Sudoku Players' Forums

THE REAL DISTRIBUTION OF MINIMAL PUZZLES

Goto page Previous 1, 2, 3 ... , 13, 14, 15 Next

newtopic postreply **Sudoku Players' Forums Forum Index -> General/puzzle**

View previous topic :: View next topic

Author	Message
denis_berthier Joined: 19 Jun 2007 Posts: 793 Location: Paris, France	Posted: Sat Jul 25, 2009 4:00 pm Post subject: quote edit **Red Ed wrote:** Mmm... not so fast. You wrote "exactly the same". The distributions are not exactly the same. I concur that they're close enough that almost no-one should ever care about the difference, though. Back to top profile pm www
Red Ed Joined: 06 Jun 2005 Posts: 695	Posted: Sat Jul 25, 2009 4:42 pm Post subject: quote Surprised, Denis? You shouldn't be. I tried to make it clear from the start of the experiment that I was aiming to prove the *existence* of a dependency on the solution grid source, without any claim about the *size* of the effect on the proportion estimates: **on the top-down sensitivity thread, I wrote:** There is some debate about whether the proportion estimates depend in any way upon the solution grid generation part, provided of course that the solution grid generator is a reasonably good/respected one like *suexg*. I'll try to show that the answer is "yes, the solution grid generator makes a difference". Note that I am not claiming that the difference is large, only that it is statistically significant, *i.e.* real and not just an artefact of sample variance. If you had not used the word "exactly" then I would not have been so sure of myself and might never have started the top-down sensitivity topic! But, hey, maybe this is just nit-picking again. When you are quoting estimates from 10000 puzzles, I doubt that it makes any practical difference (for the proportions question) whether *suexg* or an unbiased solution grid source is used,

because the variance due to the source will -- I think -- be small compared to variance due to the limited sample size. And since the latter type of variance is always going to be pretty high (because we're both pretty bad at finding 29- and 30-clue minimals), the upshot is that the bias in *suexg* is probably never really going to bite you. It'll bite me, because I want to count minimals, not just get their relative proportions; but it won't bite you.

Truce?

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 793
Location: Paris, France

Posted: Sat Jul 25, 2009 5:10 pm Post subject: quote edit

> **Red Ed wrote:**
> maybe this is just nit-picking again.

Would you? 😀

Back to top profile pm www

Red Ed

Joined: 06 Jun 2005
Posts: 695

Posted: Sat Jul 25, 2009 5:23 pm Post subject: quote

Is that a request?! 😀

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 793
Location: Paris, France

Posted: Wed Aug 26, 2009 7:07 am Post subject: quote edit

UPDATED RESULTS WITH THE CONTROLLED-BIAS GENERATOR

While I was away, the controlled-bias generator has slowly continued its work. It has now produced 50,000 minimal puzzles. In order to do this, it had to consider about 11 billion complete grids.

Here are the updated results (almost unchanged).

Number of clues of minimal puzzles:
unbiased-mean = 26.57

SER
unbiased-mean = 4.49 unbiased-sd = 2.53

NRCZT
unbiased-mean = 2.31 unbiased-sd = 1.38

As could be expected, there are more fluctuations in the distribution of clues,

especially in the tail, but it is nevertheless reasonably stable.
For comparison with the standard, non-controlled, top-down generator:
- the second column ("top-down") recalls the results obtained from the 1,000,000 puzzles generated with suexg-x.x (sudogen0_1M);
- the third column ("controlled") gives the raw results for the sample of 50,000 puzzles from the controlled-bias generator;
- the fourth column ("unbiased") gives the unbiased results, obtained by applying the correction factors; it is scaled to 1,000,000 puzzles in order to facilitate comparison with the classical top-down generators.
The mean values for each case are also recalled.
Of course, "top-down" appears to be much more biased in favour of fewer clues than "controlled".

Code:

```
#clues   top-down    controlled      unbiased
19       0            0              0.0 (*)
20       44           0              0.0 (*)
21       2428         2              0.46 (*)
22       34548        62             38.58 (*)
23       172512       975            1556
24       342335       6057           23368
25       297838       15327          134819
26       122116       16741          317169
27       25315        8429           325300
28       2686         2131           158609
29       168          260            35367
30       10           16             3772 (*)
31       0            0              0.0 (*)

mean     24.38        25.66         26.57

* values relying on a small sample should be taken
with caution.
```

Back to top profile pm www

Red Ed

Joined: 06 Jun 2005
Posts: 695

Posted: Wed Aug 26, 2009 9:51 pm Post subject: quote

Going by those figures, the number-of-clues distribution is a remarkably good fit to Normal(26.55,1.125) .

Can you think of any heuristic that might explain the observed Normal distribution?
- *i.e. the family, Normal, not the particular parameters*

Back to top profile pm

denis_berthier

Joined: 19 Jun 2007
Posts: 793
Location: Paris, France

Posted: Thu Aug 27, 2009 4:55 am Post subject: quote edit

Red Ed wrote:

> Going by those figures, the number-of-clues distribution is a remarkably good fit to Normal(26.55,1.125) .
>
> Can you think of any heuristic that might explain the observed Normal distribution?
> - *i.e. the family, Normal, not the particular parameters*

As Normal is unbounded, it can only be an approximation of the real law, which seems to make any heuristic justification of Normal unlikely.
Considering instead a bounded law, Binomial could also be a good fit to the above distribution. But I can't see any heuristic to justify it either. In addition, although it is bounded, we have the same problem as for Normal because it is not correctly bounded downwards (it allows 1-, 2-, ... 16- clue minimal puzzles).

Given the above results, the question of finding an analytical law for minimal puzzles amounts to asking: is there any analytical law which:
- is correctly* bounded downwards,
- is correctly** bounded upwards,
- can be approximated by N(26.55, 1.125)?

* correctly =? no 16-clue
** correctly =?

At the present time, I can't imagine any such law.

FYI, I let the controlled-bias generator run a little longer (mainly because my computer is not currently overloaded), but I'm not really expecting anything new from it (except perhaps a better precision for the proportions of 20- and 30- clue minimal puzzles). I would have liked it to find a 31-clue, but this seems unlikely.

Back to top profile pm www

Red Ed

Posted: Thu Aug 27, 2009 6:52 am Post subject: quote

Joined: 06 Jun 2005
Posts: 695

Well, sure, Normal's obviously not the true distribution (for a start it's continuous!) -- but it's a good fit in the middle range and that is the thing that surprised me. It's not as though there's an obvious independent additive effect going on through which, by the Central Limit Theorem, we could say the distribution is approximately Normal.

Perhaps I'll try modelling the "controlled" column with a discrete distribution later on. (btw, Binomial's a poor fit for the "unbiased" column.)

Back to top profile pm

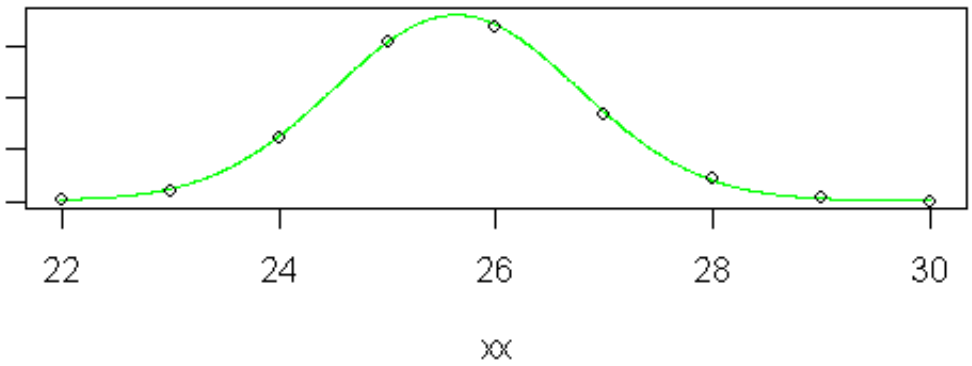
Red Ed

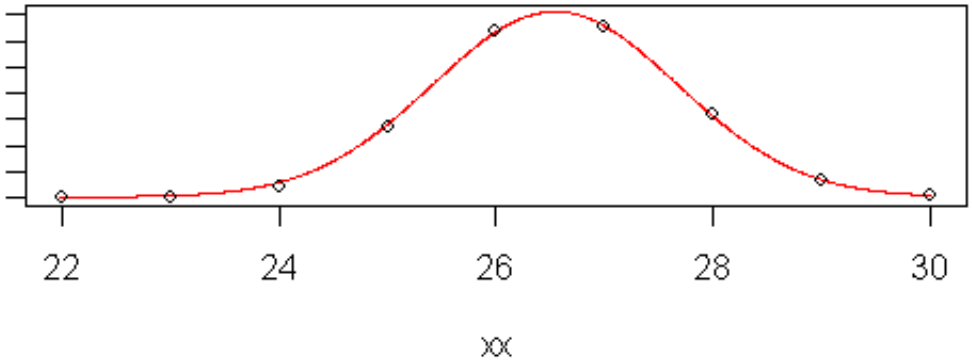
Posted: Thu Aug 27, 2009 12:38 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 695

It turns out that both the "controlled" (green, below) and "unbiased" (red) columns are well approximated by the Normal distribution.

Still no closer to understanding why, though.





Back to top profile pm

Red Ed

Posted: Thu Aug 27, 2009 1:19 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 695

A possibly even better fit to the "controlled" column is Gamma(shape=524,scale=0.0489).

That's equivalent to the total waiting time for 524 exponential random variables each of which has mean 0.0489. (And since it's a long sum of independent rvs, CLT implies near-normality.) I suppose conceivably there's a heuristic along the lines of waiting times for coverage of all unavoidable sets, but I doubt that I can make that idea at all precise.

Still no ideas about discrete distributions.

Back to top profile pm

Red Ed

Posted: Fri Aug 28, 2009 8:19 pm Post subject: quote

Joined: 06 Jun 2005
Posts: 695

denis_berthier wrote:

> FYI, I let the controlled-bias generator run a little longer (mainly because my computer is not currently overloaded), but I'm not really expecting anything new from it (except perhaps a better precision for the proportions of 20- and 30- clue minimal puzzles). I would have liked it to find a 31-clue, but this seems unlikely.

The Gamma distribution is such a good fit that you can solve for its parameters exactly given only the three highest counts in the "controlled" column (for clues=25,26,27) and it will make pretty good predictions as far as clues=30:

Code:

```
clues=22 : gamma=57.5  actual=62
clues=23 : gamma=969   actual=975
clues=24 : gamma=6072  actual=6057
clues=25 : perfect fit
clues=26 : perfect fit
clues=27 : perfect fit
clues=28 : gamma=2070  actual=2131
clues=29 : gamma=261   actual=260
clues=30 : gamma=17.6  actual=16
clues=31 : gamma=0.67  actual=0
```

... so maybe you weren't far off finding a 31-clue minimal after all.

The fit isn't great back as far as clues=17, in the sense that it predicts orders of magnitude too many 17s, but then we can't hope for miracles. (And, no, the situation isn't much better with the unbiased grid source.)

Back to top

denis_berthier

Posted: Fri Sep 18, 2009 4:55 am Post subject:

Joined: 19 Jun 2007
Posts: 793
Location: Paris, France

A BETTER PROOF OF THE MAIN FORMULA FOR THE CONTROLLED-BIAS GENERATOR

Reading again what I had written about the controlled-bias generator (here: http://www.sudoku.com/boards/viewtopic.php?t=14615&start=134), I noticed that the proof was not completely correct. To make it correct, one has to modify slightly my definition of the forest of puzzles. Instead of indexed puzzles, it will be made of doubly indexed puzzles. This mathematical trick is only useful to take account of what happens below B (i.e. in the virtual part of the generator). Here is the correct proof.

Let us introduce the notion of a doubly indexed puzzle. We consider only (single or multi solution) consistent puzzles P. The double index of a doubly indexed puzzle P has a clear intuitive meaning: the first index is one of its solution grids and the second index is a sequence (notice: not a set, but a sequence, i.e. an ordered set) of clue deletions leading from this solution to P. In a sense, the double index keeps track of the generation process.
Given a doubly indexed puzzle Q, there is an underlying singly-indexed puzzle: the ordinary puzzle obtained by forgetting the second index of Q, i.e. by remembering the solution grid from which it came and by forgetting from which sequence of deletions Q was reached from this solution.
Given a doubly indexed puzzle Q, there is also a non indexed puzzle, obtained by forgetting the two indices.
Notice that, for a single solution doubly indexed puzzle, the first index is useless as it can be computed from the puzzle; in this case singly indexed and non indexed are equivalent. (In terms of the generator, it could as well output minimal puzzles or couples minimal-puzzle-plus-solution.)

Consider now the following layered structure (a forest of trees with branches pointing downwards), the nodes being (single or multi solution) doubly indexed puzzles:
- floor 81 : the N different complete solution grids (considered as puzzles), each indexed by itself and by the empty sequence; notice that all the puzzles at floor 81 have 81 clues;
- floor 80: each doubly indexed puzzle Q at floor 81 sprouts 81 branches pointing to floor 80, one for each clue C in Q; the other end of this C branch will be the doubly indexed puzzle obtained from Q by removing clue C and indexed by the same complete grid as Q and by the 1-element sequence (C); notice that all the puzzles at floor 80 have 80 clues;
- recursive step: given floor n+1 (each doubly indexed puzzle of which has n+1 clues and is indexed by a complete grid that solves it and by a sequence of length 81-(n+1)), build floor n as follows:
each doubly indexed puzzle Q at floor n+1 sprouts n+1 branches; for each clue C in Q, there is a branch leading to a doubly indexed puzzle R at floor n: R is obtained from Q by removing clue C; its first index is identical to that of Q and its second index is the (81-n)-element sequence obtained by appending C to the end of the second index of Q; notice that all the doubly indexed puzzles at floor n have n clues and the length of their second index is equal to 1 + (81-(n+1)) = 81-n.

It is easy to see that, at floor n, each doubly indexed puzzle has an underlying singly indexed puzzle identical to that of (81 - n)! doubly indexed puzzles with the same first index at the same floor (including itself).
This is equivalent to saying that, at any floor n < 81, any singly-indexed puzzle Q can be reached by exactly (81 - n)! different paths from the top (all of which start necessarily from the complete grid defined as the first index of Q). These paths are the (81 - n)! different ways of deleting one by one its missing 81-n clues from its solution grid.
Notice that this would not be true for non indexed puzzles that have multiple solutions. This is where the first index is useful.

Let N be the number of complete grids. At each floor n, there are:
N * 81! / n! doubly indexed puzzles,
N * 81! / (81-n)! / n! singly indexed puzzles.
For each n, there is therefore a uniform probability P(n) = 1/N * 1/81! * (81-n)! * n! that a singly indexed puzzle Q at floor n is reached by a random (uniform) search starting from the associated complete grid (its first index) at the top.

What is important here is the ratio: P(n+1) / P(n) = (n + 1) / (81 - n).
This formula is valid globally if we start from all the complete grids, as above, but it is also valid for all the single solution puzzles if we start from a single complete grid (just forget N in the proof above). (Notice however that it is not valid if we start with a subgrid instead of a complete grid.)

Now, call B the set of (non indexed) minimal puzzles. On B, all the puzzles are minimal. Any puzzle strictly above B has redundant clues and a single solution. Notice that, for all the puzzles on B and above B, singly indexed and non indexed puzzles are in one-to-one correspondence.
On the set B of minimal puzzles there is a probabily Pr naturally induced by the different Pn's (and renormalised to sum up to 1) and it is the probability that a minimal puzzle Q is reached by our controlled-bias generator. It is defined, up to a multiplicative constant k by Pr(Q) = k P(n), if Q has n clues. k must be chosen so that the probabilities of all the minimal puzzles sum up to 1.
But we need not know k. What is important here is that, by construction of Pr on B (a construction which models the workings of the controlled bias generator), the relation **Pr(n+1) / Pr(n) = (n + 1) / (81 - n)** holds for any two minimal puzzles, with respectively n+1 and n clues.

The rest of the original post is unchanged.

Back to top profile pm www

Red Ed

Posted: Fri Sep 18, 2009 6:55 am Post subject: quote

Joined: 06 Jun 2005
Posts: 695

Here's a more direct proof.

- Let Z ~= 6.67e21 be the total number of all solution grids
- Consider the following algorithm equivalent to yours in its output:
  - **1:** Pick a random solution grid G
  - **2:** Remove clues one at a time until nothing's left, recording the 81 subgrids so created
  - **3:** Output the unique subgrid (if any) that is a proper minimal puzzle
- A particular proper minimal puzzle, P, with n clues is output by the algorithm iff:
  - its solution grid is picked in step **1** (prob = 1/Z); and
  - the first 81-n clues removed in step **2** are the empty cells in P (prob = 1/choose(81,n))
  - So Pr(n) = 1 / (choose(81,n)*Z)
- So Pr(n+1)/Pr(n) = choose(81,n)/choose(81,n+1) = (n+1)/(81-n)

Back to top profile pm

denis_berthier

Posted: Fri Sep 18, 2009 7:03 am Post subject: quote edit

Joined: 19 Jun 2007
Posts: 793
Location: Paris, France

It isn't really more direct, it's just the same proof without the details.

Anyway, what's important is the definition of the controlled-bias generator and the result P(n+1)/P(n).

Back to top profile pm www

David P Bird

Joined: 17 Sep 2008
Posts: 132
Location: Middle England

Posted: Fri Sep 18, 2009 7:22 am Post subject: quote

It’s a question of view point isn't it? We can either try to visualise a filing system of 6.67e21 solution grids and scratch our heads about the order we should delete the givens from one of them, or an upside down-forest of trees. Pity Salvador Dali isn't with us any more.

Back to top profile pm

Display posts from previous: All Posts Oldest First Go

newtopic postreply Sudoku Players' Forums Forum Index -> General/puzzle

All times are GMT

Goto page Previous 1, 2, 3 ... , 13, **14**, 15 Next

Page 14 of 15

Stop watching this topic

Jump to: General/puzzle Go

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** vote in polls in this forum