Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal., 1(2009), 1 – 10
ISSN 2070-5948, DOI 10.1285/i20705948v2n1p1

ON THE DISTRIBUTION OF THE LARGEST PART OF UNRESTRICTED PARTITIONS OF SMALL INTEGERS

Simon Brown[*]

School of Human Life Sciences, University of Tasmania,
Launceston, Tasmania, Australia

Received 18 June 2009; accepted 6 July 2009
Available online 15 August 2009

Abstract: *Several theoretical estimates of the distribution of the parts of integer partitions have been published. Generally these are asymptotically correct for large integers, but practical applications require that the distribution be known for small integers ($n \le 1000$). The largest part (or the number of parts) of an unrestricted partition of the integer n has the extreme value distribution, in agreement with the theoretical estimates. Expressions approximating the mode and variance of the distribution are given for $n \le 1000$ that represent significant improvements over the asymptotically correct theoretical expressions.*

Keywords: *integer partition, extreme value distribution, approximation.*

1. Introduction

A partition of an integer n is a finite, nonincreasing sequence of positive integers $\lambda_1, \lambda_2, \ldots, \lambda_r$, such that $\sum_{i=1}^{r} \lambda_i = n$, where the λ_i are called the parts of the partition. A partition can be written $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$, for example, one partition of $n = 26$ is $\{8, 6, 6, 5, 1\}$, which has 5 parts, and can be illustrated using a Ferrars graph, in which λ_i dots are drawn on the ith row (Fig 1A). The conjugate of a partition is formed from parts λ_i' which are the number of parts of the partition greater than or equal to i (Fig 1B). There is a one-to-one correspondence between a partition and its conjugate, so that the largest part of one (λ_1) is the number of parts (r) of the other. Therefore, the number of partitions of n with at most m parts ($p(m, n)$) equals $q(m, n)$, the number of

[*] Email: Simon.Brown@utas.edu.au

partitions of *n* in which no part exceeds *m* [1]. Consideration of the distribution of the largest parts of the partitions of *n* is necessarily equivalent to considering that of the number of parts.
Partitions may be applied wherever a uniform sample is distributed among different classes and have been applied in several disciplines, including combinatorics, group theory, algebraic geometry, chemistry, particle physics, lattice theory, population genetics and protein sequence analysis [1, 2, 4, 9, 13, 14]. The largest part of the partitions of *n* (λ_1) tends to an extreme value distribution as $n \to \infty$ [3, 10, 11]. However, the distribution of λ_1 for small *n* has not been considered, despite being used in the analysis of protein sequences [9]. Here I compare the empirical distribution of λ_1 for $n \leq 1000$ with the theoretical estimates, and show that the extreme value distribution can be appropriate for *n* in this range using novel semi-empirical approximations of the mode and the variance, with smaller residuals than the theoretical estimates.

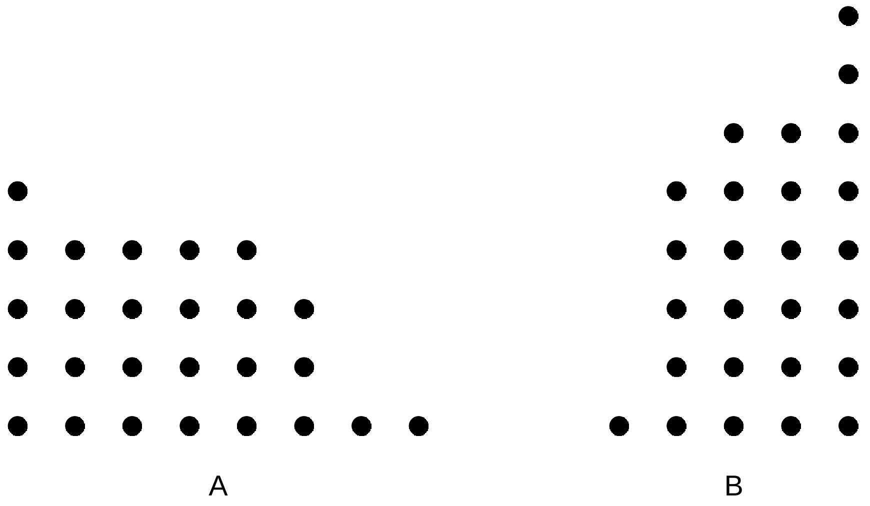

Figure 1. Ferrars graphs of the partition of *n* = 26 {8, 6, 6, 5, 1} (A) and its conjugate {5, 4, 4, 4, 4, 3, 1, 1} (B).

2. Calculation of the distributions

For any *n*, $p(1, n) = 1$ and

$$p(m,n) = q(m,n) = p(m-1,n) + p(m,n-m)\,, \tag{1}$$

Where

$$p(k,l) = \begin{cases} p(l,l), & k \geq l \\ 1, & l = 0 \end{cases}.$$

As is apparent from equation (1), the number of partitions with exactly *m* parts is

$$P(m,n) = p(m,n-m) = P(m,n-m) + P(m-1,n-1) \tag{2}$$

where

$$P(k,l) = \begin{cases} 0, & 0 < k < l \text{ or } k < 0 \\ 1, & k = l = 0 \end{cases}$$

[5].

The calculation of *p*(*m*, *n*) for even quite small values of *n* requires careful consideration because the total number of partitions increases rapidly, for example, *p*(10, 10) = 42, *p*(200, 200) = 3972999029388 and *p*(1000, 1000) = 24061467864032622473692149727991. Here it was unnecessary to list all of the partitions of *n*, because only the number of times that $\lambda_1 = i$, for *i* = 1, 2, …, *n*, occurs is significant and the largest number required (for $n \le 1000$) was *P*(81, 1000) = 4017794288116412246751907682 42.

This was obtained from equation (2) in Python 2.6.2, in which long integers of unlimited precision are available. This code worked for values of *n* of at least 3000, but I restrict the present work to $n \le 1000$. To ensure the correctness of the code four tests were applied:

(i) The same calculations (for $n \le 600$) were carried out in Pascal using quadruple reals.

(ii) Values of $p(n,n) = \sum_{m=1}^{n} P(m,n)$, for $n \le 600$, were compared with published values [1, 5, 6].

(iii) As can be seen by substituting *m*, where $0.5n \le m \le n$, into (1):

$$P(m,n) = p(n-m, n-m), \quad k \le m \le n \quad k = \begin{cases} 0.5n, & n \text{ even} \\ 0.5(n+1), & n \text{ odd} \end{cases} \tag{3}$$

so the values of *P*(*m*, *n*), for $m > 0.5n$, correspond to the values of $p(n-m, n-m)$, a random sample of which was checked against the published values ($n \le 600$).

(iv) For small values of *m*:

$$p(m,n) = \frac{1}{m!(m-1)!} n^{m-1} + R_{m-2}(n),$$

where $R_{m-2}(n)$ is a polynomial in *n* of degree at most $m-2$. For example:

$$\begin{aligned} p(1,n) &= 1 \\ p(2,n) &= \frac{1}{2}n + 1 + (-1)^n \\ p(3,n) &= \frac{1}{12}n^2 + \frac{1}{2}n + \frac{47}{72} + \frac{1}{8}(-1)^n + \frac{1}{9}\left(a_3^n + a_3^{2n}\right) \\ p(4,n) &= \frac{1}{144}n^3 + \frac{5}{48}n^2 + \frac{15}{32}n + \frac{175}{288} + \frac{1}{32}(-1)^n(n+5) \\ &\quad + \frac{i}{9\sqrt{3}}\left(a_3^{n-1} - a_3^{2n-2}\right) + \frac{1}{16}\left(i^n + i^{3n}\right) \end{aligned} \tag{4}$$

where $a_3 = \exp(2i\pi/3)$is a cube root of 1, and expressions for larger values of m are available [6]. The values of $p(m, n)$ were calculated using equation (4) for m = 1, 2, 3 and 4 and compared with the values obtained from the distributions.

These tests ensured that any computational errors were minimized and that the values calculated for larger values of n, which could not be necessarily be confirmed by reference to independent sources, were correct.

The empirical characterization of the distribution of λ_1 was based on the mean (μ) and the central moments (μ_i) of the distributions that were calculated in Python for each value of n. The skewness (γ_1) and kurtosis (γ_2) were calculated from these using:

$$\gamma_1(n) = \frac{\mu_3(n)}{\mu_2^3(n)} \text{ and } \gamma_2(n) = \frac{\mu_4(n)}{\mu_2^4(n)} \qquad (5)$$

for each value of n.

3. The distribution of λ_1

The distribution of λ_1 is unimodal [3]. Furthermore, since the number of partitions with at least 2 parts increases with n and the number of partitions with at least 3 parts increases with n^2 (4), but the number of partitions with exactly $n - m$ parts is independent of n, for $0.5n \leq m$ (3), it is clear that (i) $p(m, n) \neq p(n - m, n)$, for m > 1, and (ii) the distribution of λ_1 is asymmetric.

3.1 The theoretical distribution of λ_1

Erdős and Lehner [3] showed that as $n \to \infty$ λ_1 tends to an extreme value distribution with mode (M):

$$M \approx \frac{\sqrt{6n}}{\pi} \ln\left(\frac{\sqrt{6n}}{\pi}\right), \qquad (6)$$

at which:

$$P(M, n) \approx \frac{\pi\sqrt{2}}{24\sqrt{n^3}} \exp\left(\frac{\pi\sqrt{6n}}{3} - 1\right)$$

Szekeres [11] refined the estimate of M for sufficiently large n to:

$$M \approx \frac{\sqrt{6n}}{\pi} \ln\left(\frac{\sqrt{6n}}{\pi}\right) + \left(\frac{3}{\pi}\right)^2 \left(1 + \left[1 - \frac{1}{6}\ln\left(\frac{\sqrt{6n}}{\pi}\right)\right] \ln\left(\frac{\sqrt{6n}}{\pi}\right)\right) - \frac{1}{2}, \qquad (7)$$

the leading term of which is identical to that given in equation (6). The mean (μ) is [7, 8, 10]:

$$\mu \approx \frac{\sqrt{6n}}{\pi}\left(\ln\left(\frac{\sqrt{6n}}{\pi}\right)+\gamma\right) \quad , \qquad (8)$$

where γ = 0.5772… is Euler's constant, and comparison of equations (7) and (8) leads to the conclusion that $\mu > M$, consistent with the asymmetry of the distribution of λ_1. The variance of this distribution is:

$$\sigma^2 = 2n, \qquad (9)$$

[3, 10, 11]. Based on Richmond's [10] expressions, the third moment about the mean is:

$$\mu_3 \approx \frac{48\sqrt{6}}{\pi^3}\zeta(3)n^{3/2},$$

where $\zeta(s)$ is Riemann's zeta function and $\zeta(3) \approx 1.20205\ldots$, so the skewness is:

$$\gamma_1 \approx \frac{12\sqrt{6}}{\pi^3}\sqrt{2}\zeta(3). \qquad (10)$$

The fourth moment about the mean is:

$$\mu_4 \approx \frac{432n^2}{\pi^4}\left(\zeta(2)^2 + 6\zeta(4)\right) = \frac{204}{5}n^2,$$

where I have used $\zeta(2) = \pi^2/6$ and $\zeta(4) = \pi^4/90$, from which the theoretical estimate of the kurtosis is:

$$\gamma_2 = 204/20. \qquad (11)$$

3.2 The empirical properties of the distribution of λ_1

The probability distribution functions (PDFs) of λ_1 calculated using equation (2) for four different values of n are plotted in Figure 2. The PDFs are unimodal and asymmetric, although they appear more symmetrical when plotted on a logarithmic scale (Fig 2). The theoretical distribution derived by Erdős and Lehner [3] and the approximation given by Szekeres [12] and are also shown in Figure 2, from which it is apparent that even for n = 1000 neither fit is perfect. In general, the theoretical distributions tend to be skewed excessively to small m and the mode is slightly too low. However, the fit improves in both respects as n increases.

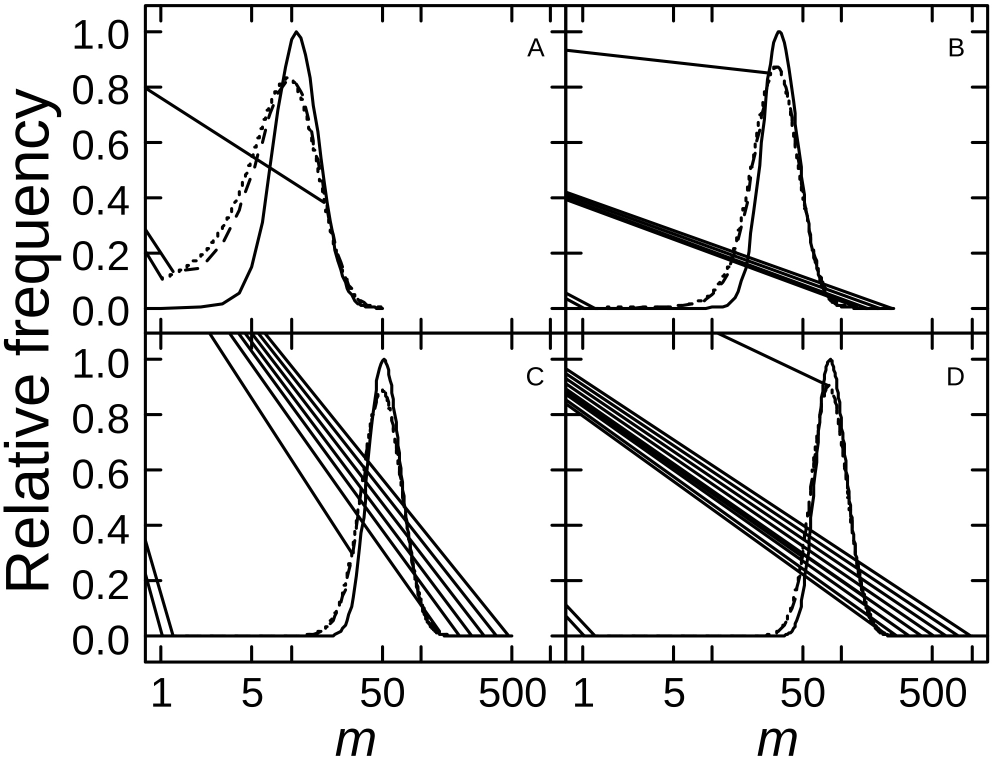

Figure 2. Distribution of λ_1 for n = 50 (A), n = 250 (B), n = 500 (C) and n = 1000 (D) (*In each case the solid line was calculated using equation (2), the dotted line is theoretical estimate of Erdős and Lehner [3] and the dashed line is the theoretical approximation of Szekeres [12] in which equation (6) is used rather than equation (7). For each n, the frequencies and the approximations are normalised to the actual modal frequency (that is P(M, n)). Note the logarithmic scale of the abscissa*)**.**

The theoretical analyses of the PDF of λ_1 indicate that μ, M and σ^2 increase with n (equations (7), (8) and (9)), whereas the skewness and kurtosis do not vary with n (equations (10) and (11)). Comparison of the empirical values with the theoretical predictions indicates that the mode is estimated well by equation (7) since the difference between the actual and the predicted values are less than 1 (Fig 3B), whereas the theoretical prediction of μ is consistently slightly lower (residual < 2) than the actual value (Fig 3, A and B).

The estimate of σ^2 (9) is considerably greater than the actual value, but the empirical expression given below (section 4) provides a significantly improved estimate (Fig 3, C and D). Both the skewness and the kurtosis appear to approach constant values (Fig 3, E and F), but they are larger than those predicted by equations (10) and (11). Moreover, it is clear that they vary considerably for very small n (say n < 200).

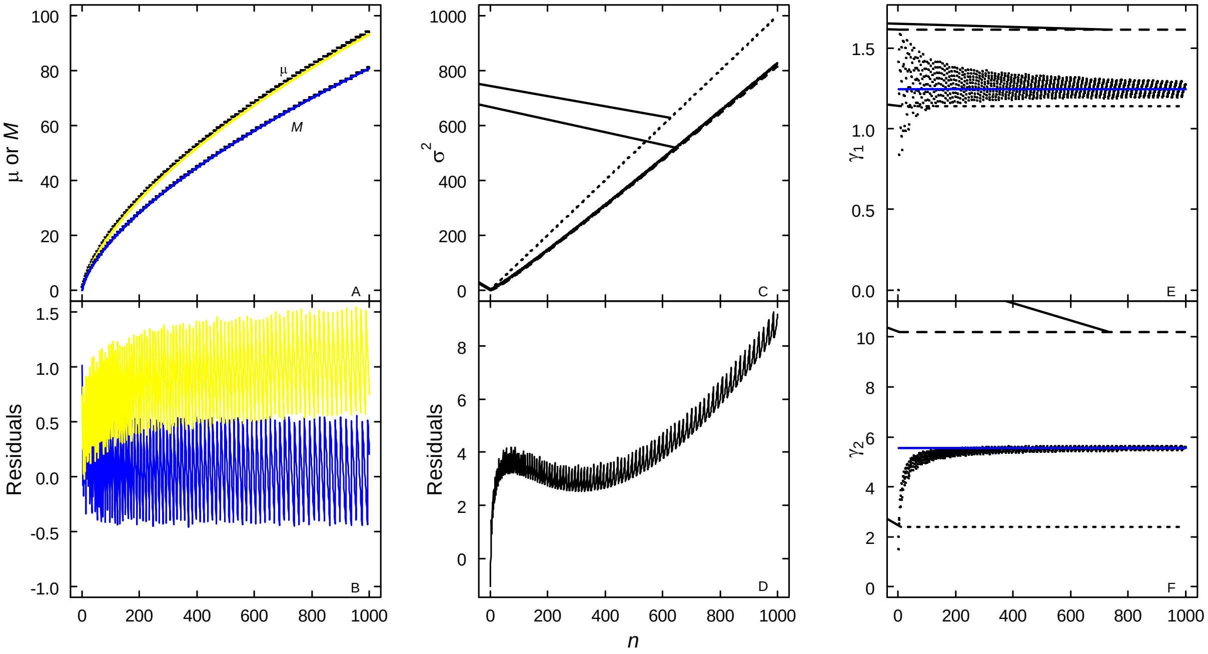

Figure 3. The empirical values and estimates of the mean and mode (A), variance (C), skewness (E) and kurtosis (F) of the distribution of λ_1 as a function of *n* (*The error in the theoretical estimates of the mean and mode (B) and of the variance (D) is also shown. (A) The points represent the actual mean (μ) and mode (M) of the calculated distributions and the solid lines are the corresponding approximations given by equations (8) and (7), respectively, and the residuals are shown in (B). In (C) the variance determined from the calculated distributions (solid line), the empirical approximation given by equation (13) (dashed line) and σ^2 = n (dotted line) are shown, and the difference between the actual variance and the empirical approximation is shown in (D). In (E) and (F), the solid line is the apparent limiting value of γ_1 and γ_2, respectively, and the dashed and dotted lines correspond to the theoretical estimates (equations (10) and (11)) and the values expected for the extreme value distribution (section 4)*).

4. Some semi-empirical improvements of the estimates for small *n*

Erdős and Lehner [3] showed that, as $n \to \infty$, λ_1 tends to an extreme value distribution, for which the PDF is:

$$P(x;a,b) = \frac{1}{b}\exp\left(\frac{a-x}{b} - \exp\left(\frac{a-x}{b}\right)\right).$$

At least two pieces of empirical evidence support this conclusion. First, for this distribution, $M = a$, $\mu = a + b\gamma$ and $\sigma^2 = \pi^2 b^2 / 6$, so a plot of μ – M against σ should have a gradient of $\gamma\sqrt{6}/\pi \approx 0.45$, which is almost the case (the value for $n \le 1000$ is 0.48, although this declines if the small values of *n* are omitted). Second, the expected skewness for the extreme value distribution is $\gamma_1 = 12\sqrt{6}\zeta(3)/\pi^3 \approx 1.139541$, which differs from the theoretical estimate (10) by a factor of $\sqrt{2}$. The expected kurtosis for the extreme value distribution is $\gamma_2 = \frac{12}{5} = 2.4$, which is approximately half of the observed limiting value (which appears to be about 5.5 (Fig 3F) and about 25% of the theoretical estimate (which is 10.2). Assuming that λ_1 has an extreme value distribution, then σ^2 can be estimated from μ and M using:

$$\sigma^2 = \frac{\pi^2}{6\gamma^2}(\mu - M)^2 .$$

While equation (7) provides a good estimate of M (Fig 3, A and B), the theoretical expression for μ (8) consistently underestimates the actual value (Fig 3, A and B), which can be partially rectified by adding 0.5 to the estimate, so it may be conjectured that:

$$\sigma^2 = \frac{\pi^2}{6\gamma^2}\left(\frac{\sqrt{6n}}{\pi}\gamma - \left(\frac{3}{\pi}\right)^2\left(1+\left[1-\frac{1}{6}\ln\left(\frac{\sqrt{6n}}{\pi}\right)\right]\ln\left(\frac{\sqrt{6n}}{\pi}\right)\right)+1\right)^2 . \tag{12}$$

The leading term of equation (12) is n, consistent with the theoretical estimates [3, 10, 11] and letting $c(n) = 1 + \left(1 - \ln\left(\sqrt{6n}/\pi\right)/6\right)\ln\left(\sqrt{6n}/\pi\right)$, which is almost constant for $200 < n < 1000$, equation (12) can be written as:

$$\sigma^2 = \left(\sqrt{n} - \frac{\left(9c(n) - \pi^2\right)}{\sqrt{6}\pi\gamma}\right)^2 ,$$

the coefficients of which are approximately constant. While this expression has some theoretical justification, it does not fit the data as well as the empirical expression of the same form:

$$\sigma^2 \cong \left(\frac{3}{\pi}\right)^2 n - \pi\sqrt{n} + \frac{\pi^2}{3} , \tag{13}$$

which is shown in Figure 3C. Equation (13) deviates from the actual value by less than 2.5% for $200 < n < 1000$ (Fig 3D).

It is clear from Figure 1 that neither the Erdős and Lehner [3] nor the Szekeres [12] approximation is a good description of the distribution of λ_1 for small values of n. In each case, the error (ε) varies systematically with m and declines with n (Fig 4, A and B), although at n = 1000, mean square error (MSE) is about 10^{-3} or greater (Fig 4D). Since the extreme value distribution appears to be an appropriate model even for small n, the improved semi-empirical estimates of σ^2 (equations (12) and (13)) and Szekeres' [11] estimate of M (7) can be used to generate a significant improvement in the estimate of the distribution for small n (Fig 4C). The semi-empirical approximation described here provides a significant reduction in ε (Fig 4C) leading to an MSE of less than 10^{-4} at n = 1000 (Fig 4D). Moreover, the MSE appears to vary relatively little for $150 \leq n \leq 1000$, compared with the theoretical estimates (Fig 4D). However, an empirical estimate of M:

$$M = \sqrt{\frac{2}{3}} n^{2/3} \qquad (14)$$

can be used in the place of Szekeres' [11] theoretical estimate (7). In this case, the MSE is smaller for $n > 600$, reaching almost 10^{-5} for $n = 1000$ (Fig 4D).

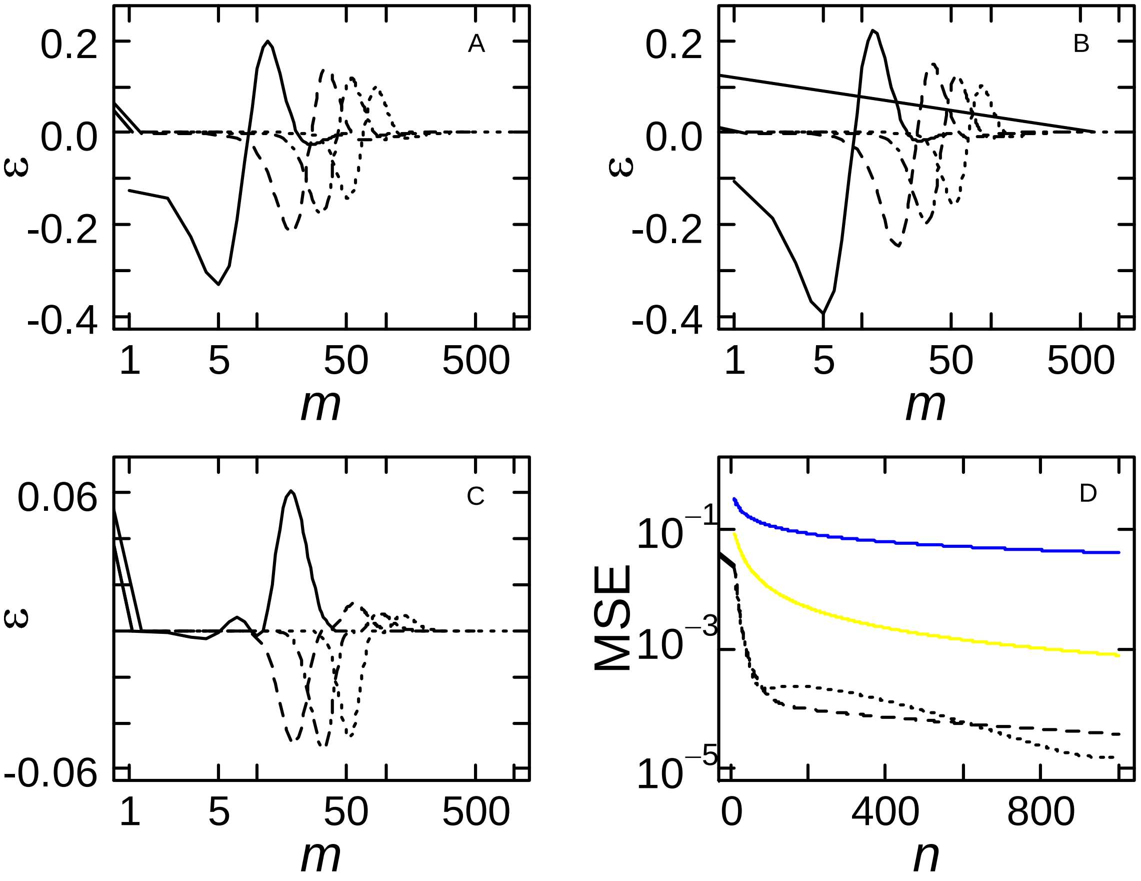

Figure 4. The error (*ε*) of the Erdős and Lehner [3] (A), Szekeres [12] (B) and empirical (C) estimates of the distribution of λ_1 (*The estimates for n = 50 (————), 250 (– – – –), 500 (– · – · –) and 1000 (·········), are plotted as a function of m (A-C), and the mean square error (MSE) for these approximations is plotted as a function of n (D). In (D) the solid blue and yellow lines correspond to the MSE for the Erdős and Lehner [3] and Szekeres [12] estimates, respectively; the dashed line is that of the empirical estimate based on equations (7) and (13); and the dotted line is that of the empirical estimate based on equations (13) and (14). Note that the abscissae of (A), (B) and (C) and the ordinate of (D) are on a logarithmic scale*)**.**

5. Conclusions

The empirical distribution of λ_1 has been analysed and shown to be consistent with the theoretical prediction that it has the extreme value distribution. However, the theoretical estimates of the parameters of the extreme value distribution of λ_1 are less reliable for small *n* and become less reliable as *n* decreases (< 1000). Semi-empirical approximations for the mode and variance of the distribution are derived that decrease the MSE at least 10-fold for small *n*.

References

[1] Andrews, G. E. (1976). *The Theory of Partitions*. Addison-Wesley Publishing Company, Reading.

[2] Auluck, F. C., Kothari, D. S. and Luthra, S. M. (1957). The degradation of high polymers and the partition theory of numbers. *Current Science* 26, 173-175.

[3] Erdős, P. and Lehner, J. (1941). The distribution of the number of summands in the partitions of a positive number. *Duke Mathematics Journal* 8, 335-345.

[4] Ewens, W. J. and Kirby, K. (1975). The eigenvalues of the neutral alleles process. *Theoretical Population Biology* 7, 212-220.

[5] Gupta, H. (1980). *Selected Topics in Number Theory*. Abacus Press, Tunbridge Wells.

[6] Gupta, H., Gwyther, C. E. and Miller, J. C. P. (1962). *Tables of partitions*. Cambridge University Press, London.

[7] Kessler, I. and Livingston, M. (1976). The expected number of parts in a partition of *n*. *Monatshefte für Mathematik* 81, 203-212.

[8] Luthra, S. M. (1957). On the average number of summands in partition of *n*. *Proceedings of the National Institute of Science of India* 23A, 483-498.

[9] Nelson, D. R. and Strobel, H. W. (1988). On the membrane topology of vertebrate cytochrome P-450 proteins. *Journal of Biological Chemistry* 263, 6038-6050.

[10] Richmond, L. B. (1975). The moments of partitions, I. *Acta Arithmetica* 26, 411-425.

[11] Szekeres, G. (1953). Some asymptotic formulae in the theory of partitions (II). *Quarterly Journal of Mathematics* 4, 96-111.

[12] Szekeres, G. (1987). Asymptotic distribution of the number and size of parts in unequal partitions. *Bulletin of the Australian Mathematical Society* 36, 89-97.

[13] Wieder, G. M. and Marcus, R. A. (1962). Dissociation and isomerization of vibrationally excited species. II. Unimolecular reaction rate theory and its application. *Journal of Chemical Physics* 37, 1835-1852.

[14] Wolfowitz, J. (1942). Additive partition functions and a class of statistical hypotheses. *Annals of Mathematical Statistics* 13, 247-279.