

Re: Choose k random lines from file

Source: <http://coding.derkeiler.com/Archive/General/comp.programming/2004-05/2054.html>

From: Nick Landsberg (*hukolau_at_NOSPAM.att.net*)

Date: 05/31/04

Date: Mon, 31 May 2004 18:09:59 GMT

Arthur J. O'Dwyer wrote:

> *On Mon, 31 May 2004, Gerry Quinn wrote:*

>

>> *gerryq@DELETETHISindigo.ie says...*

>>

>>> *ajo@nospam.andrew.cmu.edu says...*

>>>>

>>>> *I don't like to use potentially-infinite procedures when there*

>>>> *are algorithms available, no matter how fast they are. Call it*

>>>> *an ideological position. ;)*

>>

>> *Something interesting just occurred to me – if you want an even*

>> *distribution you may *need* a non-deterministic algorithm, because the*

>> *alternative will be an infinite, or at least horribly long one!*

>>

>> *How so? Well, suppose you have seventeen pages, and you want to choose*

>> *one at random, but your RNG has a period that's a power of two, or*

>> *anyway some number that's not a multiple of seventeen.*

>

>

> *It's not the period that would be the problem; it's the range.*

> *Suppose our RNG (doesn't even have to be a PRNG, but any RNG) gives*

> *us numbers in the range $0..2^N-1$. And we want a number in the range*

> *$0..16$ (seventeen choices total). There's no algorithm to give us*

> *that number without a bias. None. There are procedures much less*

> *wasteful than the canonical*

>

> *again:*

> *R = rand();*

> *if (R >= (RAND_MAX/17)*17) goto again;*

> *return (R % 17);*

>

> *[which I see you showed below],*

> *but there's no way to get an unbiased number with a range of M*

> *out of an unbiased generator with a range of N, if M and N are*

> *coprime. Or something like that. I'm not going to prove my*

> *wild assertions on a federal holiday! ;)*

comp.programming: Re: Choose k random lines from file

I'm no expert, but a thought experiment using small values for N and M (say 7 and 3) seems to indicate a bias towards the low end.

On the other hand, would something like the following give less of a bias using an extra float or double conversion? :

```
R = rand();
X = (double) R / (RAND_MAX+1); /* X is of type double*/

/* this assumes that (RAND_MAX +1) does not
cause integer overflow (UB). The result should give a
uniform distribution between 0.0 <= X < 1.0 */

R2 = X * M; /* R2 and M are the appropriate integer types */

/* I think this should yield a value 0 <= R2 < M
with a uniform distribution */
```

Or did I get it wrong?

```
>
>
>>To get an equally good deterministic algorithm, I suppose you could call
>>rand() seventeen times and add them. To be honest I'm not 100% sure
>>that this is valid, but anyway if you have a lot of pages the amount of
>>calculation rises sharply...
>
>
> Not valid. That gets a random number in the range 0..(17*RAND_MAX),
> all right, but it's no longer unbiased. There are a lot of numbers
> clustered around 8.5*RAND_MAX, and only one way to get 0. So just
> taking the sum of 17 rand()s and modding by 17 doesn't get you an
> unbiased distribution either.
```

As I remember from decades back, summing uniformly distributed random numbers yields something like a Gaussian distribution for the sum. IIRC, the more numbers you sum up, the smaller the standard deviation of the sum will be.

```
>
>
>
>>Creating a PRNG with a period of 17 has been suggested.
>
>
> Has it? :) s/period/range/ and you have a point that I don't
> recall being made in this thread; but I'm not sure such a beast
> exists for any given prime. Sounds like a job for Knuth to me!
```

Re: Choose k random lines from file

comp.programming: Re: Choose k random lines from file

>
> *-Arthur*
>

Nick L.

--
"It is impossible to make anything foolproof
because fools are so ingenious"
- A. Bloch