A PROBLEM IN STATISTICAL CLASSIFICATION THEORY

Philippe Flajolet
(Version of January 14, 1997)

This problem discussed here is at the origin of the whole Combstruct package. On October 8, 1992, Bernard Van Cutsem, a statistician at the University of Grenoble wrote to us:

In classification theory, we make use of hierarchical classification trees. I would need to generate at random such classification trees according to the uniform law. The elements to be classified may be taken as distinguished integers say from 1 to n. Do you know of an algorithm for doing this?

This led to a cooperation involving Paul Zimmermann, Bernard Van Cutsem, and Philippe Flajolet, out of which the general theory and the algorithms of Combstruct evolved, see *Theoretical Computer Science*, vol. 132, pp. 1-35. A first implementation was designed by Paul Zimmermann in 1993, under the name Gaia (*Maple Technical Newsletter*, 1994 (1), pp. 38-46).

Van Cutsem's original question was motivated by the following problem: Classification programmes in statistics build classification trees, usually proceeding by successive aggregations of closest neighbours amongst existing classes. How can we measure the way a classification carries useful information and not just "random noise"? Certainly, "good" classification trees should exhibit characteristics that depart significantly from random ones. Hence the need to simulate and analyse parameters of random classification trees.

Statistics and classification theory

Specification

We start by loading the combstruct package.

```
> with(combstruct);
```

[*allstructs*, *count*, *draw*, *finished*, *gfeqns*, *gfseries*, *gfsolve*, *iterstructs*, *nextstruct*, *prog_gfeqns*, *prog_gfseries*, *prog_gfsolve*]

A classification is either: 1) an atom; 2) a set of classification trees of degree at least 2. Atoms are distiguishable, hence we are in a labelled universe. Note that the Set construction translates a pure graph-theoretic structure with no ordering between descendents of a node.

```
> hier:=[H,{H=Union(Z,Set(H,card>1))},labelled]:
```

The original problem of Van Cutsem is solved by single commands like

```
> draw(hier,size=10);
```

$$\mathrm{Set}(\mathrm{Set}(\mathrm{Set}(\mathrm{Set}(Z_6, Z_2, \mathrm{Set}(Z_4, Z_8)), Z_7), Z_1), \mathrm{Set}(Z_5, Z_3, Z_{10}), Z_9)$$

We may adopt a more concise representation format:

```
> lreduce:=proc(e) eval(subs({Set=proc() {args} end,
  Sequence=proc() [args] end},e)) end:
> lreduce(draw(hier,size=20));
```

$\{\{\{\{\{\{$

$\{\{\{\{\{\{\{\{\{Z_9, Z_{13}\}, \{Z_7, Z_{17}\}\}, Z_{14}\}, \{Z_2, Z_{16}\}\}, Z_{20}\}, Z_6\}, Z_{15}, Z_{19}\}, Z_3\}, Z_{12}\},$

$Z_{10}\}, Z_5\}, Z_8\}, Z_{18}\}, Z_{11}\}, Z_1\}, Z_4\}$

Random generation takes only a few seconds while counting tables (that serve to determine splitting probabilities) are set up on the fly.

```
> for j from 20 by 20 to 100 do j,lreduce(draw(hier,size=j))
  od;
```

$20, \{\{Z_6, Z_3, Z_{13}\}, \{\{\{Z_7, Z_{11}\}, Z_{16}\},$

$\{\{\{\{Z_2, Z_{18}\}, Z_{17}\}, \{\{\{Z_4, Z_{20}\}, Z_{10}\}, \{\{\{Z_5, Z_{19}, Z_{12}\}, Z_9\}, Z_8, Z_1\}\}\}, Z_{15}\}\}, Z_{14}\}$

$40, \{\{Z_{36}, \{Z_{15}, Z_{16}, \{\{Z_{10}, Z_{30}\}, \{Z_2, \{Z_{20}, Z_{31}, \{Z_{11}, \{Z_{25}, \{\{Z_4, Z_{12}\},$

$\{Z_5, \{\{Z_{23}, Z_{33}\}, \{Z_1, Z_{34}\}\}, \{Z_{29}, \{Z_{17}, \{Z_{28}, \{Z_3, Z_{27}, Z_{21}, \{Z_8, Z_9, Z_{26}\}\}\}\}\}\}\}\}\}$

$\}\}\}\}, \{Z_{24}, \{Z_{22}, \{Z_{35}, \{\{Z_{40}, Z_{37}\}, \{\{Z_{19}, Z_{32}\}, \{Z_6, \{Z_7, Z_{39}, Z_{38}\}\}\}\}\}\}\}\},$

$\{Z_{14}, Z_{13}, Z_{18}\}\}$

$60, \{\{Z_9, \{\{Z_{32}, \{\{Z_{41}, Z_{39}\}, \{Z_{22}, Z_{45}\}, \{Z_{25}, \{Z_{47}, Z_{56}\}\}\}\},$

$\{Z_{19}, \{\{Z_{28}, \{Z_{18}, \{Z_{11}, Z_{33}, Z_{50}, \{Z_{12}, Z_{30}\}\}\}\}, Z_{40}\}, \{Z_{60}, Z_{36}, \{Z_{48}, \{Z_{43}, Z_{38}\}\}\}$

$\}\}, \{\{Z_8, Z_{20}\}, Z_{49}\}, \{Z_{27}, \{Z_{21}, \{$

$\{Z_{13}, \{Z_{10}, Z_{29}, Z_{35}, \{Z_{31}, \{Z_{58}, \{\{Z_{34}, \{Z_7, Z_{53}\}, \{Z_{37}, Z_{52}\}\}, Z_{46}\}, \{Z_{16}, Z_{57}\}\}\}\}\},$

$\{Z_5, \{Z_3, Z_{23}\}\}\}, \{Z_4, \{\{Z_2, Z_{14}\}, \{Z_{15}, Z_{17}, \{\{Z_{42}, \{\{Z_6, Z_1, Z_{24}\}, Z_{54}\}\}, Z_{51}\}\}\}\},$

$Z_{59}, Z_{44}\}, \{Z_{26}, Z_{55}\}\}\}$

$80, \{\{\{\{\{Z_{74}, \{\{\{Z_{70}, \{Z_{69}, \{Z_{80}, \{Z_{19}, Z_{41}\}, \{Z_{12}, Z_{29}, Z_{39}\}\}, \{Z_{75}, \{Z_{68}, Z_{28}\}\}\},$

$\{Z_{72}, \{Z_{73}, Z_{59}, Z_{51}\}\}\}, Z_{44}\}, \{Z_{62}, \{\{\{$

$\{Z_8, \{\{\{\{Z_{18}, \{Z_{64}, \{Z_{13}, \{Z_{32}, Z_{42}\}, Z_{24}, Z_{56}\}, Z_{31}\}\}, Z_{30}\}, Z_{45}\}, Z_{50}\}\}, \{Z_5, Z_{35}\},$

$\{Z_{76}, \{\{Z_{79}, Z_{47}\}, \{Z_{77}, Z_{66}\}, \{Z_{36}, Z_{58}\}\}\}\},$

$\{Z_{34}, \{Z_{14}, Z_{63}, \{\{\{Z_4, Z_{78}\}, Z_{21}\}, \{Z_{23}, Z_{26}, Z_{27}\}\}\}\}\}, \{Z_{15}, Z_{53}\}, \{Z_{17}, Z_{16}\}\}\},$

$\{Z_7, Z_{65}\}\}, \{Z_9, \{Z_2, \{\{\{Z_{71}, \{\{Z_{20}, \{Z_3, Z_{10}, Z_{11}, \{\{Z_{43}, Z_{22}\}, Z_{57}\}\}\}, \{Z_{61}, Z_{38}\}\},$

$\{\{Z_{67}, \{\{Z_1, Z_{48}\}, \{\{Z_{60}, Z_{46}\}, \{Z_6, Z_{49}\}\}\}\}, Z_{52}\}\}, Z_{55}\}, Z_{40}\}\}\}\}, Z_{37}\}, Z_{33}\}, Z_{54}$

$\}, Z_{25}\}$

$100, \{\{\{\{Z_3, \{\{\{\{\{\{\{\{\{\{Z_{47}, Z_{52}\}, \{\{Z_{91}, \{\{Z_{84}, Z_{35}\}, Z_{57}\}\}, Z_{75}\}\}, \{\{\{\{\{\{\{\{$

$\{\{$

$\{\{Z_5, \{\{\{\{\{Z_{65}, Z_{55}\}, \{Z_{15}, Z_{16}\}\}, Z_{96}\}, Z_{39}\}, \{Z_{93}, Z_{56}\}, Z_{38}\}\}, \{Z_{99}, Z_{97}, Z_{46}\}\}, \{$

$\{\{Z_2, \{\{\{\{\{Z_{72}, Z_{33}, Z_{29}\}, Z_{27}\}, \{\{Z_1, Z_{18}\}, Z_{36}, Z_{58}\}\}, \{Z_{13}, Z_{53}\}, Z_{98}\}, Z_{37}\}\}, Z_{60}$

$\}, \{\{Z_{54}, Z_{100}\}, Z_{64}, Z_{88}\}\}\}, \{Z_{68}, Z_{24}\}\}, Z_{23}\},$

$\{\{\{\{Z_8, \{\{Z_7, Z_{28}\}, \{Z_{78}, Z_{43}\}\}, Z_{62}\}, \{Z_{81}, Z_{77}\}\}, Z_{26}, Z_{42}\}, Z_{48}\}\},$

$\{\{Z_{87}, Z_{63}\}, Z_{79}\}, Z_{73}\}, Z_{83}\}, \{Z_{67}, Z_{85}\}, Z_{14}\}, Z_{59}\},$

$\{\{\{Z_{80}, Z_{50}\}, Z_{21}\}, \{\{\{Z_9, \{Z_6, Z_{30}, Z_{45}\}\}, Z_{25}\}, Z_{41}\}\}, Z_{34}\}, Z_{69}\}, Z_{51}\}, Z_{61}\}, Z_{32}\},$
$\{\{\{\{\{Z_{86}, Z_{49}\}, \{\{Z_{71}, Z_{31}\}, Z_{10}\}, \{Z_4, Z_{20}, Z_{95}\}, Z_{76}\}, Z_{17}\}, \{Z_{92}, Z_{94}\}\}, Z_{90}\}\}, Z_{89}$
$\}, Z_{19}\}, Z_{66}, Z_{74}\}, Z_{70}\}\}, Z_{12}, Z_{22}, Z_{44}\}, Z_{82}\}, Z_{11}, Z_{40}\}$

The number of objects of size *n* grows fast

```
> seq(count(hier,size=j),j=0..40);
```

0, 1, 1, 4, 26, 236, 2752, 39208, 660032, 12818912, 282137824, 6939897856,
188666182784, 5617349020544, 181790703209728, 6353726042486272,
238513970965257728, 9571020586419012608, 408837905660444010496,
18522305410364986906624, 887094711304119347388416,
44782218857752794987708416, 2376613641928863263785541632,
132280106444795539197625827328, 7705008716729749963527732396032,
468744135800126572558268335357952, 29730054390033099477714382005796864,
1962586033137616773187258991535456256,
134637659404625757681335270499748020224,
9584963644881810156457282812023186653184,
707173340451261419106233361561741760135168,
54005481349178592760992820984887698159828992,
4264097052284773334721826922349063450644185088,
347717494441208655889609784742705293689836535808,
29254882744213252920618676866373646493034580279296,
2537062817232412229880934405017394261055100581576704,
226588568079973535422904792685429246690212740119658496,
20823498054974293114261371204370275986079266118677037056,
1967585335605067331816881515789858061752205810690447900672,
191008011389013043866128326364206801607790424006556876013568,
19037012084887601494957603241545176663513597195454823228506112

This appears to be sequence **M3613** of the *Encyclopedia of Integer Sequences* and it corresponds to "Schroeder’s fourth problem". When the count is not too large, we can do exhaustive listings. This is made possible by Combstruct that is able to build canonical forms and generate elements under unique standard forms.

```
> for j to 4 do map(lreduce,allstructs(hier,size=j)) od;
```

$$[Z_1]$$

$$[\{Z_2, Z_1\}]$$

$$[\{\{Z_1, Z_3\}, Z_2\}, \{Z_3, \{Z_2, Z_1\}\}, \{Z_2, Z_1, Z_3\}, \{\{Z_2, Z_3\}, Z_1\}]$$

$[\{Z_1, \{Z_2, Z_4, Z_3\}\}, \{Z_1, \{Z_2, \{Z_4, Z_3\}\}\}, \{Z_3, \{Z_2, Z_4, Z_1\}\}, \{Z_1, \{Z_3, \{Z_2, Z_4\}\}\},$
$\{Z_1, \{\{Z_2, Z_3\}, Z_4\}\}, \{\{Z_4, Z_3\}, \{Z_2, Z_1\}\}, \{Z_2, \{\{Z_1, Z_3\}, Z_4\}\},$
$\{Z_2, \{Z_4, Z_1, Z_3\}\}, \{Z_3, \{Z_4, \{Z_2, Z_1\}\}\}, \{Z_4, Z_3, \{Z_2, Z_1\}\}, \{Z_2, Z_1, \{Z_4, Z_3\}\},$

$\{\{\{Z_1, Z_3\}, Z_2, Z_4\}, \{\{\{Z_2, Z_3\}, Z_1\}, Z_4\}, \{Z_2, \{Z_1, \{Z_4, Z_3\}\}\}, \{Z_2, Z_4, Z_1, Z_3\},$
$\{\{Z_2, Z_1, Z_3\}, Z_4\}, \{\{Z_3, \{Z_2, Z_1\}\}, Z_4\}, \{\{Z_2, Z_3\}, \{Z_4, Z_1\}\},$
$\{\{Z_1, Z_3\}, \{Z_2, Z_4\}\}, \{Z_2, \{Z_3, \{Z_4, Z_1\}\}\}, \{\{Z_2, Z_3\}, Z_4, Z_1\},$
$\{Z_3, \{Z_2, \{Z_4, Z_1\}\}\}, \{Z_3, \{Z_1, \{Z_2, Z_4\}\}\}, \{Z_1, Z_3, \{Z_2, Z_4\}\},$
$\{\{\{Z_1, Z_3\}, Z_2\}, Z_4\}, \{Z_2, Z_3, \{Z_4, Z_1\}\}\}$

Asymptotic analysis

We get generating function equations by combstruct[gfeqns]

```
> gfeqns(op(2..3,hier),z);
```

$$[\mathrm{Z}(z) = z, \mathrm{H}(z) = \mathrm{Z}(z) + \mathbf{e}^{\mathrm{H}(z)} - 1 - \mathrm{H}(z)]$$

And combstruct[gfsolve] attempts different strategies to solve the system

```
> gfsolve(op(2..3,hier),z);
```

$$\{\mathrm{Z}(z) = z, \mathrm{H}(z) = -\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,z-1/2)}\right) + \frac{1}{2}z - \frac{1}{2}\}$$

The solution involves Lambert's W function that is known to Maple: by definition, this is the solution of

$$\mathrm{W}(z)\,\mathbf{e}^{\mathrm{W}(z)} = z.$$

```
> H_z:=subs(",H(z));
```

$$H\_z := -\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,z-1/2)}\right) + \frac{1}{2}z - \frac{1}{2}$$

Objects being labelled, this is an exponential generating function (EGF).

```
> H_ztayl:=series(H_z,z=0,20);
```

$$H\_ztayl := z + \frac{1}{2}z^2 + \frac{2}{3}z^3 + \frac{13}{12}z^4 + \frac{59}{30}z^5 + \frac{172}{45}z^6 + \frac{4901}{630}z^7 + \frac{10313}{630}z^8 + \frac{400591}{11340}z^9 + \frac{8816807}{113400}z^{10} + \frac{27108976}{155925}z^{11} + \frac{1473954553}{3742200}z^{12} + \frac{43885539223}{48648600}z^{13} + \frac{710119934413}{340540200}z^{14} + \frac{12409621176731}{2554051500}z^{15} + \frac{35834430733963}{3143448000}z^{16} + \frac{9346699791424817}{347351004000}z^{17} + \frac{199627883623263677}{3126159036000}z^{18} + \frac{695699572204213751}{4569001668000}z^{19} + \mathrm{O}(z^{20})$$

As usual, we also obtain the corresponding ordinary generating functions by a Laplace transform applied to the series expansion

```
> series(subs(w=1/w,w*inttrans[laplace](H_ztayl,z,w)),w,20);
```

$$w + w^2 + 4\,w^3 + 26\,w^4 + 236\,w^5 + 2752\,w^6 + 39208\,w^7 + 660032\,w^8 + 12818912\,w^9 + 282137824\,w^{10} + 6939897856\,w^{11} + 188666182784\,w^{12} + 5617349020544\,w^{13} + 181790703209728\,w^{14} + 6353726042486272\,w^{15} + 238513970965257728\,w^{16} + 9571020586419012608\,w^{17} + 408837905660444010496\,w^{18} + \mathrm{O}(w^{19})$$

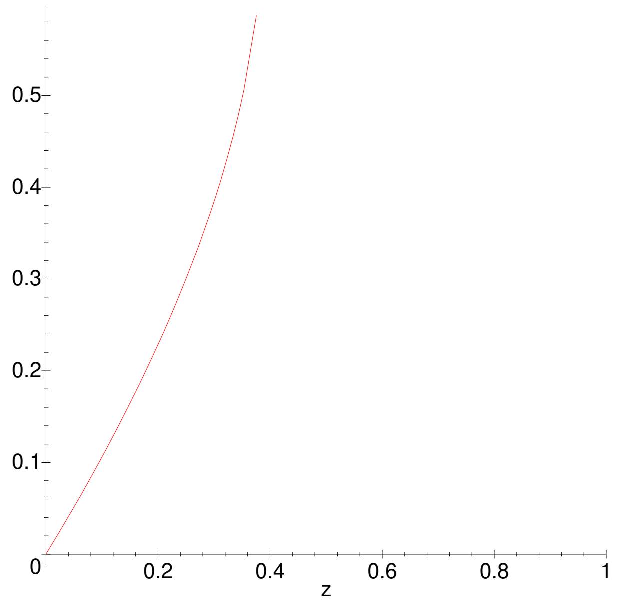
The result is then directly comparable to the counting coefficients:

```
> seq(count(hier,size=j),j=1..18);
```

1, 1, 4, 26, 236, 2752, 39208, 660032, 12818912, 282137824, 6939897856, 188666182784, 5617349020544, 181790703209728, 6353726042486272, 238513970965257728, 9571020586419012608, 408837905660444010496

In order to analyse the number of hierarchies, we must find the dominant singularity of their generating function. A plot detects a vertical slope near 0.4

```
> plot(H_z,z=0..1);
```



Here is a cute way to get the singularity "automatically". We express that the function ceases to be differentiable at its singularity.

```
> diff(H_z,z);
```

$$-\frac{1}{2}\frac{\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,z-1/2)}\right)}{1+\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,z-1/2)}\right)}+\frac{1}{2}$$

```
> rho:=solve(denom(")=0); evalf(rho,30);
```

$$\rho := -1+2\ln(2)$$

$$.386294361119890618834464242920$$

Next, we know that the singular expansion determines the asymptotic form of coefficients. Thus, we look at

```
> H_s:=subs(z=rho*(1-Delta^2),H_z);
```

$$H\_s := -\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,(-1+2\ln(2))\,(1-\Delta^2)-1/2)}\right)+\frac{1}{2}(-1+2\ln(2))\,(1-\Delta^2)-\frac{1}{2}$$

```
> H_sing:=map(simplify,series(H_s,Delta=0,5));Delta=sqrt(``(1-z/rho));
```

$H\_sing :=$

$$\ln(2) - \sqrt{-1+2\ln(2)}\,\Delta + \left(\frac{1}{6} - \frac{1}{3}\ln(2)\right)\Delta^2 - \frac{1}{36}(-1+2\ln(2))^{3/2}\Delta^3 + O(\Delta^4)$$

$$\Delta = \sqrt{\left(1 - \frac{z}{-1+2\ln(2)}\right)}$$

With this, we can get an asymptotic expansion for coefficients to any order, which is an interesting fact per se. Here is the first one:

```
> H_n_asympt:=n!*asympt(coeff(H_sing,Delta,1)*rho^(-n)*subs(
{cos(Pi*n)=1,O=0},simplify(asympt(binomial(1/2,n),n,2))),n
);
evalf(",20);
```

$$H\_n\_asympt := \frac{1}{2}\,\frac{n!\sqrt{-1+2\ln(2)}\left(\frac{1}{n}\right)^{3/2}}{\sqrt{\pi}\,(-1+2\ln(2))^n}$$

$$.17532920044983416513\,\frac{n\,\Gamma(n)\left(\frac{1}{n}\right)^{3/2}}{.38629436111989061 88^n}$$

And for n=50, we get

```
> round(evalf(subs(n=50,H_n_asympt),20));
count(hier,size=50); evalf(""/");
```

6800732403666468178700000000000000000000000000000000000000000\
00000000000000

6850001173973904748957636607981747909339510635183093671870187559786 8\
96636537470976

.9928074800

The error is of about 1% for $n = 50$. A complete asymptotic expansion can be obtained by this method, by taking successive singular terms into account.

The number of classification stages

We examine the number of internal nodes in a classification. This corresponds to the number of classes actually created. The idea is to make use of "marks" in the form of Epsilon structures that have size 0 (and thus do not affect the combinatorial model).

```
> hier2:=[H,{H=Union(Z,Prod(class,Set(H,card>1))),class=Epsilon
},labelled]:
```

Such marks do not affect the combinatorial model:

```
> seq(count(hier,size=j),j=0..11);
```

0, 1, 1, 4, 26, 236, 2752, 39208, 660032, 12818912, 282137824, 6939897856

```
> seq(count(hier2,size=j),j=0..11);
```

0, 1, 1, 4, 26, 236, 2752, 39208, 660032, 12818912, 282137824, 6939897856

(We change the formatting procedure to take Prod into account.)

```
> lreduce:=proc(e) eval(subs({Set=proc() {args} end,
Sequence=proc() [args] end, Prod=``},e)) end:
```

Here is a random object with "class" marking classification nodes:

```
> lreduce(draw(hier2,size=20));
```

$$(class, \{(class, \{(class, \{(class, \{(class, \{Z_{10}, Z_{17}, (class, \{Z_{19}, (class, \{Z_{11}, Z_{12}\})\})\}), (class, \{Z_{15}, (class, \{(class, \{Z_4, Z_7, (class, \{Z_3, Z_{13}\})\}), (class, \{Z_6, Z_{14}\}), (class, \{Z_{20}, Z_9\})\})\}), (class, \{Z_{18}, Z_{16}\})\}), Z_2\}), Z_5\}), Z_1, Z_8\})$$

The system determined by equations over bivariate generating functions can be solved by Maple:

```
> gfeqns(op(2..3,hier2),z,[[u,class]]);
```

$$[\mathrm{Z}(z,u) = z, \mathrm{class}(z,u) = u, \mathrm{H}(z,u) = \mathrm{Z}(z,u) + \mathrm{class}(z,u)\,(\mathbf{e}^{\mathrm{H}(z,u)} - 1 - \mathrm{H}(z,u))]$$

```
> H_zu:=solve(H=z+u*(exp(H)-1-H),H);
```

$$H\_zu := \frac{-\mathrm{LambertW}\left(-\frac{u\,\mathbf{e}^{\left(\frac{z-u}{1+u}\right)}}{1+u}\right) - \mathrm{LambertW}\left(-\frac{u\,\mathbf{e}^{\left(\frac{z-u}{1+u}\right)}}{1+u}\right) u + z - u}{1+u}$$

One gets averages by differentiation:

```
> H1_z:=subs(u=1,diff(H_zu,u));
```

$$H1\_z := 2\,\frac{\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,z-1/2)}\right)\left(-\frac{1}{4}\mathbf{e}^{(1/2\,z-1/2)} - \frac{1}{2}\left(-\frac{1}{4}-\frac{1}{4}z\right)\mathbf{e}^{(1/2\,z-1/2)}\right)}{\left(1+\mathrm{LambertW}\left(-\frac{1}{2}\mathbf{e}^{(1/2\,z-1/2)}\right)\right)\mathbf{e}^{(1/2\,z-1/2)}} - \frac{1}{4} - \frac{1}{4}z$$

Numerically, the mean number of nodes in a random classification tree of size n, when divided by n, is for $n = 1 .. 20$,

```
> Digits:=5: evalf(series(H1_z,z=0,22)):
  seq(coeff(",z,j)/count(hier,size=j)/j*j!,j=1..20);
```

0, .50000, .58336, .63463, .66610, .68727, .70248, .71387, .72271, .72990, .73567, .74063, .74469, .74824, .75134, .75398, .75634, .75849, .76039, .76206

This suggests that a random classification may have about .76 n classification stages.
We can in fact analyse this rigorously, using the asymptotic method already employed for counts.

```
> H1_s:=subs(z=rho*(1-Delta^2),H1_z):
  H1_sing:=map(simplify,series(H1_s,Delta=0,5));
  H1_n_asympt:=n!*asympt(coeff(H1_sing,Delta,-1)*rho^(-n)*subs(
  {cos(Pi*n)=1,O=0},simplify(asympt(binomial(-1/2,n),n,2))),n);
```

$$H1\_sing := -\frac{1}{2}\frac{\ln(2)-1}{\sqrt{-1+2\ln(2)}}\Delta^{-1} + \left(-\frac{1}{6}\ln(2) - \frac{1}{3}\right) - \frac{1}{24}\frac{-15\ln(2) + 2\ln(2)^2 + 7}{\sqrt{-1+2\ln(2)}}\Delta + \mathrm{O}(\Delta^2)$$

$$H1\_n\_asympt := -\frac{1}{2}\frac{n!\,(\ln(2)-1)\sqrt{\frac{1}{n}}}{\sqrt{-1+2\ln(2)}\sqrt{\pi}\,(-1+2\ln(2))^n}$$

```
> C_classif:=asympt(H1_n_asympt/H_n_asympt,n,1); evalf(",20);
```

$$C\_classif := -\frac{(\ln(2)-1)\,n}{-1+2\ln(2)}$$

$$.7943497247810449l547\,n$$

Thus, we have obtained (easily!) a new **Theorem**. *In a random classification tree, the number of classification stages (internal nodes) is asymptotic to*

$$-\frac{(\log(2)-1)\,n}{2\log(2)-1} = .794349724\,n.$$

Degrees in random classification trees

The corresponding generating functions are now outside of the range of implicit functions that Maple knows about. Thus, a separate mathematical analysis is needed. However, an empirical analysis based on small sizes is already quite informative. The following code builds a specification where nodes of degree k are marked. The principle is the obvious set-theoretic equation

$$\mathrm{Set}(X) = \mathrm{Union}(\mathrm{Set}(X, card < k), \mathrm{Set}(X, card = k), \mathrm{Set}(X, k \le card)).$$

The code uses combstruct[gfeqns] to generate the system of equations for each degree that is then expanded. In passing, it prints the corresponding generating function:

```
> deg_hier:=proc(k) local j,spec,n,dHH;
      spec:=[H,{
  
  H=Union(Z,Union(Set(H,card>k)),Prod(classif,Set(H,card=k)),se
  q(Set(H,card=j),j=2..k-1)),
          classif=Epsilon},labelled];
  
  dHH:=subs(u=1,diff(RootOf(subs({Z(z,u)=z,classif(z,u)=u,H(z,u
  )=H},
  
  H(z,u)=subs(gfeqns(op(2..3,spec),z,[[u,classif]]),H(z,u))),H)
  ,u));
      print(dHH);
  
  seq(evalf(coeff(series(subs(u=1,dHH),z,27),z,n)/count(spec,si
  ze=n)*n!/n,5),n=1..25)
  end:
> deg_hier(2);
```

$$\frac{\mathrm{RootOf}(4\,\_Z-2\,z-2\,\mathbf{e}^{\_Z}+2)^2}{4-2\,\mathbf{e}^{\mathrm{RootOf}(4\,\_Z-2\,z-2\,\mathbf{e}^{\_Z}+2)}}$$

0, .50000, .50000, .52885, .54661, .55875, .56754, .57419, .57940, .58358, .58702, .58989, .59233, .59442, .59623, .59782, .59922, .60047, .60159, .60260, .60351, .60434, .60510, .60580, .60644

```
> deg_hier(3);
```

$$\frac{\mathrm{RootOf}(12\_Z-6z-6\,\mathbf{e}^{\_Z}+6)^3}{12-6\,\mathbf{e}^{\mathrm{RootOf}(12\_Z-6z-6\,\mathbf{e}^{\_Z}+6)}}$$

0, 0, .083333, .096154, .10593, .11234, .11689, .12028, .12291, .12501, .12672, .12815, .12935, .13038, .13127, .13205, .13274, .13335, .13390, .13439, .13484, .13524, .13561, .13595, .13626

```
> deg_hier(4);
```

$$\frac{\mathrm{RootOf}(48\_Z-24z-24\,\mathbf{e}^{\_Z}+24)^4}{48-24\,\mathbf{e}^{\mathrm{RootOf}(48\_Z-24z-24\,\mathbf{e}^{\_Z}+24)}}$$

0, 0, 0, .0096154, .012712, .014838, .016323, .017424, .018272, .018947, .019497, .019954, .020339, .020668, .020953, .021202, .021422, .021617, .021791, .021947, .022089, .022217, .022335, .022442, .022541

```
> deg_hier(5);
```

$$\frac{\mathrm{RootOf}(240\_Z-120z-120\,\mathbf{e}^{\_Z}+120)^5}{240-120\,\mathbf{e}^{\mathrm{RootOf}(240\_Z-120z-120\,\mathbf{e}^{\_Z}+120)}}$$

0, 0, 0, 0, .00084746, .0012718, .0015813, .0018135, .0019944, .0021393, .0022580, .0023570, .0024408, .0025127, .0025751, .0026297, .0026778, .0027207, .0027591, .0027936, .0028248, .0028533, .0028792, .0029030, .0029249

Thus a random classification on n elements seems to have on average about

- about .6 n binary nodes;
- about .14 n ternary nodes;
- about .02 n quaternary nodes.

These results are consistent with the proved result that the total number of internal nodes is on average .79 n. The simple pattern revealed by this computation suggests a formal proof (by singularity analysis) that the distribution of degrees in fact obeys a modified Poisson law. The following theorem, first found while developing this worksheet, appears to be new:

Theorem. *The probability that a random internal node in a random hierarchy of size n has degree k satisfies asymptotically a truncated Poisson law*

```
> tau:=sqrt(2*log(2)-1);
  S:=expand(sum(exp(-tau)*tau^(k-1)/(k-1)!,k=2..infinity));
  Pr(deg=k)=normal(1/S*exp(-tau)*tau^(k-1)/(k-1)!);
```

$$\tau := \sqrt{-1+2\ln(2)}$$

$$S := 1-\frac{1}{\mathbf{e}^{(\sqrt{-1+2\ln(2)})}}$$

$$\mathrm{Pr}(deg=k)=\frac{\mathbf{e}^{(\sqrt{-1+2\ln(2)})}\,\mathbf{e}^{(-\sqrt{-1+2\ln(2)})}\,(\sqrt{-1+2\ln(2)})^{(k-1)}}{(\mathbf{e}^{(\sqrt{-1+2\ln(2)})}-1)\,(k-1)!}$$

Equivalently, the mean number of nodes of degree $2 \le k$ is asymptotic to

```
> C_classif/S*exp(-tau)*tau^(k-1)/(k-1)!;
```

$$-\frac{(\ln(2)-1)\, n\, \mathbf{e}^{(-\sqrt{-1+2\ln(2)})}\, (\sqrt{-1+2\ln(2)})^{(k-1)}}{(-1+2\ln(2))\left(1-\frac{1}{\mathbf{e}^{(\sqrt{-1+2\ln(2)})}}\right)(k-1)!}$$

Numerically, this evaluates to

```
> evalf([seq(",k=2..10)]);
```

$[.5729032234\, n, .1780370765\, n, .03688488077\, n, .005731226563\, n, .0007124210721\, n, .00007379801670\, n, .6552481970\ 10^{-5}\, n, .5090671021\ 10^{-6}\, n, .3515537274\ 10^{-7}\, n]$

These figures are consistent with what was found on sizes near 20. They show that nodes of degree 5 and higher have negligible chances of occurring.

Alternative models

Unlabelled hierarchies

A number of related models can be similarly analyzed. We examine here:

- Unlabelled hierarchies: these represent the types of trees when one considers the elements to be classified as "indistinguishable". What we obtain is then reminiscent of chemical molecules (with an unrealistic element that would be capable of an arbitray valency).
- Planar hierarchies, where one distinguishes the order between decsendents of classification node.

For unlabelled, hierarchies, we just need to change the qualifier of specifications to "unlabelled".

```
> hier4:=[H,{H=Union(Z,Set(H,card>1))},unlabelled]:
> ureduce:=proc(e) eval(subs({Set=proc() {[args]}
  end,Prod=proc() ``(args) end, Sequence=proc() [args]
  end},e)) end:
> ureduce(draw(hier4,size=20));
```

{[{[{[{[Z, Z, {[Z, {[Z, Z]}]}]}]}, {[Z, Z, Z]}, Z, Z, Z]}, {[Z, Z, {[Z, Z]}]}]}, {[{[Z, Z]}, {[Z, {[Z, Z]}]}]}]}

Notice that internally, the setting up of counting tables is more complex as it involves a fragment of Polya's theory. The counting results grow much more slowly, since we distinguish fewer configurations.

```
> seq(count(hier4,size=j),j=0..30);
```

0, 1, 1, 2, 5, 12, 33, 90, 261, 766, 2312, 7068, 21965, 68954, 218751, 699534, 2253676, 7305788, 23816743, 78023602, 256738751, 848152864, 2811996972, 9353366564, 31204088381, 104384620070, 350064856815, 1176693361956, 3963752002320, 13378623786680, 45239588651121

```
> for j to 6 do j,map(ureduce,allstructs(hier4,size=j)) od;
```

1, [Z]

2, [{[Z, Z]}]

3, [{[Z, Z, Z]}, {[Z, {[Z, Z]}]}]

4, [{[Z, Z, {[Z, Z]}]}, {[{[Z, Z]}, {[Z, Z]}]}, {[Z, Z, Z, Z]},
{[Z, {[Z, {[Z, Z]}]}]}, {[{[Z, Z, Z]}, Z]}]

5, [{[Z, Z, {[Z, {[Z, Z]}]}]}, {[{[Z, Z]}, {[Z, {[Z, Z]}]}]},
{[{[{[Z, Z]}, {[Z, Z]}]}, Z]}, {[Z, {[Z, Z, Z, Z]}]}, {[Z, Z, Z, Z, Z]},
{[{[Z, Z, Z]}, Z, Z]}, {[{[Z, Z, {[Z, Z]}]}, Z]}, {[Z, {[Z, {[Z, {[Z, Z]}]}]}]},
{[{[Z, Z, Z]}, {[Z, Z]}]}, {[Z, Z, Z, {[Z, Z]}]}, {[Z, {[Z, Z]}, {[Z, Z]}]},
{[Z, {[{[Z, Z, Z]}, Z]}]}]

6, [{[{[Z, Z, {[Z, {[Z, Z]}]}]}, Z]}, {[{[{[Z, Z]}, {[Z, {[Z, Z]}]}]}, Z]},
{[{[Z, Z, Z]}, {[Z, {[Z, Z]}]}]}, {[Z, Z, {[Z, {[Z, {[Z, Z]}]}]}]},
{[{[Z, {[Z, Z]}]}, {[Z, {[Z, Z]}]}]}, {[Z, Z, {[{[Z, Z, Z]}, Z]}]},
{[{[Z, Z]}, {[Z, {[Z, {[Z, Z]}]}]}]}, {[{[Z, Z, {[Z, Z]}]}, {[Z, Z]}]},
{[Z, {[Z, Z, Z, {[Z, Z]}]}]}, {[{[{[{[Z, Z]}, {[Z, Z]}]}, Z]}, Z]},
{[{[Z, Z, Z]}, Z, Z, Z]}, {[{[Z, Z, Z]}, Z, {[Z, Z]}]},
{[{[Z, Z, Z]}, {[Z, Z, Z]}]}, {[{[Z, Z]}, {[Z, Z, Z, Z]}]},
{[Z, {[Z, {[Z, {[Z, {[Z, Z]}]}]}]}]}, {[Z, {[Z, {[Z, Z, Z, Z]}]}]},
{[{[Z, Z, {[Z, Z]}]}, Z, Z]}, {[{[{[Z, Z]}, {[Z, Z]}]}, Z, Z]},
{[Z, {[Z, Z, Z, Z, Z]}]}, {[Z, {[Z, {[{[Z, Z, Z]}, Z]}]}]},
{[Z, Z, {[Z, Z]}, {[Z, Z]}]}, {[{[{[Z, Z]}, {[Z, Z]}]}, {[Z, Z]}]},
{[Z, Z, Z, Z, {[Z, Z]}]}, {[Z, {[{[Z, Z, Z]}, {[Z, Z]}]}]},
{[Z, {[Z, Z]}, {[Z, {[Z, Z]}]}]}, {[Z, {[Z, {[Z, Z]}, {[Z, Z]}]}]},
{[{[Z, Z]}, {[Z, Z]}, {[Z, Z]}]}, {[{[Z, Z]}, {[{[Z, Z, Z]}, Z]}]},
{[Z, {[{[Z, Z, Z]}, Z, Z]}]}, {[Z, Z, Z, {[Z, {[Z, Z]}]}]},
{[Z, {[{[Z, Z, {[Z, Z]}]}, Z]}]}, {[Z, Z, Z, Z, Z, Z]}, {[Z, Z, {[Z, Z, Z, Z]}]}]

Planar hierarchies

We only need to change Set into Sequence to get the right classification:

```
> hier5:=[H,{H=Union(Z,Sequence(H,card>1))},unlabelled]:
> ureduce:=proc(e) eval(subs({Set=proc() {[args]}
  end,Prod=proc() ``(args) end, Sequence=proc() [args]
  end},e)) end:
> ureduce(draw(hier5,size=50));
```

[[Z, Z, [Z, Z], Z, [
[Z, Z, [[Z, Z, [Z, Z, Z, [Z, Z]], Z, [[Z, Z], Z, [Z, Z, Z]]], Z, [Z, Z]]], [[
[Z, [Z, [[Z, Z], Z]]]],
[[[Z, Z, Z], [[Z, Z, Z], Z, Z, Z], Z, Z], [[[Z, Z], [Z, Z]], [[Z, Z], Z]], Z]], Z]]], Z]

The counting sequence is

```
> seq(count(hierS,size=j),j=0..30);
```

0, 1, 1, 3, 11, 45, 197, 903, 4279, 20793, 103049, 518859, 2646723, 13648869, 71039373, 372693519, 1968801519, 10463578353, 55909013009, 300159426963, 1618362158587, 8759309660445, 47574827600981, 259215937709463, 1416461675464871, 7760733824437545, 42624971294485657, 234643073935918683, 1294379445480318899, 7154203054548921813, 39614015909996567325

This is found as Sequence **M2898** in the *Encyclopedia of Integer Sequences* by Sloane and Plouffe and is known as Schroeder's second sequence.This sequence has a dignified history and Stanley noticed recently that the element count(*hier5*, *size* = 10) = 103049 already appears in Plutarch's [AD50- AD120 (!)] biographical notes on Hipparchus.

```
> gfsolve(op(2..3,hier5),z);
```

$$\{\mathrm{Z}(z)=z,\ \mathrm{H}(z)=\frac{1}{4}+\frac{1}{4}z-\frac{1}{4}\sqrt{1-6z+z^2}\}$$

```
> H5_z:=subs(",H(z));
```

$$H5\_z:=\frac{1}{4}+\frac{1}{4}z-\frac{1}{4}\sqrt{1-6z+z^2}$$

```
> series(H5_z,z=0,11);
```

$$z+z^2+3z^3+11z^4+45z^5+197z^6+903z^7+4279z^8+20793z^9+103049z^{10}+\mathrm{O}(z^{11})$$

Here is finally one quick way to obtain a simple recurrence for these numbers: first guess the recurrence, then check your guess. This, and many alternatives are encapsulated in the Gfun package.

```
> with(gfun):
  listtorec([seq(count(hier5,size=j),j=0..30)],u(n));
```

$$[\{(n-n^2)\,\mathrm{u}(n)+(5n+7n^2)\,\mathrm{u}(n+1)+(-18-23n-7n^2)\,\mathrm{u}(n+2)+(6+5n+n^2)\,\mathrm{u}(n+3),\ \mathrm{u}(2)=1,\ \mathrm{u}(0)=0,\ \mathrm{u}(1)=1\},\ ogf]$$

```
> rectodiffeq(op(1,"),u(n),Y(z));
```

$$\{\mathrm{Y}(0)=0,\ \mathrm{D}(Y)(0)=1,\ -2\,\mathrm{Y}(z)+(2z+2)\left(\frac{\partial}{\partial z}\mathrm{Y}(z)\right)+(z^3-7z^2+7z-1)\left(\frac{\partial^2}{\partial z^2}\mathrm{Y}(z)\right)\}$$

```
> dsolve(",Y(z));
```

$$\mathrm{Y}(z)=\frac{1}{4}+\frac{1}{4}z-\frac{1}{4}\sqrt{1-6z+z^2}$$

Conclusion

Various models of random classification trees can be analysed both theoretically and empirically. Random generation is easy and the experiments lead to new conjectures (like the degree distribution) and even theorems (like the analysis of the number of classification stages). Returning to statistics, some properties of random trees appear to be present accross all models:

for instance nodes of even moderately large degrees, $5 \leq deg$, are highly infrequent, and branching is predominantly binary. General observations of this type may be used to help distinguish classification trees without informational content ("random" trees) from meaningful ones.