerland

[10] Whittle, Peter, 1992, 'Probability via Expectation [3rd Edition]', Springer-Verlag New York

[11] Wikipedia contributors,2014, 'Principal of Indifference', Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Principle_of_indifference&oldid=628117267 (accessed 07 April 2015)

[12] Zabell, S.L., 2005, 'Symmetry and its discontents: essays on the history of inductive probability', Cambridge Studies in Probability, Induction and Decision Theory, August 2005, Cambridge University Press

[13] Zipf, G.K., 1949, 'Human Behaviour and the Principle of Least Effort', Addison-Wesley