Note

Taxonomic Classes of Sets

JAMES F. LYNCH*

Department of Mathematics, Clarkson University, Potsdam, New York 13676

Communicated by the Managing Editors

Received August 21, 1987

A taxonomic class is a family of nonempty sets such that either every two of its sets are disjoint or one is contained in the other. A complete taxonomic class is one in which every set is the union of the minimal sets contained in it. A connected taxonomic class has exactly one maximal set. The main results are exact and asymptotic formulas for the number of these classes and maximal taxonomic classes. © 1990 Academic Press, Inc.

1. Introduction

By a class of sets we mean a set whose members are sets. A number of papers have studied classes where the sets obey restrictions involving intersection and containment (surveyed in [4, 5, 7]). S. M. Ulam suggested the following restriction. A class $\mathscr{C}$ of nonempty sets is *taxonomic* if for every $A, B \in \mathscr{C}$,

$$A \subseteq B \qquad \text{or} \qquad B \subseteq A \qquad \text{or} \qquad A \cap B = \varnothing.$$

The $\subseteq$ relation on such a class has a hierarchical or tree-like structure. This structure occurs whenever a collection of objects is classified into disjoint categories, subcategories, and so on. An example from biology is the taxonomy of organisms. The largest categories are the kingdoms, followed by the phyla, and continuing down to species. In many cases, particularly with domesticated animals and plants, there are further refinements. Since there are organisms that have not been discovered or completely classified, this taxonomy is incomplete. Formally, a taxonomic class $\mathscr{C}$ is *complete* if every $a \in \bigcup \mathscr{C}$ is a member of some minimal $S \in \mathscr{C}$. An instance of this

* Research partially supported by NSF Grants DCR-8402206 and DCR-8605551. This paper extends and improves results given in [8].

0097-3165/90 $3.00

would be the taxonomy of organisms whose classification is established. We also consider an extreme form of completeness—$\mathscr{C}$ will be said to be *maximal* if for every nonempty $S \subseteq \bigcup \mathscr{C}$, either $S \in \mathscr{C}$ or $\mathscr{C} \cup \{S\}$ is not taxonomic.

Taxonomic classes with exactly one maximal set, i.e., $\bigcup \mathscr{C} \in \mathscr{C}$, will play an important role in our proofs. For obvious reasons, such classes are said to be *connected.* More generally, a *component* of $\mathscr{C}$ is a maximal connected subclass of $\mathscr{C}$, i.e., the subclass of all sets contained in a maximal set.

Let $n \geqslant 0$. $\mathscr{C}$ is a *class on n* if $\bigcup \mathscr{C} = \{0, 1, ..., n-1\}$. Our main results are exact formulas and asymptotic estimates for

T_n = number of taxonomic classes on n

U_n = number of connected taxonomic classes on n

C_n = number of complete taxonomic classes on n

D_n = number of connected complete taxonomic classes on n

M_n = number of maximal taxonomic classes on n.

Following standard conventions, $T_0 = C_0 = 1$ and $U_0 = D_0 = 0$, i.e., $\varnothing$ is the only (complete) taxonomic class on 0, and there is no connected taxonomic class on 0.

If $\varnothing$ were allowed as a member of taxonomic classes, then T_n, U_n, C_n, and D_n would become exactly twice as large, since the presence or absence of $\varnothing$ in a class does not affect any of the definitions.

The formulas for T_n, U_n, C_n, and D_n are derived by solving functional equations for their exponential generating series. The proofs of the exact formulas are similar to Polya's enumeration of rooted trees [6, 10], and the asymptotic estimates use methods described in [2]. This approach also works for M_n, but a more elementary proof will be given which yields additional information: for $n \geqslant 1$, every maximal taxonomic class on n has exactly $2n-1$ sets.

Lastly, using a theorem of Compton [3], we give easy derivations of asymptotic formulas for the probability that a random (complete) taxonomic class has a given number of components and the expected number of components.

2. Exact Formulas

Let $T(x)$, $U(x)$, $C(x)$, $D(x)$, and $M(x)$ be the respective exponential generating series for the sequences T_n, U_n, C_n, D_n, and M_n, e.g.,

$$T(x) = \sum_{n \geqslant 0} \frac{T_n}{n!} x^n.$$

We begin by studying the relationship between $T(x)$ and $U(x)$.

2.1. LEMMA. *We have*

$$T(x) = e^{U(x)} \tag{2.2}$$

$$2U(x) + 1 = e^x T(x). \tag{2.3}$$

Proof. (2.2) is an application of a well-known principle [1]. The central idea is that every taxonomic class is the disjoint union of connected taxonomic classes. The exponential generating series for the number of taxonomic classes that are the disjoint union of exactly m connected classes is $U(x)^m/m!$.

To prove (2.3), for every connected taxonomic class $\mathscr{C}$ let

$$\mathscr{C}' = \mathscr{C} - \left\{\bigcup \mathscr{C}\right\},$$

$$S = \bigcup \mathscr{C} - \bigcup \mathscr{C}',$$

and

$$f(\mathscr{C}) = \langle \mathscr{C}', S \rangle.$$

It is easily seen that f is one to one. Thus to count C_n it suffices to count ordered pairs $\langle \mathscr{C}', S \rangle$ that are in the range of f.

There are three cases:

(I) $\mathscr{C}' = \varnothing$

(II) $\mathscr{C}'$ is connected

(III) $\mathscr{C}' \neq \varnothing$ and $\mathscr{C}'$ is not connected.

The exponential generating series for $\mathscr{C}'$ is

$$1 \qquad \text{for case (I)}$$

$$U(x) \qquad \text{for case (II)}$$

and

$$T(x) - U(x) - 1 \qquad \text{for case (III)}.$$

In case (I) S is nonempty because $U_0 = 0$. In case (II) S is also nonempty; otherwise $\bigcup \mathscr{C} = \bigcup \mathscr{C}'$. But $\bigcup \mathscr{C} \notin \mathscr{C}'$ by definition of $\mathscr{C}'$ and $\bigcup \mathscr{C}' \in \mathscr{C}'$ since $\mathscr{C}'$ is connected. Conversely, for any nonempty S, $\langle \mathscr{C}', S \rangle$ is in the range of f. Therefore in these two cases the exponential generating series for S is

$$\sum_{n \geqslant 1} \frac{x^n}{n!} = e^x - 1.$$

In case (III) it is

$$\sum_{n\geqslant 0}\frac{x^n}{n!}=e^x.$$

Since the three cases are mutually exclusive,

$$U(x)=(e^x-1)+U(x)(e^x-1)+(T(x)-U(x)-1)e^x$$

and (2.3) follows. ▮

2.4. THEOREM. *We have*

$$U_n=\frac{1}{2}\sum_{k\geqslant 1}k^n\left(\frac{1}{k!}\right)\left(\frac{k}{2}\right)^{k-1}e^{-k/2}\qquad \textit{for}\quad n\geqslant 1$$

$$T_n=\sum_{k\geqslant 1}(k-1)^n\left(\frac{1}{k!}\right)\left(\frac{k}{2}\right)^{k-1}e^{-k/2}.$$

Proof. It will be convenient to use auxiliary variables

$$w=2U(x)+1\qquad\text{and}\qquad y=e^x$$

Then by Lemma 2.1

$$w=ye^{(w-1)/2}\tag{2.5}$$

Express w as an infinite series in y and x:

$$w=\sum_{k\geqslant 1}a_k y^k=\sum_{n\geqslant 0}b_n x^n.$$

By Lagrange's Inversion Formula [6]

$$\begin{aligned}a_k&=\left(\frac{1}{k!}\right)\left[\left(\frac{d}{dw}\right)^{k-1}(e^{(w-1)/2})^k\right]_{w=0}\\&=\left(\frac{1}{k!}\right)\left(\frac{k}{2}\right)^{k-1}e^{-k/2}.\end{aligned}\tag{2.6}$$

Therefore

$$b_n=\frac{1}{n!}\left[\sum_{k\geqslant 1}k^n\left(\frac{1}{k!}\right)\left(\frac{k}{2}\right)^{k-1}e^{-k/2}\right],$$

and the result for U_n follows from $U(x)=(w-1)/2$.

To solve for T_n, by Lemma 2.1

$$T(x) = \frac{2U(x)+1}{e^x} = \frac{w}{y}.$$

By (2.6),

$$\begin{aligned}\frac{w}{y} &= \sum_{k \geqslant 1} \left(\frac{1}{k!}\right)\left(\frac{k}{2}\right)^{k-1} e^{-k/2} y^{k-1} \\ &= \sum_{n \geqslant 0} \frac{x^n}{n!} \left[\sum_{k \geqslant 1} (k-1)^n \left(\frac{1}{k!}\right)\left(\frac{k}{2}\right)^{k-1} e^{-k/2}\right]\end{aligned}$$

and the result for T_n follows. ▌

2.7. Lemma. *For $n \geqslant 1$, $2D_n = C_n + 1$.*

Proof. Let **C** (**D**) be the collection of all complete (connected complete) taxonomic classes on n other than $\{\{0, 1, ..., n-1\}\}$. Define a function $f\colon \mathbf{C} \to \mathbf{D}$ by $f(\mathscr{C}) = \mathscr{C} \cup \{\{0, 1, ..., n-1\}\}$. It is evident that the range of f is **D**, and for every $\mathscr{C} \in \mathbf{D}$,

$$|f^{-1}(\mathscr{C})| = |\{\mathscr{C}, \mathscr{C} - \{\{0, 1, ..., n-1\}\}\}| = 2.$$

The lemma follows. ▌

2.8. Corollary. *We have*

$$C(x) = e^{D(x)} \tag{2.9}$$

$$2D(x) + 2 = e^x + C(x). \tag{2.10}$$

Proof. (2.9) has the same proof as (2.2), and (2.10) follows immediately from the Lemma, noting that $D_0 = 0$ and $C_0 = 1$. ▌

2.11. Theorem. *We have*

$$C_n = \sum_{k \geqslant 0} \frac{k^n}{2^k k!} \left[\sum_{m \geqslant 1} m^k \left(\frac{1}{m!}\right)\left(\frac{m}{2}\right)^{m-1} e^{-m}\right].$$

Proof. We will use the auxiliary variables $u = C(x)$ and $v = e^{e^x/2}$. By Corollary 2.8

$$u = ve^{u/2-1}.$$

Expressing u as an infinite series in v, we have

$$u = \sum_{m \geqslant 1} a_m v^m,$$

and using Lagrange's Inversion Formula,

$$\begin{aligned} a_m &= \left(\frac{1}{m!}\right)\left[\left(\frac{d}{du}\right)^{m-1} (e^{u/2-1})^m\right]_{u=0} \\ &= \left(\frac{1}{m!}\right)\left(\frac{m}{2}\right)^{m-1} e^{-m}, \end{aligned}$$

which implies the Theorem. ▌

2.12. Corollary. *For $n \geqslant 1$*

$$D_n = \frac{1}{2}\left[1 + \sum_{k \geqslant 0} \frac{k^n}{2^k k!}\left[\sum_{m \geqslant 1} m^k \left(\frac{1}{m!}\right)\left(\frac{m}{2}\right)^{m-1} e^{-m}\right]\right].$$

Proof. Immediate from Lemma 2.7. ▌

We could use the functional equation $M(x) = x + M(x)^2/2$ (where we define $M_0 = 0$) to obtain a solution for M_n. However, a more direct approach is easier and also yields more information:

2.13. Proposition. *Every maximal taxonomic class on $n \geqslant 1$ contains exactly $2n-1$ sets.*

Proof. We use induction on n. For $n=1$, the Proposition is obvious. Assume it has been proven for all $m \leqslant n$. Let $\mathscr{C}$ be a maximal taxonomic class on $n+1$. Since $n \geqslant 1$ and $\mathscr{C}$ is maximal, there must be some maximal $A \in \mathscr{C}$ such that $A \neq \{0, 1, ..., n\}$. Then $B = \{0, 1, ..., n\} - A \in \mathscr{C}$; otherwise $\mathscr{C} \cup \{B\}$ would be a larger taxonomic class than $\mathscr{C}$. Let

$$\begin{aligned} \mathscr{C}_0 &= \{S \in \mathscr{C} : S \subseteq A\} \\ \mathscr{C}_1 &= \{S \in \mathscr{C} : S \subseteq B\}. \end{aligned}$$

It can be seen that $\mathscr{C}_0$ and $\mathscr{C}_1$ are maximal taxonomic classes on A and B respectively. Therefore by induction

$$|\mathscr{C}_0| = 2\,|A| - 1 \qquad \text{and} \qquad |\mathscr{C}_1| = 2\,|B| - 1.$$

Since $\mathscr{C} = \mathscr{C}_0 \uplus \mathscr{C}_1 \uplus \{\{0, 1, ..., n\}\}$,

$$|\mathscr{C}| = |\mathscr{C}_0| + |\mathscr{C}_1| + 1,$$

and the result follows. ▌

2.14. Theorem. *For $n \geqslant 1$,*

$$M_n = \prod_{i=1}^{n-1} (2i-1).$$

Proof. We use induction on n. For $n=1$, we follow the usual convention that the null product is 1, and the result holds.

Assume the result holds for n.

Let $\mathbf{M}_n$ be the collection of all maximal taxonomic classes on n. Define a function f on $\mathbf{M}_{n+1}$ by

$$f(\mathscr{C}) = \{S - \{n\} : S \in \mathscr{C} \text{ and } S \neq \{n\}\} - \{\{n\}\}$$

for $\mathscr{C} \in \mathbf{M}_{n+1}$. It is easily seen that the range of f is $\mathbf{M}_n$. In fact, for every $\mathscr{D} \in \mathbf{M}_n$, $|f^{-1}(\mathscr{D})| = 2n-1$. To see this, take any $A \in \mathscr{D}$ and let

$$\mathscr{C}_A = \{S : A \nsubseteq S \in \mathscr{D}\} \cup \{S \cup \{n\} : A \subseteq S \in \mathscr{D}\} \cup \{\{n\}\}$$

Then $f(\mathscr{C}_A) = \mathscr{D}$ and for $B \in \mathscr{D}$, $B \neq A$ implies $\mathscr{C}_B \neq \mathscr{C}_A$. Then $|f^{-1}(\mathscr{D})| = |\mathscr{D}| = 2n-1$ by Proposition 2.13. Therefore

$$\begin{aligned} M_{n+1} &= M_n(2n-1) \\ &= \prod_{i=1}^{n} (2i-1) \end{aligned}$$

by the induction hypothesis. ▌

3. Asymptotic Estimates

To estimate T_n, U_n, C_n, and D_n, we apply the following result of Bender (Theorem 5 in [2]) to Lemma 2.1 and Corollary 2.8. Alternatively, we could use the exact formulas given by Theorems 2.4 and 2.11, and approximate the sums using Stirling's formula (see [2]). Either way, we get the same asymptotic estimates.

3.1. Theorem. *Assume that the power series $w(z) = \sum a_n z^n$ with nonnegative coefficients satisfies $F(z, w) \equiv 0$. Suppose there exist real numbers $r > 0$ and $s > a_0$ such that*

(i) *for some $\delta > 0$, $F(z, w)$ is analytic whenever $|z| < r + \delta$ and $|q| < s + \delta$*

(ii) $F(r, s) = F_w(r, s) = 0$

(iii) *if* $|z| \leqslant r$, $|w| \leqslant s$, *and* $F(z, w) = F_w(z, w) = 0$, *then* $z = r$ *and* $w = s$

(iv) $F_z(r, s) \neq 0$ *and* $F_{ww}(r, s) \neq 0$.

Then

$$a_n \sim ((rF_z)/(2\pi F_{ww}))^{1/2} n^{-3/2} r^{-n},$$

where the partial derivatives F_z *and* F_{ww} *are evaluated at* $z = r$ *and* $w = s$.

The proof uses the Weierstrass Preparation Theorem [9] to characterize the behavior of F near (r, s). The asymptotic formula then follows from Darboux's Theorem [11].

3.2. THEOREM. *We have*

(i) $U_n \sim n^{n-1} e^{-n} (\ln 2 - \frac{1}{2})^{-n+1/2}$

(ii) $T_n \sim n^{n-1} e^{-n+1/2} (\ln 2 - \frac{1}{2})^{-n+1/2}$

(iii) $D_n \sim (\ln 2)^{1/2} n^{n-1} e^{-n} (\ln \ln 4)^{-n+1/2}$

(iv) $C_n \sim 2(\ln 2)^{1/2} n^{n-1} e^{-n} (\ln \ln 4)^{-n+1/2}$.

Proof. (i) Let $F(z, w) = 2w - e^z e^w + 1$. By Lemma 2.1,

$$F(z, w) \equiv 0 \qquad \text{when} \quad w = U(z) = \sum_{n \geqslant 0} \frac{U_n}{n!} z^n.$$

Clearly F is analytic,

$$F_z(z, w) = -e^z e^w, \qquad F_w(z, w) = 2 - e^z e^w, \qquad \text{and} \qquad F_{ww}(z, w) = -e^z e^w.$$

Letting $r = \ln 2 - \frac{1}{2} > 0$ and $s = \frac{1}{2} > 0 = U_0/0!$, we have

$$F(z, w) = F_w(z, w) = 0 \qquad \text{iff} \quad z = r \quad \text{and} \quad w = s.$$

Lastly,

$$F_z(r, s) = -2 \qquad \text{and} \qquad F_{ww}(r, s) = -2.$$

Therefore the conditions of Bender's Theorem are met, and

$$\begin{aligned} U_n &\sim n!\left[(\ln 2 - \tfrac{1}{2}) F_z(\ln 2 - \tfrac{1}{2}, \tfrac{1}{2})/(2\pi F_{ww}(\ln 2 - \tfrac{1}{2}, \tfrac{1}{2}))\right]^{1/2} \\ &\quad \times n^{-3/2} (\ln 2 - \tfrac{1}{2})^{-n} \\ &= n!(2\pi)^{-1/2} (\ln 2 - \tfrac{1}{2})^{-n+1/2} n^{-3/2} \\ &\sim n^{n-1} e^{-n} (\ln 2 - \tfrac{1}{2})^{-n+1/2}. \end{aligned}$$

(ii) Let $F(z, w) = ew^2 - e^{we^z}$, $r = \ln 2 - \frac{1}{2}$, and $s = e^{1/2}$. The proof proceeds exactly as above.

(iii) Let $F(z, w) = e^z + e^w - 2w - 2$, $r = \ln \ln 4$, and $s = \ln 2$. The proof is the same as (i) and (ii) except that Corollary 2.8 is used instead of Lemma 2.1.

(iv) follows from (iii) and Lemma 2.7. ▌

3.3. THEOREM. *We have*

$$M_n \sim 2^{n-1/2} \left(\frac{n-1}{e}\right)^{n-1}.$$

Proof. By Theorem 2.14,

$$M_n \prod_{i=1}^{n-1} 2i = (2n-2)!.$$

Therefore

$$M_n = \frac{(2n-2)!}{2^{n-1}(n-1)!}$$

and the result follows by Stirling's formula. ▌

Our final theorem pertains to probabilities that a random (complete) taxonomic class has a given number of components. It uses a theorem of Compton [3] on component distribution probabilities for labeled relational structures. This theorem extends without difficulty to taxonomic classes. For a set of structures $\mathbf{R}$, let $\mathbf{C}$ be the random variable such that for $S \in \mathbf{R}$, $\mathbf{C}(S)$ is the number of components of S, and for $m \geqslant 1$ let

$$\mu_n(\mathbf{C} = m) = \frac{|\{S \in \mathbf{R} : S \text{ has } n \text{ elements and } m \text{ components}\}|}{|\{S \in \mathbf{R} : S \text{ has } n \text{ elements}\}|}$$

$$E(\mathbf{C}, \mu_n) = \sum_{m \geqslant 1} m\mu_n(\mathbf{C} = m).$$

That is, $E(\mathbf{C}, \mu_n)$ is the expected number of components under the uniform probability measure μ_n.

3.4. THEOREM (Compton). *Suppose* $\mathbf{R}$ *is closed under disjoint unions, components, and isomorphism, and it has exponential generating series* $a(x) = \sum_{n \geqslant 0} (a_n/n!)x^n = e^{c(x)}$ *and* $c(x) = \sum_{n \geqslant 0} (c_n/n!)x^n$. *If*

(i) $\lim_{n \to \infty} na_{n-1}/a_n = R$, $0 < R \leqslant \infty$, *and*

(ii) *for some nonnegative integer* r *and real* $K > 0$ *we have for large enough* n *and* $r \leqslant k \leqslant n$ *that* $(a_k/(k-r)!)R^k \leqslant K(a_n/(n-r)!)R^n$,

then $\lim_{n \to \infty} \mu_n(\mathbf{C} = m) = c(R)^{m-1}/((m-1)!\, a(R))$ *and* $\lim_{n \to \infty} E(\mathbf{C}, \mu_n) = 1 + c(R)$.

Taking $\mathbf{R}$ to be the collection of taxonomic classes on n, for arbitrary n, let μ_n be as above. Similarly, let $\lambda_n(\mathbf{C}=m)$ be the probability that a randomly selected complete taxonomic class on n has m components.

3.5. THEOREM. *We have*

(i) $\lim_{n\to\infty} \mu_n(\mathbf{C}=m)=(\frac{1}{2})^{m-1}/((m-1)!e^{1/2})$ *and* $\lim_{n\to\infty} E(\mathbf{C},\mu_n) = \frac{3}{2}$

(ii) $\lim_{n\to\infty} \lambda_n(\mathbf{C}=m)=(\ln 2)^{m-1}/(2(m-1)!)$ *and* $\lim_{n\to\infty} E(\mathbf{C},\lambda_n) = \ln 2+1$.

Proof. (i) From Theorem 3.2(ii), it is easily seen that the conditions of Compton's theorem hold with $R=\ln 2-\frac{1}{2}$ and $r=2$. From Lemma 2.1, $U(R)+\frac{1}{2}=e^{-1/2+U(R)}$, so $U(R)=\frac{1}{2}$. The conclusion then follows.

(ii) is similar. ▌

REFERENCES

1. E. A. BENDER AND J. R. GOLDMAN, Enumerative uses of generating functions, *Indiana Univ. Math. J.* **20** (1971), 753–765.
2. E. A. BENDER, Asymptotic methods in enumeration, *SIAM Rev.* **16** (1974), 485–515.
3. K. J. COMPTON, Some methods for computing component distribution probabilities in relational structures, *Discrete Math.* **66** (1987), 59–77.
4. P. ERDÖS, Problems and results in combinatorial analysis, *in* "Combinatorics" (T. S. Motzkin, Ed.), pp. 77–89, Proceedings of Symposia in Pure Mathematics, Vol. XIX, Amer. Math. Soc., Providence, 1971.
5. P. ERDŐS AND D. J. KLEITMAN, Extremal problems among subsets of a set, *Discrete Math.* **8** (1974), 281-294.
6. F. HARARY AND E. PALMER, "Graphical Enumeration," Academic Press, New York, 1978.
7. G. KATONA, Extremal problems for hypergraphs, *in* "Combinatorics, Proc. NATO Advanced Study Inst., Breukelen, The Netherlands, 1974, Part 2: Graph Theory; Foundations, partitions and combinatorial geometry" (M. Hall, Jr., and J. H. Van Lint, Eds.), pp. 13–42, Math. Centre Tracts, No. 56, Math. Centrum, Amsterdam, 1974.
8. J. F. LYNCH, Enumeration of taxonomic classes of sets, *Congressus Numerantium* **47** (1985), 315–316.
9. A. I. MARKUSHEVICH, "Theory of Functions of a Complex Variable," Vol. II (R. A. Silverman, Ed.)., Prentice–Hall, Englewood Cliffs, NJ, 1965.
10. G. PÓLYA, Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen, *Acta Math.* **68** (1937), 145–254.
11. G. SZEGÖ, "Orthogonal Polynomials," Amer. Math. Soc. Coll. Publ., Vol. XXIII, rev. ed., Amer. Math. Soc., New York, 1959.