mathematics

MDPI

Article

On Surprise Indices Related to Univariate Discrete and Continuous Distributions: A Survey

Indranil Ghosh * and Tamara D. H. Cooper

Department of Mathematics and Statistics, University of North Carolina, Wilmington, NC 28403, USA; coopert@uncw.edu

* Correspondence: ghoshi@uncw.edu

Abstract: The notion that the occurrence of an event is surprising has been discussed in the literature without adequate details. By definition, a surprise index is an index by which how surprising an event is may be determined. Since its inception, this index has been evaluated for univariate discrete probability models, such as the binomial, negative binomial, and Poisson probability distributions. In this article, we derive and discuss using numerical studies, in addition to the above-mentioned probability models, surprise indices for several other univariate discrete probability models, such as the zero-truncated Poisson, geometric, Hermite, and Skellam distributions, by adopting a established strategy and using the `Mathematica, version 12` software. In addition, we provide symbolical expressions for the surprise index for several univariate continuous probability models, which has not been previously discussed. For illustrative purposes, we present some possible real-life applications of this index and potential challenges to extending the notion of the surprise index to bivariate and higher dimensions, which might involve ubiquitous normalizing constants.

Keywords: discrete distributions; continuous univariate distributions; surprise index

MSC: 62G05; 62G32

Citation: Ghosh, I.; Cooper, T.D.H. On Surprise Indices Related to Univariate Discrete and Continuous Distributions: A Survey. *Mathematics* **2023**, *11*, 3234. https://doi.org/10.3390/math11143234

Academic Editor: Joseph Ngatchou-Wandji

Received: 21 June 2023
Revised: 15 July 2023
Accepted: 17 July 2023
Published: 23 July 2023

1. Introduction

The notion of the surprise index (SI) is not new in the literature, but has not been discussed thoroughly due to a lack of applicability and complexity in deriving for probability models that do not conform to well-known generating functions. The scarcity of scholarly works in this direction is reminiscent of this fact. The earliest reference dates back to 1948, when [1] asserted that an event with a low probability may be rare but is not surprising.

Interestingly enough, research on this topic is very limited. Some pertinent references are given as follows: Ref. [2] generalized and derived the SI for the multivariate normal distribution, but it has a different expression and notion. Ref. [3] derived the SIs for the binomial and Poisson distributions but without adequate details. Ref. [4] discussed the SI for the negative binomial distribution. Ref. [5] discussed the role of the SI in the context of macro-surprises from a monetary economics perspective. From the above-cited references, one may arrive at the conclusion that finding the SI is difficult to achieve analytically and subsequently requires the assistance of a powerful and efficient computing environment, such as `Mathematica`, which is utilized in this paper to obtain closed-form expressions for probability distributions in both the discrete domain and in the continuous domain other than those that have already been discussed.

In this article, we aim to discuss, in adequate details, the computation of SIs for various discrete probability distributions, including the binomial, negative binomial, and Poisson distributions (i.e., those that have been at least discussed in the literature), and SIs for zero-truncated Poisson, geometric, Hermite, and Skellam distributions, which are new contributions to the current topic. In addition, we also provide an analogous expression

for deriving the SI for univariate continuous probability models using the definition given in Equation (19) and defined later. For illustrative purposes, we compute SIs for various well-known univariate absolutely continuous probability models using Equation (19). It appears that, in most of the cases, the resulting expression of the SI associated with each of the discrete probability distributions is available in closed form, involving special functions and infinite series wherever applicable. Furthermore, we provide some empirical studies of the SIs corresponding to several discrete probability models. We conjecture that a similar development can be made in terms of identifying SIs for bivariate and/or multivariate continuous probability models, which will be the subject matter of a separate article. In summary, the major contributions of this article on the topic of SIs can be summarized as follows:

- We revisit the computation of the SIs for the binomial, Poisson, and negative binomial distributions and provide the correct expression of the SI for the Poisson distribution using `Mathematica`.
- Surprise indices are computed for the geometric and negative binomial, zero-truncated Poisson, and Hermite (for which closed-form expressions involving special functions and/or infinite series are available) distributions, while for the generalized Poisson distribution, the associated SI is not available in closed form, and a numerical solution is to be searched for. All of these derivations are new contributions to this topic.
- In addition, we provide the derivation of SIs for univariate continuous probability models using an analogous expression based on the geometric mean of a random variable.
- Finally, we conduct empirical studies on SIs for several of the discrete distributions with varying parameter choices, and several useful observations are derived accordingly.

The remainder of this article is organized in the following manner: In Section 2, we provide the computational details of deriving the SI for each of the univariate discrete probability models assumed in this paper with empirical studies on several of such probability models. In Section 3, we derive the SI for a continuous probability model based on the definition according to [2] and provide some useful conjectures on the properties of SIs. Section 4 presents several potential applications of the SI in a practical setting along with some potential challenges to extending this definition in bivariate and higher domains. Finally, some concluding remarks are presented in Section 5.

2. Surprise Index Derivation: Preliminaries

We begin this section by providing the definition of SI. According to [1], the SI, S_i, is defined as the comparison of the expected probability and the observed probability, which has the following form:

$$S_i = \frac{\sum p_m^2}{p_i}, \tag{1}$$

where $p_m = P(X = m)$ and p_i represents the probability that an event E_i has actually occurred. The expression in Equation (1) of the SI is from [3]. This feature can be obtained for discrete probability distributions in computing their corresponding probability generating functions, a strategy which is discussed later. Based on a suggestion by an anonymous reviewer, alternatively, Equation (1) can be rewritten as

$$S_i = \frac{E(p_X)}{p_i}.$$

Noticeably, this form is also independently obtained in [1].

Next, we revisit the computation of the SIs for the binomial, negative binomial, and Poisson distributions that have been independently discussed and derived in [3,4]. Proceeding in the same manner, we derive SIs for the zero-truncated Poisson, geometric, Hermite, and Skellam distributions. The process of obtaining the SI involves the following steps (for details, see [3]):

- Step 1: Calculate the generating function of p_m, which is of the form $\sum_{m\geq 0} p_m x^m$ from a given probability mass function (p.m.f.).
- Step 2: Set $x = e^{i\theta}$, and $e^{-i\theta}$, to obtain the following quantity $\sum_{m\geq 0} p_m^2 = \sum_{m\geq 0} p_m e^{-im\theta}$ $p_m e^{im\theta}$, which is the numerator of Equation (1), where $i = \sqrt{-1}$.
- Step 3: Integrate the simplified quantity on the R.H.S. obtained in Step 2, from 0 to 2π.

Then, substitute the value obtained in Step 3 to the numerator of Equation (1). Observe that, since the rationale behind this strategy of obtaining the SI has already been discussed in [3], it is not discussed here.

Next, this simple process is carried out below, for each of the discrete probability distributions selected for this purpose. It is important to note that the goal of the above steps is to obtain an expression for the sum of p_m^2, which involves solving the integral in step 3. In the next subsection, we begin by revisiting the SI for a binomial distribution at first.

2.1. Surprise Index for a Binomial Distribution

The binomial distribution is denoted as $B(n,\ p)$, with $n \in \{0,\ 1,\ 2, \ldots\}$ being the number of trials and $p \in [0,\ 1]$ being the probability of success resulting from each trial. The associated probability mass function (p.m.f.) is

$$p_m = \binom{n}{m} p^m q^{n-m},$$

where $m \in \{0,\ 1,\ 2,\ \ldots,\ n\}$ is the number of successes, with $p + q = 1$. The associated generating function will be

$$\sum_{m=0}^{n} p_m x^m = (q + px)^n.$$

Then, following steps two and three (given earlier) and simplifying, we obtain

$$\begin{aligned} \sum_{m=0}^{n} p_m^2 &= \sum_{m=0}^{n} (p_m \exp(-im\theta))(p_m \exp(im\theta)) \\ &= \frac{1}{2\pi} \int_0^{2\pi} (q^2 + 2qp\cos(\theta) + p^2)^n \, d\theta \\ &= (p-q)^{2n} \, {}_2F_1\left(\frac{1}{2}, -n; 1; -\frac{4pq}{(p-q)^2}\right), \quad \text{on using Mathematica,} \end{aligned} \tag{2}$$

where

$${}_2F_1(a, b; c, d) = \frac{(a)_n (b)_n d^n}{n!(c)_n}, \tag{3}$$

is the Gauss hypergeometric function, and $(W)_n = W(W+1)(W+2)\ldots(W+n-1)$ if $n > 0$ and $(W)_n = 1$ if $n = 0$.

Therefore, the SI for the binomial distribution related to the i-th probability is (on substituting Equation (2) in the numerator of Equation (1)):

$$S_i = \frac{(p-q)^{2n} \, {}_2F_1\left(\frac{1}{2}, -n; 1; -\frac{4pq}{(p-q)^2}\right)}{p_i}. \tag{4}$$

For illustrative purposes, we assume some representative values of p and subsequently compute the associated values of S_i for a fixed value of $n = 10$ and for varying choices m, p, and q in Equation (3), which are reported in Table 1.

Table 1. Surprise index values for binomial distribution for various choices of m, p, and q.

n	m	p	q	p_i	S_i
10	1	0.01	0.99	0.0914	9.04
10	3	0.01	0.99	0.0001	7387.44
10	5	0.01	0.99	0.00803	34,478,242.41
10	8	0.01	0.99	4.41×10^{-15}	1.87×10^{14}
10	10	0.01	0.99	1.00×10^{-20}	8.26×10^{19}
10	1	0.25	0.75	0.1877	1.09
10	3	0.25	0.75	0.2503	0.82
10	5	0.25	0.75	0.0584	3.52
10	8	0.25	0.75	0.0004	531.61
10	10	0.25	0.75	0.000001	215,301.13
10	1	0.8	0.2	0.000004	54,639.75
10	3	0.8	0.2	0.0008	284.58
10	5	0.8	0.2	0.0264	8.47
10	8	0.8	0.2	0.30199	0.74
10	10	0.8	0.2	0.1074	2.08

From Table 1, we can observe the following:

- For fixed n, with p_i decreasing, the corresponding SI values increase, which is expected.
- For fixed values of p and q, as the number of successes increase and with p_i decreasing, the SI values increase.

2.2. Surprise Index for a Negative Binomial Distribution

The negative binomial distribution is denoted as $NB(r,\ p)$, with $r > 0$ as the number of successes until the experiment is terminated and $p \in [0,\ 1]$ being the probability of success for each experiment. The associated p.m.f. is

$$p_m = \binom{m+r-1}{m} p^r q^m,$$

where $m \in \{0,\ 1,\ 2,\ \ldots\}$ is the number of failures. Consequently, the generating function will be

$$\sum p_m x^m = \left(\frac{p}{1-qx}\right)^r.$$

Proceeding as before, we obtain

$$\begin{aligned} \sum_{m=0}^{n} p_m^2 &= \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{p^2}{(q^2 - 2q\cos(\theta) + 1)}\right)^r d\theta \\ &= p^2(q+1)^{-2r}\, {}_2F_1\left(\frac{1}{2}, r; 1; \frac{4q}{(q+1)^2}\right), \end{aligned} \tag{5}$$

using `Mathematica`, where ${}_2F_1()$ is defined in Equation (3).

Thus, the SI for the negative binomial distribution is, on substituting Equation (5) in the numerator of Equation (1),

$$S_i = \frac{p^2(q+1)^{-2r}\,{}_2F_1\left(\frac{1}{2}, r; 1; \frac{4q}{(q+1)^2}\right)}{p_i}. \tag{6}$$

Assuming several representative values of p and q, and substituting various values for $r$and p_i in Equation (6), we find the following values of S_i for this distribution, which are presented in Table 2.

Table 2. Surprise index values for negative binomial distribution for various choices of r, m, and p.

n	m	p	q	p_i	S_i
1	9	0.01	0.99	0.0091	0.55
3	7	0.01	0.99	0.00003	5,616,123,374.28
5	5	0.01	0.99	0.00000001	1.15×10^{21}
8	2	0.01	0.99	3.53×10^{-15}	2.98×10^{39}
10	0	0.01	0.99	1.00×10^{-20}	9.32×10^{52}
1	9	0.25	0.75	0.0188	7.61
3	7	0.25	0.75	0.0751	185.21
5	5	0.25	0.75	0.0292	88,714.07
8	2	0.25	0.75	0.0003	2.63×10^{10}
10	0	0.25	0.75	0.000001	1.93×10^{15}
1	9	0.5	0.5	0.0010	341.33
3	7	0.5	0.5	0.0352	61.81
5	5	0.5	0.5	0.1230	203.05
8	2	0.5	0.5	0.0352	34,684.81
10	0	0.5	0.5	0.0010	17,668,300.52

From Table 2, one may observe the following:

- The SI values are dependent on the magnitude of either or both of p and p_i.
- For fixed p, q as p_i increases, the SI values decrease for varying r, m.
- For $r > m$, $p < q$, with q increasing, the SI value increases.
- For $r < m$, with $p < q$, and q decreasing, as m decreases, the SI values increase.

2.3. Surprise Index for a Poisson Distribution

The associated p.m.f. is

$$p_m = \frac{\lambda^m e^{-\lambda}}{m!},$$

where $m \in \{0,\ 1,\ 2, \ldots\}$ is the number of occurrences and $\lambda \in (0, \infty)$. The associated generating function is

$$\sum_{m \geq 0} p_m x^m = e^{-\lambda} e^{\lambda x}.$$

Proceeding as before,

$$\begin{aligned} \sum_{m=0}^{n} p_m^2 &= \frac{e^{-2\lambda}}{2\pi} \int_0^{2\pi} e^{2\lambda \cos(\theta)}\, d\theta \\ &= e^{-2\lambda} I_0(2\lambda), \end{aligned} \tag{7}$$

where $I_0()$ is the zero-order modified Bessel function of the first kind.

Therefore, the SI for the Poisson distribution on substituting Equation (7) in the numerator of Equation (1) is

$$S_i = \frac{e^{-2\lambda} I_0(2\lambda)}{p_i}. \tag{8}$$

Substituting various values for λ and m in Equation (8), we find the following values of S_i for this distribution, given in Table 3.

Table 3. Surprise index values for Poisson distribution for various choices of λ.

λ	m	p_i	S_i
0.5	1	0.3033	1.54
0.5	3	0.0126	36.86
0.5	5	0.0002	2948.77
0.5	8	0.00000006	7,926,282.59
0.5	10	1.63×10^{-10}	2,853,461,732.66
1	1	0.3679	0.84
1	3	0.0613	5.03
1	5	0.0031	100.63
1	8	9,123,994.08	33,812.86
1	10	0.0000001	3,043,157.28
2.5	1	0.2052	0.89
2.5	3	0.2138	0.86
2.5	5	0.0668	2.75
2.5	8	0.0031	59.08
2.5	10	0.0002	850.81

From Table 3, it appears that

- For a fixed λ, with m increasing and p_i decreasing, the SI values increase.
- For a fixed m, with λ increasing, the SI values decrease.

For a comprehensive view of the SI in this case, further empirical studies are required.

2.4. Surprise Index for a Zero-Truncated Poisson Distribution

The zero-truncated Poisson distribution is denoted as ZTP(λ) with parameter $\lambda \in (0, \infty)$. The p.m.f. is

$$p_m = \frac{e^{-\lambda}\left(\frac{\lambda^m}{m!}\right)}{1 - e^{-\lambda}} = \frac{\lambda^m}{(e^{\lambda} - 1)m!},$$

where $m \in \{1, 2, 3, \ldots\}$ is the number of occurrences; for a detailed study on this distribution, see [6]. The associated generating function will be

$$\sum_{m=1}^{\infty} p_m x^m = \frac{e^{\lambda x}}{e^{\lambda} - 1}.$$

Proceeding as before, the numerator of Equation (1) in this case, will be

$$\sum_{m=0}^{\infty} p_m^2 = \frac{1}{2\pi(e^{2\lambda} - 2e^{\lambda} + 1)} \int_0^{2\pi} e^{2\lambda\cos(\theta)}\, d\theta$$
$$= (e^{2\lambda} - 2e^{\lambda} + 1)^{-1} I_0(2\lambda), \quad (9)$$

where $I_0()$ has been defined earlier in the previous subsection. Therefore, upon substituting Equation (9) in the numerator of Equation (1), the SI for the zero-truncated Poisson distribution will be

$$S_i = \frac{(e^{2\lambda} - 2e^{\lambda} + 1)^{-1} I_0(2\lambda)}{p_i}. \quad (10)$$

Substituting various representative values for λ and m in Equation (10), we find the following values of S_i for this distribution, which is presented in Table 4.

Table 4. Surprise index values for zero-truncated Poisson distribution for various choices of λ.

λ	m	p_i	S_i
0.5	1	0.7707	0.79
0.5	3	0.0321	19.05
0.5	5	0.0004	1523.94
0.5	8	0.0000001	4,096,347.51
0.5	10	4.14838×10^{-10}	1,474,685,102.05
1	1	0.5820	0.69
1	3	0.0970	4.14
1	5	0.0048	82.89
1	8	0.00001	27,850.21
1	10	0.0000002	2,506,518.52
2.5	1	0.2236	0.89
2.5	3	0.2329	0.85
2.5	5	0.0728	2.73
2.5	8	0.0034	58.65
2.5	10	0.0002	844.61

From Table 4, one can observe the following:

- The SI values are slightly different from the Poisson distribution's SI values. Also, we see that smaller values of λ generate greater differences between the zero-truncated Poisson and the Poisson SI values.
- The behavior/changing pattern of the SI values are exactly the same (except for the magnitude) as in the previous case (Poisson distribution), for varying choices of λ, m and p_i.

2.5. Surprise Index for a Geometric Distribution

The geometric distribution is denoted as Geo(p), with $p \in \{1, 2, 3, \ldots\}$ being the number of Bernoulli trials needed to achieve one success. The associated p.m.f. is

$$p_m = (1-p)^{m-1} p = pq^{m-1},$$

where $m \in \{1, 2, 3, \ldots\}$ is the number of successes. The generating function is then found to be

$$\sum p_m x^m = \frac{px}{1-qx}.$$

Consequently, the numerator of Equation (1) will be

$$\sum_{m=0}^{\infty} p_m^2 = \frac{p^2}{2q^2\pi} \int_0^{2\pi} (q^2 - 2q\cos(\theta) + 1)^{-1} \, d\theta$$
$$= \frac{p^2}{q^2(1-q^2)}, \tag{11}$$

on using `Mathematica`.

Hence, on substituting Equation (11) in the numerator of Equation (1), we have the following expression for the SI for the geometric distribution:

$$S_i = \frac{p^2}{p_i q^2 (1-q^2)}. \tag{12}$$

Assuming various representative values for m and p, in Equation (12), we find the following values of S_i for this distribution, which are given in Table 5.

Table 5. Surprise index values for the Geometric distribution for various choices of p.

m	p	q	p_i	S_i
1	0.01	0.99	0.01	0.51
5	0.01	0.99	0.0096	0.53
10	0.01	0.99	0.00914	0.56
20	0.01	0.99	0.0083	0.62
50	0.01	0.99	0.0061	0.84
1	0.25	0.75	0.25	1.02
5	0.25	0.75	0.0791	3.21
10	0.25	0.75	0.0188	13.53
20	0.25	0.75	0.0011	240.26
50	0.25	0.75	0.0000002	1,345,356.92
1	0.8	0.2	0.8	20.83
5	0.8	0.2	0.0013	13,020.83
10	0.8	0.2	0.0000004	40,690,104.16
20	0.8	0.2	4.19×10^{-14}	3.97×10^{14}
50	0.8	0.2	4.50×10^{-35}	3.70×10^{35}

From Table 5, one may observe the following:

- For fixed p, q with $p < q$ and with m increasing, the SI values exhibit an increasing pattern.
- For fixed m, with q decreasing, the SI values increase.

2.6. Surprise Index for a Hermite Distribution

The Hermite distribution is denoted as Herm(a_1, a_2) with parameters $a_1 \geq 0$ and $a_2 \geq 0$. This distribution is used to measure count data using more than one parameter and has been used in biological research. There are several scholarly studies related to this distribution that exist in the literature. For example, Ref. [7] discussed several useful structural properties of the Hermite distribution and they established the fact that this distribution is the generalized Poisson distribution. Ref. [8] have discussed the utility of this distribution in the context of a zero-inflated overdispersed probability model. Ref. [9]

developed an R package `hermite` to apply generalized hermite distribution in modeling real-world scenario(s) of fitting count data in the presence of overdispersion or multimodality with a lot more added flexibility in terms of inference under the classical method. The associated p.m.f. of the random variable $Y = X_1 + X_2$ is

$$p_m = e^{-(a_1+a_2)} \sum_{j=0}^{\lfloor m/2 \rfloor} \frac{a_1^{m-2j} a_2^j}{(m-2j)!j!}.$$

where $m = 0, 1, 2, \ldots$ and $\lfloor m/2 \rfloor$ is the integer part of $m/2$, and $a_1, a_2 \geq 0$ are the parameters associated with the two independent Poisson variables $X_1$and X_2, respectively. The associated generating function is given by

$$\sum_{m=0}^{\lfloor m/2 \rfloor} p_m x^m = e^{a_1(x-1)+a_2(x^2-1)}.$$

Proceeding as before, the numerator of Equation (1) will be

$$\begin{aligned}
\sum_{m=0}^{n} p_m^2 &= \frac{1}{2\pi} \int_0^{2\pi} e^{2a_1(cos(\theta)-1)+2a_2(cos(\theta)-1)}\, d\theta \\
&= \frac{1}{2\pi} \sum_{j=0}^{\infty} \frac{1}{j!} \int_0^{2\pi} [2a_1(cos(\theta)-1) + 2a_2(cos(\theta)-1)]^j\, d\theta \\
&= \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{(-a_1)^j \Gamma(2j+1) \left(\frac{a_1}{a_1+4a_2}\right)^{-j} {}_2\tilde{F}_1\left(-j, j+\frac{1}{2}; j+1; \frac{4a_2}{a_1+4a_2}\right)}{\Gamma(j+1)} \right),
\end{aligned} \tag{13}$$

where ${}_2\tilde{F}_1()$ is the regularized hypergeometric distribution, obtained using `Mathematica`. Therefore, upon substituting Equation (17) in the numerator of Equation (1), the SI for the Hermite distribution will be

$$S_i = \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{(-a_1)^j \Gamma(2j+1) \left(\frac{a_1}{a_1+4a_2}\right)^{-j} {}_2\tilde{F}_1\left(-j, j+\frac{1}{2}; j+1; \frac{4a_2}{a_1+4a_2}\right)}{p_i \Gamma(j+1)} \right). \tag{14}$$

Substituting various values for m, a_1, anda_2 in Equation (14), one can find values of S_i for this distribution, which is not reported in this paper for brevity. Also, it is quite difficult to obtain numerically, as it involves infinite sums and special functions.

2.7. Surprise Index for a Skellam Distribution

The Skellam distribution, also known as the Poisson difference distribution, is derived from the difference of two Poisson random variables (for details, see [10]) and is denoted as Skellam(μ_1, μ_2) with parameters $\mu_1 \geq 0$ and $\mu_2 \geq 0$. This distribution may be used for describing the point spread distribution for sports such as hockey, where all points scored are equal, describing the statistics of the difference of two images with simple photon noise, or studying treatment effects, as discussed in [10]. The p.m.f. when considering two Poisson random variables is given by

$$p_m = e^{-(\mu_1+\mu_2)} \left(\frac{\mu_1}{\mu_2}\right)^{m/2} I_m(2\sqrt{\mu_1\mu_2}),$$

where m is an integer and $I_m(z)$ is the m-th order modified Bessel function of the first kind. The associated generating function will be

$$\sum p_m x^m = e^{-(\mu_1+\mu_2)+\mu_1 m+\mu_2/m}.$$

Again, by proceeding as before, the numerator of Equation (1) can be derived using the infinite series expression for the exponential function and using `Mathematica`, as follows:

$$\begin{aligned}\sum_{m=0}^{n} p_m^2 &= \frac{1}{2\pi}\int_0^{2\pi} e^{(2\mu_1+2\mu_2)(cos(\theta)-1)}\,d\theta \\ &= \frac{1}{2\pi}\sum_{j=0}^{\infty}\frac{(2(\mu_1+\mu_2))^j}{j!}\int_0^{2\pi}(cos(\theta)-1)^j d\theta \\ &= \sum_{j=0}^{\infty}\frac{(2(\mu_1+\mu_2))^j}{j!}\left(\frac{(-2)^j\Gamma\left(j+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma(j+1)}\right),\end{aligned} \tag{15}$$

Subsequently, upon substituting Equation (15) in the numerator of Equation (1), the SI for the Skellam distribution can be written as

$$S_i = \sum_{j=0}^{\infty}\frac{(2(\mu_1+\mu_2))^j}{j!}\left(\frac{(-2)^j\Gamma\left(j+\frac{1}{2}\right)}{p_i\sqrt{\pi}\Gamma(j+1)}\right). \tag{16}$$

Substituting various values for m, μ_1, and μ_2 in Equation (16), one can find expressions of the SI for this distribution. However, from Equation (16), it is clear that it would be difficult to obtain numerical values as the expression involves infinite sum and gamma functions.

2.8. Surprise Index for a Generalized Poisson Distribution

The generalized Poisson distribution is denoted as GDP$(\theta,\ \lambda)$ with parameters θ and λ, $0 \le \lambda < 1$ and $\theta > 0$. To allow us to differentiate between the parameter and the integration variable, we change θ to α, and then, the p.m.f. is

$$p_m = \frac{\alpha(\alpha+n\lambda)^{m-1}e^{-m\lambda-\alpha}}{m!},$$

where $m \in \{0,\ 1,\ 2,\ \ldots\}$ is the number of occurrences. The associated generating function is then, according to [11],

$$\sum p_m x^m = \exp\left\{-\frac{\alpha}{\lambda}\mathrm{W}(-\lambda x\exp[-\lambda]) + \lambda\right\},$$

where $\mathrm{W}(\cdot)$ is the Lambert W function. Continuing with the prescribed process, we found the following integral form:

$$\sum p_m^2 = \frac{1}{2\pi}\int_0^{2\pi}\exp\left\{-\frac{\alpha}{\lambda}\Big(\mathrm{W}(-\lambda\exp[i\theta]\exp[-\lambda]) + \mathrm{W}(-\lambda\exp[-i\theta]\exp[-\lambda])\Big) + 2\lambda\right\}d\theta. \tag{17}$$

Consequently, the associated SI for a GPD, upon substituting Equation (17) in the numerator of Equation (1), will be

$$\begin{aligned}S_j = &\left(\frac{1}{2\pi}\int_0^{2\pi}\exp\left\{-\frac{\alpha}{\lambda}\Big(\mathrm{W}(-\lambda\exp[i\theta]\exp[-\lambda]) + \mathrm{W}(-\lambda\exp[-i\theta]\exp[-\lambda])\Big) + 2\lambda\right\}d\theta\right) \\ &\times\left(\frac{\alpha(\alpha+n\lambda)^{j-1}e^{-i\lambda-\alpha}}{j!}\right)^{-1}.\end{aligned} \tag{18}$$

Noticeably, from Equation (18), it can be observed that this integral is difficult to solve in order to obtain a closed and analytically tractable form because of the involvement of the Lambert W function which has both real and imaginary parts. Numerical methods must be adopted, which we have not considered for brevity.

In addition, for illustrative purposes, we have also provided graphs of the SI for several discrete probability distributions discussed in this section in Appendix B.

3. Surprise Index for Continuous Probability Models

For a continuous random variable (r.v.), the associated expression for the SI is given by [2] and has the following form:

$$\zeta = \frac{E(p^*|H)}{p},$$

where p^* is the r.v. that is the probability density function (p.d.f.) of the original r.v., p is a realization of p^*, and H is a simple statistical hypothesis. Equivalently, we may rewrite the definition as follows. Let X be a continuous random variable with density function $f()$. Then, for all $x \in S(X)$, the SI is given by

$$S_x = \frac{E[f(X)]}{f(x)}.$$

However, an alternative version which does involve the geometric expectation (it is termed as a generalization of the SI) is given by

$$\zeta_0 = \frac{GE(p^*)}{p} = \frac{\exp\Big(E(\log X)\Big)}{p}, \tag{19}$$

where GE stands for the geometric expectation which will be equivalently evaluated using $E(\log X)$. For computation of the SI for various continuous probability models, we use Equation (19). In Table 6, we provide the expression of Equation (19), which can be viewed as an expression of the SI (according to [2]) for various univariate absolute continuous distributions. The symbolic computations are all carried out using `Mathematica`.

Table 6. Surprise index expressions for several continuous probability models.

Distribution	Surprise Index
Uniform (a, b)	$\left(\frac{1}{(b-a)}\right)^{-1} \times \exp\left(\left(\frac{b^b}{a^a}\right)^{1/(b-a)} \times e^{-1}\right)$
Beta(a, b)	$\left(\exp\left((\Gamma[a]\Gamma[b](PolyGamma[0,a] - PolyGamma[0,a+b]))/\Gamma[a+b]\right)\right) \times \left(\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\right)^{-1}$
Beta (type-II)(α, β)	$\left(\exp\left(\frac{\Gamma(\alpha+1)\Gamma(\beta-1)\left(H_\alpha - H_{\beta-2}\right)}{\Gamma(\alpha+\beta)B(\alpha,\beta)}\right)\right) \times \left(\frac{B(\alpha,\beta)(1+x)^{\alpha+\beta}}{x^\alpha}\right)$
Pareto (type-II)	$\left(\exp\left(\psi^{(0)}(-\alpha) - \log(\sigma) + \gamma\right)\right) \times \left(\frac{\alpha}{\sigma}\left(1+\frac{x}{\sigma}\right)^{-(\alpha+1)}\right)^{-1}$
Gamma (α, β)	$\left[\left(\exp\left(\left(\frac{1}{\beta^\alpha\Gamma(\alpha)}\right) \times \left(-\beta^\alpha\Gamma(\alpha)\left(\log\left(\frac{1}{\beta}\right) - \psi^{(0)}(\alpha)\right)\right)\right)\right)\right] \times \left(\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}\exp(-\frac{x}{\beta})\right)^{-1}$
Weibull(k, λ)	$\left[\exp\left(-\frac{\log\left(\left(\frac{1}{\lambda}\right)^k\right)+\gamma}{k}\right)\right] \times \left(\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp(-(\frac{x}{\lambda})^k)\right)^{-1}$
Log-normal (μ, σ)	$\left(\exp(\mu)\right) \times \left(\frac{1}{x\sqrt{2\pi\sigma}}\exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right)\right)^{-1}$
Exponentiated-exponential (α, β)	$\left(\exp\left(\sum_{j=0}^{\infty}\binom{\alpha-1}{j}(-1)^j\left(-\frac{\alpha\lambda(\log((j+1)\lambda)+\gamma)}{j\lambda+\lambda}\right)\right)\right) \times \left(\alpha\lambda(1-\exp(-\lambda x)^{\alpha-1}\exp(-\lambda x)\right)^{-1}$

Note: For a Pareto (type-IV) distribution, the associated integral for the numerator of Equation (19) diverges.

From Table 6, one can make the following observations for fixed $X = x$:

- For uniform (a, b), and b increasing and a decreasing, the SI will increase.
- For Beta (a, b), as a increases and b increases, SI decreases. On the other hand, when both a and b increase, the SI increases.
- For Beta (type-II) (α, β), when both α, β increase, the SI will increase.
- For Pareto (type-II) distribution, because of the nature of the polygamma function as obtained from `Mathematica`, for any choices of the parameter α, regardless of the other permissible choices of the other two parameters, it is divergent and, therefore, it cannot be computed.
- For the Log-normal(μ, σ) distribution, as both μ and σ increase, the associated SI increases.
- For the Gamma(α, β) distribution—(i) when α is fixed, with β increasing, the SI will increase and (ii) with β fixed and α increasing, the SI will increase.
- For the Weibull(k, λ) distribution, the following can be observed:
  - For a fixed k as λ and γ increase, the SI will increase.
  - For a fixed γ as k and λ increase, the SI will increase.
  - For any choice of $\lambda < 1$ and decreasing with k increasing, for a fixed choice of γ, the corresponding SI will decrease.

Next, we make the following conjectures. The proofs seem obvious, but we leave this up to the reader.

- **Conjecture 1.** The SI, if available, uniquely determines a discrete and/or continuous probability distribution.
- **Conjecture 2.** The SI for a truncated model differs only by a scalar quantity (involving model parameter(s)) corresponding to the non-truncated version of the assumed discrete probability model and is bigger than the SI computed for the non-truncated version. For example, the authors of [6] have shown that the SI for the truncated Poisson is bigger than that for the usual Poisson distribution.
- **Conjecture 3.** The SI is invariant under all non-singular linear transformations. Equivalently, we can state the following. Let X and Y be two non-degenerate random variables with valid probability distributions that are well-defined on $\mathbb{R}$. Further, let $Y = aX + b$, with $a \neq 0$, and $b \in (-\infty, \infty)$, and let SI_X and SI_Y be the surprise indices for the r.v. X and Y, respectively. Then, $SI_Y = aSI_X + b$.

  Proof. The result follows immediately by using the invariance property of a generating function. We provide the proof for a discrete r.v.; however, a similar approach can be made to establish the result for a continuous r.v. If $G_Y(s)$ and $G_X(s)$ are the probability generating functions of X and Y, respectively, then

$$\begin{aligned} G_Y(s) &= E\left[s^Y\right] \\ &= E\left[s^{aX+b}\right] \\ &= s^b E\left[(s^a)^X\right] \\ &= s^b G_X(s^a). \end{aligned}$$

  Hence, the proof. □

Note that in Appendix A, we provide the `Mathematica` codes for computing the SI for both univariate discrete and continuous probability models.

4. Potential Applications and Challenges/Open Problems

The use of Weaver's SI as an alternative to the use of tail area probabilities was suggested by [2]. Some applications of the SI have been presented such as determining if certain events are surprising; i.e., being dealt the same hand of cards consecutively in a

game of bridge [1] or a fair coin toss with edges of a particular size landing on its edge when flipped [1]. Although these applications are interesting, they are not particularly useful. For example, Ref. [4] suggests using the SI for outlier detection which we find intriguing since detecting outliers can be difficult, and by applying this feature to various data sets, we established the fact that it can be considered another tool for detecting outliers.

The Hermite distribution is used in the distribution of counts of bacteria in leucocytes. We assume that applying the surprise index for this distribution could be useful in determining that the counts of bacteria in white blood cells (leucocytes) are alarmingly high. This information could be helpful in choosing follow-up tests, determining diseases, or expediting patient care for patients who need urgent medical attention.

Several potential challenges in extending this definition in bivariate and higher domains might be summarized as follows:

(i) Ref. [2] states, "for multivariate normal distributions, $P(p^* < p)$, the distribution of the likelihood density, does not seem to be expressible in elementary terms" (p. 1133);
(ii) The special functions are difficult to determine for the univariate case, which leads to even more difficulty when more variables are considered;
(iii) The long runtimes when finding the closed-form expressions for several of such distributions suggest that a multivariate analysis of the SI will require highly efficient computing environments.

5. Concluding Remarks

In this article, we discuss with adequate details, the derivation of the SI for several univariate discrete probability distributions that had not been discussed earlier along with a re-evaluation of the surprise indices for the binomial, Poisson, and the geometric distributions. Using the `Mathematica` software, we obtain closed-form expressions for the SI for the binomial, negative binomial, and Poisson distributions including that of the zero-truncated Poisson, geometric, Hermite, and Skellam distributions involving either special functions and/or infinite sums or series. Also, we have computed the SI for univariate continuous probability models via an analogous expression (similar to the discrete case, but not exactly the same), which involves computing the geometric mean of a random variable. Extension to the bivariate and higher dimensions will be the topic of a separate article. However, the SI is not above criticism. For example, it is conjectured that in the definition of the SI, the numerator given in Equations (1) and (2) is arbitrary. Furthermore, the value of SI drastically changes when the results of an experiment are lumped together in a different way (discrete case) and/or there is a change in the values of stochastically independent r.v.s in the continuous case.

Author Contributions: Conceptualization, I.G.; Formal analysis, T.D.H.C. and I.G.; Investigation, T.D.H.C. and I.G.; Methodology, I.G. and T.D.H.C.; Supervision, I.G.; Writing—original draft, I.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

In this section, we provide the `Mathematica` codes for obtaining the **numerator** of Equation (1) of the surprise indices for several univariate discrete distributions and a couple of continuous distributions for illustrative purposes.

- **Binomial distribution (Equation (3) numerator) Integrate**$[(q^2 + 2qp\cos\theta + p^2)^n, \{\theta, 0, 2\pi\}]$**.**
- **Poisson distribution (Equation (7) numerator)**

  Integrate$[\exp\left(2\lambda \times \cos\theta\right), \{\theta, 0, 2\pi\}]$.

- **Negative binomial distribution (Equation (6) numerator numerator)**
  $\frac{p^2}{\pi}$Integrate$[(1-2q\cos\theta+q^2)^{-r},\{\theta,0,2\pi\}]$.
- **Geometric distribution (Equation (5) numerator)**
  $\frac{p^2}{q^2\pi}$Integrate$[(1-2q\cos\theta+q^2)^{-1},\{\theta,0,2\pi\}]$.
- **Pareto (type II) distribution (Table 6, row 4)**
  Integrate$[(1+\frac{x}{\sigma})^{\alpha-1}(\frac{\alpha}{\sigma})*\log[x],\{\theta,0,2\pi\}]$.
- **For a two parameter beta distribution (Table 6, row 2)**
  Integrate$[x^{a-1}*(1-x)^{b-1}*\log[x],\{\theta,0,2\pi\}]$.

Appendix B

In this section, we provide several graphs related to the SI for discrete distributions for illustrative purposes.

From these figures, one can make the following observation:

1. **Observations from Figure A1**:
   - For $p=0.01,0.25$ as m increases, the $\log(SI)$ value increases, i.e., equivalently, the SI values increase.
   - For $p=0.8$ as m increases, the $\log(SI)$ value decreases, i.e., equivalently, the SI values decrease.
2. **Observations from Figure A2**: For all fixed choices of p, as m increases, the $\log(SI)$ value increases, i.e., equivalently, the SI values increase.
3. **Observations from Figure A3**: For all fixed choices of λ, as m increases, the $\log(SI)$ value increases, i.e., equivalently, the SI values increase; however, the magnitude of increment decreases as λ becomes larger.
4. **Observations from Figure A4**: The pattern is almost similar to Figure A3.
5. **Observations from Figure A5**:
   - When $p=0.01$, $\log(SI)$ takes a constant value for all choices m.
   - For $p=0.25,0.8$, as m increases, the $\log(SI)$ value increases, i.e., equivalently, the SI values increase.

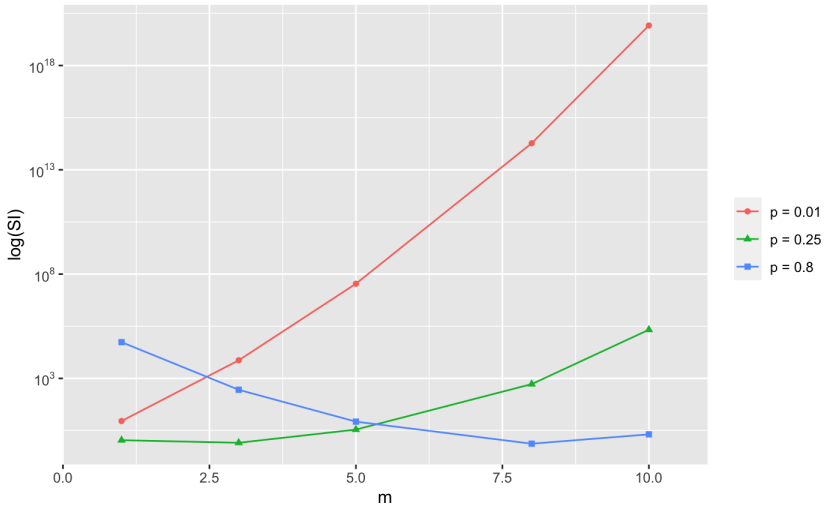

Figure A1. Surprise index values for binomial distribution, n = 10.

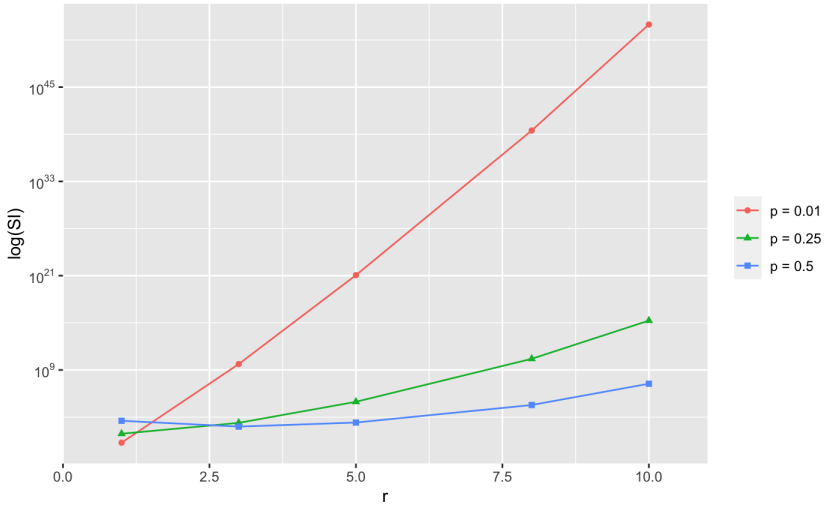

Figure A2. Surprise index values for negative binomial distribution, n = 10.

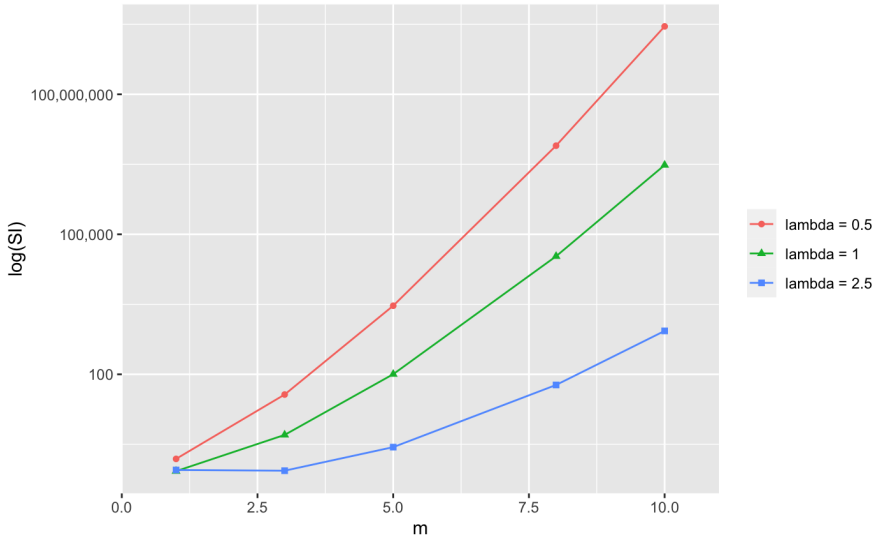

Figure A3. Surprise index values for Poisson distribution.

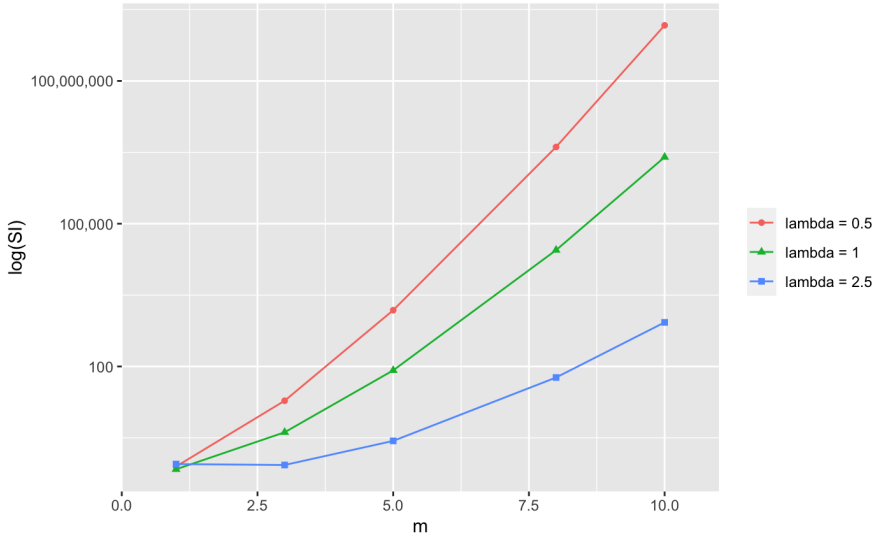

Figure A4. Surprise index values for zero-truncated Poisson distribution.

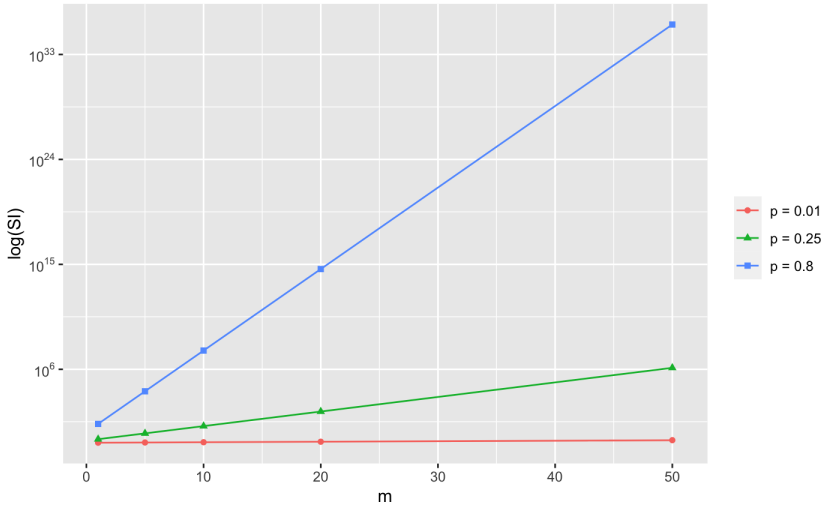

Figure A5. Surprise index values for geometric distribution.

References

1. Weaver, W. Probability, rarity, interest, and surprise. *Sci. Mon.* **1948**, *67*, 390–392. [CrossRef]
2. Good, I.J. The surprise index for the multivariate normal distribution. *Ann. Math. Stat.* **1956**, *27*, 1130–1135. [CrossRef]
3. Redheffer, R.M. A note on the surprise index. *Ann. Math. Stat.* **1951**, *22*, 128–130. [CrossRef]
4. Borja, M.C. Outliers in Long-Tailed Discrete Data. 2012. Available online: https://web-archive.lshtm.ac.uk/csm.lshtm.ac.uk/wp-content/uploads/sites/6/2016/04/Mario-Cortina-Borja-16-11-2012.pdf (accessed on 16 June 2023).
5. Scotti, C. Surprise and uncertainty indexes: Real-time aggregation of real-activity macro-surprises. *J. Monet. Econ.* **2016**, *82*, 1–19. [CrossRef]
6. David, F.N.; Johnson, N.L. The truncated poisson. *Biometrics* **1952**, *8*, 275–285. [CrossRef]
7. Kemp, C.D.; Kemp, A.W. Some properties of the 'Hermite' distribution. *Biometrika* **1965**, *52*, 381–394. [PubMed]
8. Kumar, S.C.; Ramachandran, R. On some aspects of a zero-inflated overdispersed model and its applications. *J. Appl. Stat.* **2020**, *47*, 506–523. [CrossRef] [PubMed]
9. Moriña, D.; Higueras, M.; Puig, P.; Oliveira Pérez, M. Generalized Hermite Distribution Modelling with the R Package Hermite. 2015. Available online: https://journal.r-project.org/archive/2015/RJ-2015-035/index.html (accessed on 22 June 2023).
10. Sellers, K.F. A distribution describing differences in count data containing common dispersion levels. *Adv. Appl. Stat. Sci.* **2012**, *7*, 35–46.
11. Vernic, R. A multivariate generalization of the generalized Poisson distribution. *ASTIN Bull. J. IAA* **2000**, *30*, 57–67. [CrossRef]