91.71 Deleting Blocks of Names from a List
Author(s): Martin Griffiths
Source: *The Mathematical Gazette*, Vol. 91, No. 522 (Nov., 2007), pp. 540-546
Published by: The Mathematical Association
Stable URL: http://www.jstor.org/stable/40378437
Accessed: 07-05-2018 16:41 UTC

(Unfortunately my referee disagrees with this over-optimistic estimation as to what constitute 'minor quirks', and has consequently downgraded my 'proof' to a mere 'argument'. While I concur, I still enjoyed writing the article.)

Reference

1. Georg Pick, Sitzungber, Lotos, *Naturwissen Zeitschrift*, Prague, **19** (1899) pp. 311-319.

J. TRAININ

Edificio La Roca 27, Paseo del Altillo 11, 18690 Almuñécar, Granada, Spain

e-mail: *traininjohn2@telefonica.net*

91.71 Deleting blocks of names from a list

Introduction

In order to write school reports for my sixth form classes I am provided with a list, in the form of a table within an electronic document, of all the students in the relevant year group. Before typing my comments into the appropriate boxes, I remove the names of all the students that I do not teach from this list by deleting the rows that they occupy. Below is a fictitious and very much scaled-down year group consisting of ten students, three of whom I teach (those in bold):

1	Alison Anderson
2	***Elaine Barker***
3	***Peter Brown***
4	George Costanidis
5	Sandeep Desai
6	Paul Hutton
7	Stuart Jackson
8	***Milana Pavlov***
9	Helen Sagar
10	Stephanie Smith

When there is a run of students in the list that I do not teach then I can highlight their names and delete them all in one go. In the above list, for example, I could have eliminated the names of these students in three block-deletions. Having just performed this procedure for my lower sixth classes, I ended up having to carry out 25 block-deletions in order to remove the 74 students that I did not teach from the year group of size 106. I wondered whether I had been somewhat unfortunate in having to perform so many deletions – what would the expected number be in this case, and in the more general situation? Of course, knowledge of the expectation alone is not sufficient to determine how unlucky (or possibly lucky) I had been. We would also need to ascertain the variance of the required number of block-deletions.

An anonymous referee pointed out to me that an equivalent version of this problem does actually appear in the literature, stated within the context of tossing a coin. If, without loss of generality, we replace each of the students that I do teach by a *tail* and each of those that I do not teach by a *head* then the number of block-deletions corresponds to the number of runs of heads. The distribution of the number of runs of heads (when a coin is tossed $h + t$ times, resulting in h heads) and its expectation appear as problems in [1] – Problem 23 of Chapter II and Problem 28 of Chapter IX respectively – although there are no hints or suggestions as to how to proceed.

We split this paper into three sections. The first contains my original solution to the problem, based on the ordered partitioning of positive integers into sums of non-negative integers. In the second section we solve the problem via an application of indicators – this arose as a result of a suggestion by the referee, who also mentioned that Ross [2, p. 312] finds the expectation using this method but does not consider the variance. In the final part we calculate some higher moments of the number of block-deletions and use these results to make a conjecture regarding its distribution when the list becomes large.

The partition method

Suppose that a year group has k students, t of whom I teach and n of whom are not taught by me. We can approach the problem of finding the expected number of block-deletions from the list by considering ordered partitions of n into sums of $t + 1$ non-negative integers. We shall refer to an ordered partition of p into q non-negative integers as a p-q partition. We may easily establish a one-to-one correspondence between n-$(t+1)$ partitions and the selections of t students from the list. For ease of reference let us call students that I do and do not teach t-students and n-students respectively. For $1 \leqslant m \leqslant t - 1$, the $(m + 1)$th term in the partition is equal to the number of n-students between the mth and $(m + 1)$th t-students, while the first and last terms in the partition are equal to the number of students preceding the first t-student and following the last t-student respectively. Zero terms in the sum are created by consecutive t-students, and also by any such student occupying the first or last position in the list. To take an example, the 7-4 partition corresponding to the list given above is $1 + 0 + 4 + 2$. The number of selections of t students from k, and consequently the number of n-$(t + 1)$ partitions, is $\binom{k}{t}$.

Note that, for a particular selection of t t-students from k, the number of zeroes in the corresponding n-$(t + 1)$ partition plus the number of blocks of n-students in the list is equal to $t + 1$. Let us define the random variable Z to be the number of zeroes in the n-$(t + 1)$ partition corresponding to a random selection of t t-students from the list, and let D be the number of blocks of n-students that need to be deleted. We then have that $D = t + 1 - Z$, giving

$\mathrm{E}(D) = t + 1 - \mathrm{E}(Z)$ and $\mathrm{Var}(D) = \mathrm{Var}(Z)$. In order firstly to obtain $\mathrm{E}(D)$ we shall calculate $\mathrm{E}(Z)$ and then use the relationship in the previous sentence.

Suppose that a particular n-$(t + 1)$ partition has exactly m zeroes. If we remove these zeroes then we will have an n-$(t + 1 - m)$ partition for which all the terms are positive. Furthermore, there are $\binom{t+1}{m}$ distinct n-$(t + 1)$ partitions with exactly m zeroes that will, on removing those zeroes, give the same n-$(t + 1 - m)$ partition. There is an obvious one-to-one correspondence between n-$(t + 1 - m)$ partitions for which all the terms are positive and the $(n - t - 1 + m)$-$(t + 1 - m)$ partitions (to obtain the latter simply take 1 from each of the terms of the former). A similar argument to the one used earlier to explain why the number of n-$(t + 1)$ partitions is $\binom{k}{t}$ tells us that the number of $(n - t - 1 + m)$-$(t + 1 - m)$ partitions is $\binom{n-1}{n-t-1+m}$. There are thus $\binom{t+1}{m}\binom{n-1}{n-t-1+m}$ distinct n-$(t + 1)$ partitions with exactly m zeroes, giving us

$$\begin{aligned}\mathrm{E}(Z)\binom{k}{t} &= \sum_{m=1}^{t} m\binom{t+1}{m}\binom{n-1}{n-t-1+m}\\ &= (t+1)\sum_{m=1}^{t}\binom{t}{m-1}\binom{n-1}{n-t-1+m}\\ &= (t+1)\binom{t+(n-1)}{t+(n-t)}\\ &= (t+1)\binom{n+t-1}{n},\end{aligned}$$

on using the results $m\binom{n}{m} = n\binom{n-1}{m-1}$ and $\sum_{k=0}^{r}\binom{r}{k}\binom{s}{n+k} = \binom{r+s}{r+n}$, to be found in [3, p. 53 and p. 58] for example. Note that $\binom{n}{m} = 0$ for integers m and n with $m < 0$ or $m > n$, so it is possible that the sums given above contain terms whose value is equal to zero. From this we obtain

$$\mathrm{E}(Z) = (t+1) \times \frac{n!\,t!}{(n+t)!} \times \frac{(n+t-1)!}{n!\,(t-1)!} = \frac{t(t+1)}{n+t},$$

and thus

$$\mathrm{E}(D) = t + 1 - \frac{t(t+1)}{n+t} = \frac{n(t+1)}{n+t}.$$

For the case in which $n = 74$ and $t = 32$ we have $\mathrm{E}(D) \approx 23$, so it would appear that I had not been particularly unfortunate in having to perform 25 such deletions. However, in order to check this, we ought to calculate the variance of D:

$$\begin{aligned}
\operatorname{Var}(D)\binom{k}{t} &= \operatorname{Var}(Z)\binom{k}{t}\\
&= \sum_{m=1}^{t} m^2\binom{t+1}{m}\binom{n-1}{n-t-1+m} - \binom{k}{t}\left(\frac{t(t+1)}{t+n}\right)^2\\
&= (t+1)\sum_{m=1}^{t} m\binom{t}{m-1}\binom{n-1}{n-t-1+m} - \binom{k}{t}\left(\frac{t(t+1)}{t+n}\right)^2\\
&= (t+1)\sum_{m=1}^{t}\left\{(m-1)\binom{t}{m-1}+\binom{t}{m-1}\right\}\binom{n-1}{n-t-1+m} - \binom{k}{t}\left(\frac{t(t+1)}{t+n}\right)^2\\
&= (t+1)\sum_{m=1}^{t}\left\{t\binom{t-1}{m-2}+\binom{t}{m-1}\right\}\binom{n-1}{n-t-1+m} - \binom{k}{t}\left(\frac{t(t+1)}{t+n}\right)^2\\
&= (t+1)\left\{t\binom{n+t-2}{n}+\binom{n+t-1}{n}\right\} - \binom{k}{t}\left(\frac{t(t+1)}{t+n}\right)^2,
\end{aligned}$$

where we have used, once more, the previously stated results from [3]. From this we obtain, after a considerable amount of simplification, that

$$\operatorname{Var}(D) = \frac{nt(n-1)(n+1)}{(n+t)^2(n+t-1)}.$$

For our particular case we have $\operatorname{Var}(D) \approx 5$, which does tend to confirm that the 25 block-deletions I had to carry out were not a particularly extreme outcome.

The indicator method

We can also approach the problem of finding $\mathrm{E}(D)$ by using random variables called indicator functions. We define the indicator I_r of the event that a run of n-students begins at the rth position, so that

$$I_1 = \begin{cases} 1 & \text{if the first in the list is an n-student} \\ 0 & \text{otherwise} \end{cases}$$

and for $r \geqslant 2$

$$I_r = \begin{cases} 1 & \text{if the } (r-1)\text{th is a t-student and the } r\text{th is an n-student} \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$\mathrm{E}(I_1) = \mathrm{P}(I_r = 1) = \frac{n}{n+t} \text{ and } \mathrm{E}(I_r) = \mathrm{P}(I_r = 1) = \frac{nt}{(n+t)(n+t-1)}$$

for $r \geqslant 2$, from which we obtain

$$\mathrm{E}(D) = \mathrm{E}\left(\sum_{r=1}^{n+t} I_r\right) = \sum_{r=1}^{n+t} (\mathrm{E}(I_r)) = \sum_{r=1}^{n+t} \mathrm{P}(I_r = 1)$$

$$= \frac{n}{n+t} + \sum_{r=2}^{n+t} \frac{nt}{(n+t)(n+t-1)} = \frac{n(t+1)}{n+t}.$$

We can also use these indicator functions to calculate $\mathrm{Var}(D)$, although this requires a little more thought since $\mathrm{Var}(D) = \mathrm{Var}\left(\sum_{r=1}^{n+t} I_r\right)$ and $\sum_{r=1}^{n+t} \mathrm{Var}(I_r)$ are not necessarily equal here because the indicator functions are not independent. For example,

$$\mathrm{P}(I_1)\,\mathrm{P}(I_2) = \frac{n^2 t}{(n+t)^2(n+t-1)} \text{ while } P(I_1 I_2) = 0,$$

the latter being a consequence of the fact that a block of n-students in a list cannot start at both the first and second positions. We have

$$\mathrm{E}(D^2) = \mathrm{E}\left(\left(\sum_{r=1}^{n+t} I_r\right)^2\right) = \sum_{r=1}^{n+t} \mathrm{E}(I_r I_r) + \sum_{|i-j|=1} \mathrm{E}(I_i I_j) + \sum_{|i-j|>1} \mathrm{E}(I_i I_j),$$

where the first sum is equal to $n(t+1)/(n+t)$, the second is equal to zero (as blocks of n-students cannot start at consecutive positions) and the third sum can be evaluated using the fact that

$$\mathrm{E}(I_i I_j) = \mathrm{P}(I_i = 1 \cap I_j = 1) = \begin{cases} \dfrac{nt(n-1)}{(n+t)(n+t-1)(n+t-2)} & \text{if } i = 1 \text{ or } j = 1 \\ \dfrac{nt(n-1)(t-1)}{(n+t)(n+t-1)(n+t-2)(n+t-3)} & \text{otherwise.} \end{cases}$$

This gives us

$$\mathrm{Var}(D) = \mathrm{E}(D^2) - \{\mathrm{E}(\mathrm{D})\}^2$$

$$= \frac{n(t+1)}{n+t} + \frac{2nt(n-1)}{(n+t)(n+t-1)} + \frac{nt(n-1)(t-1)}{(n+t)(n+t-1)} - \left(\frac{n(t+1)}{n+t}\right)^2$$

$$= \frac{nt(n-1)(t+1)}{(n+t)^2(n+t-1)},$$

as before.

Limiting distributions

It is interesting to consider the behaviour of $\mathrm{E}(D)$ and $\mathrm{Var}(D)$ as one or both of n and t become large. On fixing t we see that $\mathrm{E}(D) \to t + 1$ and $\mathrm{Var}(D) \to 0$ as $n \to \infty$, which is as we might expect intuitively since the larger n becomes the less likely we are to find consecutive t-students in the list and the less likely t-students are to occupy the first or last positions in the list. On the other hand, if we allow both n and t to increase without limit, subject to the constraint $n = pt$ for some positive integer p, then we have $\mathrm{E}(D) \to \infty$ and $\mathrm{Var}(D) \to \infty$. We can, however, say something about the relative behaviour of $\mathrm{E}(D)$ and $\mathrm{Var}(D)$ in this case. Since

$$\mathrm{Var}(D) = \frac{nt(n-1)(t+1)}{(n+t)^2(n+t-1)} = \frac{t(n-1)}{(n+t)(n+t-1)}\,\mathrm{E}(D),$$

we see that

$$\frac{\mathrm{E}(D)}{\mathrm{Var}(D)} \to 2 + p + \frac{1}{p} \text{ as } t \to \infty.$$

Finally, let us see if we can obtain any more information about the nature of the scaled limiting distribution of D. In other words, we consider the distribution of $D/\mathrm{Var}(D)$ as n and t become large. In order to do this we calculate scaled higher central moments of D. Adopting the notation and definitions given in [4, p. 51], the kth central moment, σ_k, of D is defined to be $\mathrm{E}\big((D - \mathrm{E}(D))^k\big)$. So, in particular, we have $\sigma_2 = \mathrm{Var}(D)$. In all of what follows we shall assume that $n = pt$ for some positive integer p, and that $n, t \geqslant 2$.

Let us consider the *skewness* and *kurtosis* of D, given by

$$\mathrm{skw}(D) = \frac{\sigma_3}{\left(\sqrt{\sigma_2}\right)^3} \text{ and } \mathrm{kur}(D) = \frac{\sigma_4}{\sigma_2^2}, \text{ respectively.}$$

Skewness provides us with a measure of the degree of asymmetry of a distribution while kurtosis gives us its degree of peakedness (see [5] and [6]). We find, after a considerable amount of effort, that

$$\sigma_3 = \frac{nt(n-1)(t+1)(n-t)(t-n+2)}{(n+t)^3(n+t-1)(n+t-2)}$$

and

$$\sigma_4 = \frac{nt(n-1)(t+1)}{(n+t)^4(n+t-1)(n+t-2)(n+t-3)}f(n,t),$$

where $f(n,t) =$

$$n^4 + \left(3t^2 - 5t - 5\right)n^3 + 3\left(t^3 + 4t + 2\right)n^2 - t\left(11t^2 + 12t + 6\right)n + t^2\left(t^2 + 7t + 6\right),$$

noting that in the case $n = t$ (that is, $p = 1$) the fourth central moment takes the particularly simple form

$$\sigma_4 = \frac{(t-1)(t+1)\left(3t^2 - 4t - 3\right)}{16(2t-1)(2t-3)}.$$

From the above we obtain

$$\mathrm{skw}(D) = \frac{nt(n-1)(t+1)(n-t)(t-n+2)}{(n+t)^3(n+t-1)(n+t-2)} \times \left(\frac{(n+t)^2(n+t-1)}{nt(n-1)(t+1)}\right)^{3/2}$$

$$= \frac{(n-t)(t-n+2)\sqrt{n+t-1}}{(n+t-2)\sqrt{nt(n-1)(t+1)}}.$$

It is clear, on disregarding the trivial cases, that skw $(D) = 0$ if, and only if, $n = t$ or $n = t + 2$. However, if we allow t (and hence n) to increase without limit, we see that $\lim_{t \to \infty} \operatorname{skw}(D) = 0$. We also have

$$\operatorname{kur}(D) = \frac{nt(n-1)(t+1)}{(n+t)^4(n+t-1)(n+t-2)(n+t-3)} f(n,t) \times \left(\frac{(n+t)^2(n+t-1)}{nt(n-1)(t+1)} \right)^2$$

$$= \frac{(n+t-1)}{nt(n-1)(t+1)(n+t-2)(n+t-3)} f(n,t).$$

Then, noting that the dominant term in $f(pt, t)$ is $3p^2(p+1)t^5$, we have

$$\operatorname{kur}(D) = \frac{(t(p+1)-1)(3p^2(p+1)t^5 + O(t^4))}{pt^2(pt-1)(t+1)(t(p+1)-2)(t(p+1)-3)},$$

from which we see that $\lim_{t \to \infty} \operatorname{kur}(D) = 3$.

Any normal distribution has skewness 0 and kurtosis 3 so we might conjecture, on the strength of the above results, that the sequence $\{D/\operatorname{Var}(D) : t = 2,3,4, \ldots\}$ converges to the standard normal distribution, in the sense that the cumulative distribution function of $D/\operatorname{Var}(D)$ converges to that of the standard normal distribution. Note that we cannot appeal directly to the central limit theorem here to provide confirmation of this since the indicator functions used to define D are not independent (nor all identically distributed). It would be interesting to conduct a series of Monte Carlo simulations in order to investigate further the possibility that the limiting distribution is standard normal.

Acknowledgement

I am grateful to the referee for suggesting that I also approach this problem via indicator functions, and for pointing out that the consideration of higher moments might be worthwhile.

References

1. W. Feller, *An Introduction to Probability Theory and its Applications*, 1, Wiley (1968).
2. S. Ross, *A First Course in Probability*, Prentice-Hall (1998).
3. D. E. Knuth, *The Art of Computer Programming*, Volume 1, Addison-Wesley (1968).
4. G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford University Press (2001).
5. E. W. Weisstein, Skewness from MathWorld – A Wolfram web resource. http://mathworld.wolfram.com/Skewness.html.
6. E. W. Weisstein, Kurtosis from MathWorld – A Wolfram web Resource. http://mathworld.wolfram.com/Kurtosis.html.

MARTIN GRIFFITHS

Colchester County High School for Girls, Norman Way, Colchester CO3 3US