BULLETINOFTHEICA
Volume**102**(2024),60–70

Good's and Mallows's enumerations of hypotheses of independence: a new proof and a discussion

Petros Hadjicostas*

Abstract. We clarify the kinds of independence hypotheses among n variables that I.J. Good considered in a 1975 paper, and we give a new and more concise proof to his exponential generating function for the total number of such hypotheses. We also examine two questions posed in 1979 by C.L. Mallows about the total number of independence hypotheses among n variables without the restrictions imposed by Good.

1 Introduction

In 1975, mathematician and statistician I.J. Good [5] enumerated the number of different hypotheses of independence that exist among the random variables in a random vector $(X_1, \ldots, X_n)$, or equivalently, the number of different hypotheses of independence for a multidimensional contingency table that cross-classifies n categorical variables $(X_1, \ldots, X_n)$. By "independence", we mean either "unconditional independence" or "conditional independence".

If a_n is the total number of different hypotheses of independence that exist among the random variables in the vector $(X_1, \ldots, X_n)$, Good [5] proved that the exponential generating function (e.g.f.) of the numbers $(a_n : n \in \mathbb{Z}_{\geq 0})$ is

$$A(y) := \sum_{n=0}^{\infty} \frac{a_n}{n!} y^n = \exp(\exp(y) + 2y - 1) - \exp(3y). \tag{1}$$

*Corresponding author: peterhadji1@gmail.com
Key words and phrases: Bell numbers, conditional independence, enumeration, exponential generating function.
AMS (MOS) Subject Classifications: 05A15, 03B48, 60A05

Received: 8 July 2023
Accepted: 29 January 2024

Clearly, Eq. (1) is valid for all $y \in (-\infty, \infty)$.

Some values of the sequence $(a_n : n \in \mathbb{Z}_{\geq 0})$ are given in Table 1.1.

Table 1.1: Good's enumeration of hypotheses of independence.

n	0	1	2	3	4	5	6	7	8
a_n	0	0	1	10	70	431	2534	14820	88267

Unfortunately, the paper by Good [5] that contains the proof of Eq. (1) is difficult to find, and most people know of the e.g.f. $A(x)$ and the values in Table 1.1 through other papers and books that cite it; e.g., see Fienberg [4, p. 72], Good [6, p. 1171], and Good [7, p. 192].

In any case, in 2022, we were able to acquire the paper. In it, Good [5] actually proved that

$$a_n = \sum_{k=0}^{n} \binom{n}{k} B_k 2^{n-k} - 3^n, \tag{2}$$

where B_n is the n^{th} *Bell number*; see Comtet [1, Section 5.4].

In 2022, G.C. Greubel (after being provided by the author the original paper by Good [5]) noticed that Eq. (2) simplifies to

$$a_n = B_{n+2} - B_{n+1} - 3^n. \tag{3}$$

See sequence A005465 in the OEIS [9] for more values of a_n and for Greubel's formula (3) (and see sequence A000110 for the Bell numbers).

Mallows [8] disagreed with Good [5], and for the case $n = 3$, he produced 17 kinds of independence among 3 variables rather than $a_3 = 10$. Indeed, indirectly, Good [7, p. 192] admitted that Mallows was correct. (In addition, I.J. Good was the section editor of the journal where the short article by Mallows [8] appeared!)

In this paper, we clarify what kinds of independence among n variables Good [5] considered and give a new and more concise proof of his formula (1) for the e.g.f. $A(y)$. The proof and the formulas given in Section 2 of this paper were obtained by the author before acquiring the paper by Good [5] (and a version of the proof appeared in 2019 in the comments for sequence A005465 in the OEIS [9]).

In addition, we discuss Mallows's [8] additional hypotheses of independence among 3 variables, and we examine his two enumeration questions that he posed for the total number of modes of independence among n variables.

For Mallows's [8] second question about the total number of independence hypotheses among n variables that ignore conditional independence, we give a formula and the e.g.f. for the total number b_n of such hypotheses that were considered by Good [5]. See Eqs. (9) and (10) in Section 4 of the paper. (Eq. (9) in this paper is also Eq. (9) in Good [5].)

Note, however, that we do not solve either one of the two problems posed by Mallows [8], who asked us to remove Good's [5] restrictions. These problems are very difficult even in the case $n = 4$.

As a matter of notation, denote by $f(x_1, \ldots, x_n)$ the joint pdf or pmf (with respect to the Lebesgue measure or the counting measure in $\mathbb{R}^n$) of the random variables $X_1, \ldots, X_n$. If $\mathcal{A}$ is a subset of the set of the indices $\{1, 2, \ldots, n\}$ of the random variables $X_1, \ldots, X_n$, we denote by $f((x_i : i \in \mathcal{A}))$ the joint marginal pdf/pmf of the random variables $(X_i : i \in \mathcal{A})$ that we get from the joint pdf/pmf $f(x_1, \ldots, x_n)$.

In addition, if $\mathcal{A}$ and $\mathcal{B}$ are two disjoint subsets of $\{1, 2, \ldots, n\}$, we denote by

$$f((x_i : i \in \mathcal{A}) \,|\, (x_j : j \in \mathcal{B}))$$

the conditional joint pdf/pmf of the random variables $(X_i : i \in \mathcal{A})$, given $(X_j : j \in \mathcal{B}) = (x_j : j \in \mathcal{B})$, that we get from the joint pdf/pmf $f(x_1, \ldots, x_n)$.

If $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are three pairwise disjoint subsets of $\{1, 2, \ldots, n\}$, then we say that the two lists of variables $(X_i : i \in \mathcal{A})$ and $(X_i : i \in \mathcal{B})$ are conditionally independent, given $(X_j : j \in \mathcal{C})$, if and only if

$$\begin{aligned} f((x_i : i \in \mathcal{A} \cup \mathcal{B}) \,|\, (x_j : j \in \mathcal{C})) = f((x_i : \ i \in \mathcal{A}) \,|\, (x_j : j \in \mathcal{C})) \\ \times f((x_i : \ i \in \mathcal{B}) \,|\, (x_j : j \in \mathcal{C})) \end{aligned}$$

for each vector $(x_j : j \in \mathcal{C})$ in the joint range of the random vector $(X_j : j \in \mathcal{C})$. The definition can be appropriately extended to more than three pairwise disjoint subsets of $\{1, 2, \ldots, n\}$.

Remark 1.1. Colin L. Mallows, famous for his C_p criterion in regression, died in November 2023 at the age of 93. See Dalal and Landwehr [2].

2 What kinds of hypotheses of independence did Good consider?

In this section, we clarify the kinds of independence that Good [5] enumerated and provide a new proof to his formula (1) for the e.g.f. $A(y)$.

Indirectly, for $n \geq 2$, Good [5] partitioned the index set $\{1, 2, \ldots, n\}$ for the random variables $X_1, \ldots, X_n$ into three disjoint groups: $\mathcal{C}$, $\mathcal{I}$, and $\mathcal{L}$. Except for $\mathcal{I}$, each of these groups of indexes of variables may be empty. The set $\mathcal{C}$ contains all the indexes for the variables on which we condition, the set $\mathcal{I}$ contains the indexes for the variables whose joint pdf or pmf (conditional on all the variables in $\mathcal{C}$) we factor in all possible ways, and $\mathcal{L}$ contains the indexes for the variables that are not used in the definition of conditional independence in the hypothesis we examine.

If $s = |\mathcal{C}|$, then we might choose the index set $\mathcal{C}$ in $\binom{n}{s}$ ways and $s \in \{0, 1, ..., n-2\}$.

Note that we can only condition on up to $n-2$ variables, because we need at least two variables to define any kind of independence: conditional ($s \geq 1$) or unconditional ($s = 0$). Thus,

$$2 \leq n - s = |\mathcal{I}| + |\mathcal{L}| \leq n \quad \text{and} \quad a_0 = a_1 = 0.$$

If we let $t = |\mathcal{I}|$, then there are $\binom{n-s}{t}$ ways of choosing the index set $\mathcal{I}$ and $t \in \{2, \ldots, n-s\}$.

If $r \in \mathbb{Z}_{\geq 2}$ and $\{S_1, \ldots, S_r\}$ is a partition of the set of indexes $\mathcal{I}$ (with $|S_i| \geq 1$ for $i = 1, 2, \ldots, r$), the hypothesis of independence corresponding to this partition (with $\mathcal{C}$ and $\mathcal{L}$ defined as above) is that

$$f\left((x_i : i \in \mathcal{I}) \,|\, (x_j : j \in \mathcal{C})\right) = \prod_{k=1}^{r} f\left((x_i : i \in S_k) \,|\, (x_j : j \in \mathcal{C})\right)$$

for each vector $(x_j : j \in \mathcal{C})$ in the joint range of the random vector

$$(X_j : j \in \mathcal{C}).$$

Given the set $\mathcal{I}$ with $t = |\mathcal{I}|$, there are $B_t - 1$ ways to form a partition $\{S_1, ..., S_r\}$ of $\mathcal{I}$ with $r \geq 2$, $|S_i| \geq 1$, and $S_1 \cup \cdots \cup S_r = \mathcal{I}$. This is a

well-known property of the Bell numbers (and the -1 is due to the fact that we exclude the partition $\{S_1\} = \{\mathcal{I}\}$ that has $r = 1$).

Thus, in the sense of Good [5], there are

$$a_n = \sum_{s=0}^{n-2} \sum_{t=2}^{n-s} \binom{n}{s} \binom{n-s}{t} (B_t - 1)$$

kinds of independence among n variables.

Letting $k = n - s$, we get

$$a_n = \sum_{k=2}^{n} \sum_{t=2}^{k} \binom{n}{n-k} \binom{k}{t} (B_t - 1) = \sum_{k=2}^{n} \sum_{t=2}^{k} \binom{n}{k} \binom{k}{t} (B_t - 1). \tag{4}$$

Note that Eq. (4) hold even for $n \in \{0, 1\}$ because empty sums are by definition 0 and we have $a_0 = a_1 = 0$.

Next we claim that

$$a_n = \sum_{k=2}^{n} (B_{k+1} - 2^k) \binom{n}{k}. \tag{5}$$

This can be proved by using Eq. (4) and the identities

$$B_{k+1} = \sum_{m=0}^{k} \binom{k}{m} B_m \quad \text{and} \quad 2^k = \sum_{m=0}^{k} \binom{k}{m}, \tag{6}$$

which are valid for all nonnegative integers k. (Note that $B_0 = B_1 = 1$.) For the first equation in (6), see Comtet [1, Section 5.4, Eq. (4c)].

Since $B_{k+1} = 2^k$ for $k = 0, 1$, we may alternatively write Eq. (5) as

$$a_n = \sum_{k=0}^{n} (B_{k+1} - 2^k) \binom{n}{k}. \tag{7}$$

Finally, we prove Eq. (1) using Eq. (7). To achieve that, we use the e.g.f. of the Bell numbers; see Comtet [1, Section 5.4, Eq. (4b)]:

$$\sum_{n=0}^{\infty} \frac{B_n}{n!} y^n = \exp(\exp(y) - 1). \tag{8}$$

Differentiating both sides of Eq. (8) with respect to y and shifting the index of summation, we get

$$\sum_{m=0}^{\infty} \frac{B_{m+1}}{m!} y^m = \sum_{n=1}^{\infty} \frac{B_n}{(n-1)!} y^{n-1} = \exp(\exp(y) + y - 1).$$

Using the above equations, we get

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{a_n}{n!} y^n &= \sum_{n=0}^{\infty} \frac{1}{n!} \left(\sum_{k=0}^{n} (B_{k+1} - 2^k) \binom{n}{k} \right) y^n \\
&= \sum_{k=0}^{\infty} \frac{y^k}{k!} (B_{k+1} - 2^k) \sum_{n=k}^{\infty} \frac{y^{n-k}}{(n-k)!} \\
&= \left(\sum_{k=0}^{\infty} \frac{y^k}{k!} B_{k+1} - \sum_{k=0}^{\infty} \frac{(2y)^k}{k!} \right) \left(\sum_{\ell=0}^{\infty} \frac{y^\ell}{\ell!} \right) \\
&= \exp\left(\exp(y) + 2y - 1\right) - \exp(3y).
\end{aligned}$$

This proves Eq. (1).

Remark 2.1. Consider a partition $\{S_1, \ldots, S_r\}$ of the index set $\mathcal{I}$ with $r \geq 2$. For $k = 1, \ldots, r$, let

$$\mathbf{X}_k = (X_i : i \in S_k).$$

Good [5] considered modes of independence of the following kind:

> Random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_r$ are conditionally independent, given each value in the joint range of the random vector $(X_j : j \in \mathcal{C})$.

3 Mallows's discussion of modes of independence among three variables

Consider three random variables X_1, X_2, X_3, and denote

- by s_1 the statement that X_1 is independent of the random vector (X_2, X_3);
- by s_1' the statement that X_2 and X_3 are (unconditionally) independent; and
- by s_1'' the statement that X_2 and X_3 are conditionally independent given each value in the range of X_1.

Define similarly the statements s_2, s_2', s_2'' and s_3, s_3', s_3''.

We invite the reader to prove the following equivalences:

- $s_1' \,\&\, s_2'' \Leftrightarrow s_1'' \,\&\, s_2' \Leftrightarrow s_1'' \,\&\, s_2'' \Leftrightarrow s_3$;
- $s_2' \,\&\, s_3'' \Leftrightarrow s_2'' \,\&\, s_3' \Leftrightarrow s_2'' \,\&\, s_3'' \Leftrightarrow s_1$;
- $s_3' \,\&\, s_1'' \Leftrightarrow s_3'' \,\&\, s_1' \Leftrightarrow s_3'' \,\&\, s_1'' \Leftrightarrow s_2$;
- $s_1 \,\&\, s_2 \Leftrightarrow s_2 \,\&\, s_3 \Leftrightarrow s_3 \,\&\, s_1 \Leftrightarrow s_1 \,\&\, s_2 \,\&\, s_3$.

The last statement,

$$s_1 \,\&\, s_2 \,\&\, s_3,$$

is also equivalent to the statement

$$\text{“}X_1, X_2, X_3 \text{ are independent”}.$$

Mallows [8] wrote that there are 17 independence hypotheses for the case $n = 3$:

(i) 1 like $s_1 \,\&\, s_2 \,\&\, s_3$

(ii) 3 like s_1

(iii) 3 like s_1'

(iv) 3 like s_1''

(v) 3 like $s_1' \,\&\, s_1''$

(vi) 3 like $s_1' \,\&\, s_2'$

(vii) 1 like $s_1' \,\&\, s_2' \,\&\, s_3'$

The previous equivalences guarantee that we did not leave any cases behind. Cases (i)–(iv) were examined by Good [5], who counted $a_3 = 1+3+3+3 = 10$ possibilities. The rest were not considered by him.

Mallows [8] then asked: “How many possibilities are there when the number of variables is 4, 5, 6, ...?” He did not answer that question and neither do we! Finding the total number of hypotheses (in the sense of Mallows [8]) for a general n is extremely difficult.

Determining which combinations of independence statements among n variables are equivalent for a general n (like what we did above for the case $n = 3$) is very difficult. One might start with the classic paper by Dawid [3] and try to follow up the thousands of papers that reference it.

Even if one determines exactly which statements of independence are equivalent, theoretically, he or she has to provide examples to show that the different possibilities of non-equivalent modes of hypotheses among n variables are *indeed* not equivalent. This, in general, is very tedious.

Remark 3.1. Note that Good [5] did not consider case (vii) above. Even though, in this case, $\mathcal{C} = \emptyset$ (since we do not condition on any variables) and $\mathcal{L} = \emptyset$ (since we do not omit any variable from consideration), we do *not* really partition the index set $\mathcal{I} = \{1, 2, 3\}$ into $\{S_1, \ldots, S_r\}$ for some integer $r \geq 2$. Here we actually have three index sets of variables that we do not omit or condition on:

- $\mathcal{I}_1 = \{2, 3\}$ (from condition s_1') with partition $\{\{2\}, \{3\}\}$;
- $\mathcal{I}_2 = \{3, 1\}$ (from condition s_2') with partition $\{\{3\}, \{1\}\}$; and
- $\mathcal{I}_3 = \{1, 2\}$ (from condition s_3') with partition $\{\{1\}, \{2\}\}$.

Good [5] did not account for such a situation. (We may make similar comments for case (vi) about each of the statements $s_1' \,\&\, s_2'$, $s_2' \,\&\, s_3'$, and $s_3' \,\&\, s_1'$.)

4 Mallows's second question about modes of independence of variables

Mallows [8] asked a second question: "What if the cases that involve conditional independence are ignored (6 of these 17)?" For $n = 3$, he is asking us to ignore cases (iv) and (v) in Section 3 above; *i.e.*, he is asking us to ignore the six non-equivalent statements s_1'', s_2'', s_3'', $(s_1' \,\&\, s_1'')$, $(s_2' \,\&\, s_2'')$, $(s_3' \,\&\, s_3'')$.

Out of 11 non-equivalent modes of independence among $n = 3$ variables that ignore conditional independence, Good [5] only considered 7, those in cases (i), (ii), and (iii) in Section 3.

Since now $\mathcal{C} = \emptyset$, by modifying the proof in Section 2, we may easily prove that the total number of modes of independence that Good [5] considered that ignore conditional independence is

$$b_n = \sum_{t=2}^{n} \binom{n}{t} (B_t - 1) = \sum_{t=0}^{n} \binom{n}{t} (B_t - 1).$$

Using Eqs. (6), we get

$$b_n = B_{n+1} - 2^n. \tag{9}$$

(Again $b_0 = b_1 = 0$ since we need at least two variables to define independence.)

By modifying again the proof in Section 2, one can easily prove that the e.g.f. of the numbers $(b_n : n \in \mathbb{Z}_{\geq 0})$ is

$$B(y) := \sum_{n=0}^{\infty} \frac{b_n}{n!} y^n = \exp(y + \exp(y) - 1) - \exp(2y). \tag{10}$$

Some values of b_n appear in Table 4.1. See also sequence A058681 in the OEIS [9].

Table 4.1: The number of Good's hypotheses of independence that ignore conditional independence.

n	0	1	2	3	4	5	6	7	8
b_n	0	0	1	7	36	171	813	4012	20891

Even though Mallows's [8] second problem has definitely an easier solution than his first one, we were still not able to solve it. Even for $n = 4$, in addition to the $b_4 = 36$ hypotheses of independence about variables X_1, X_2, X_3, X_4 considered by Good [5], we need to consider complicated statements such as the one below:

> Each of the random vectors (X_1, X_2, X_3), (X_1, X_4), (X_2, X_4), and (X_3, X_4) consists of independent random variables.

Remark 4.1. Eq. (9) above was originally obtained by Good [5, Eq. (9), p. 80], who called the number b_n "the total number of *purely* marginal independence hypotheses, which is also the number of purely conditional independence hypotheses".

5 Conclusion

In this paper, we clarified the modes of independence among n variables that were examined by Good [5] and Mallows [8], and we provided a new and more concise proof of some of Good's [5] formulas. In addition, we explained the challenges in answering Mallows's [8] two questions about the enumeration of different kinds of independence among n variables.

Answering Mallows's [8] interesting questions seems very difficult (even for the case $n = 4$), but we hope our paper will inspire future researchers in examining these problems in a new perspective. Maybe with the use of symbolic computation software one may achieve that in the future (and we believe that only the second question will be answered).

References

[1] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions* (*rev. ed.*), Kluwer, 2010.

[2] S. Dalal and J. Landwehr, *Obituary: Colin Mallows 1930–2023*, (December 2023), available electronically at https://imstat.org/2023/12/15/obituary-colin-mallows-1930-2023/.

[3] A.P. Dawid, Conditional independence in statistical theory, *J. Roy. Statist. Soc. Ser. B*, **41** (1979), 1–31.

[4] S.E. Fienberg, *The analysis of cross-classified categorical data* (*2nd ed.*), The MIT Press, 1987.

[5] I.J. Good, The number of hypotheses of independence for a random vector or for a multidimensional contingency table, and the Bell numbers, *Iranian J. Sci. Tech.*, **4** (1975), 77–83.

[6] I.J. Good, On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, *Ann. Statist.*, **4** (1976), 1159–1189.

[7] I.J. Good, C370. A compromise between credibility and subjective probability, *J. Stat. Comput. Simul.*, **36** (1990), 186–193.

[8] C.L. Mallows, C48. How many independence hypotheses are there?, *J. Stat. Comput. Simul.*, **9** (1979), 235–236.

[9] N. J. A. Sloane, *The on-line encyclopedia of integer sequences*, available electronically at http://oeis.org.

Petros Hadjicostas
University of Nevada, Las Vegas, Nevada, USA
peterhadji1@gmail.com