Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Vol. 11, Issue 1 (June 2016), pp. 115 - 126

Applications and Applied Mathematics:
An International Journal
(AAM)

Survival Analysis of the Men's 100 Meter Dash Record

Farzad Noubary[1] and Reza Noubary[2]

[1]Biostatistics, Epidemiology, and Research Design (BERD) Center
Tufts Medical Center
35 Kneeland Street, 9th Floor
Boston, MA 02111 USA

[2]Department of Mathematics/Statistics
Bloomsburg University
Bloomsburg, Pennsylvania USA
rnoubary@bloomu.edu

Received: August 15, 2014; Accepted: December 2, 2015

ABSTRACT

In the 2012 Summer Olympics in London seven out of eight finalists in the men's 100 meter dash crossed the finish line in under 10 seconds. This result and other recent performances of exceptional sprinters such as Bolt have made experts wonder, not whether a new record will be set, but when and how much it will lower the present record. Seeking an answer, some researchers have tried to model the available data with the goal of using them to predict future records. This article presents a different approach based on theory of records for independent and identically distributed observations. It modifies the number of attempts to break a record to make the results of the theory of records applicable to this situation. The modification is necessary because many sports records have been broken more frequently than what this theory predicts. Two modifications of the number of attempts are considered, fixed rate via a geometric increase, and random rate via a non-homogeneous Poisson process.

Keywords: Survival; 100 meter Dash; Records; Non-homogeneous Poisson Process

MSC 2010 No: 46N30

1. Introduction

The athletic ability of human beings is an issue of great interest to physiologists, physical educators, health professionals, sport fans, and the general public. Records set in different sports shed light on human strengths and limitations and provide data for scientific investigations and training or treatment programs. Research in this area can be divided into two categories: short-term prediction and long-term (ultimate record) prediction. See Krzysztof and Mero (2013), Stephenson and Tawn (2013), Barrow (2012), Solow and Smith (2005), Noubary (2005), Gulati and Padgett (2003), Bennett (1998), Blest (1996), and references therein. In what follows we present some of the statistical approaches employed for prediction of records and add some new insights.

As in most other applications, a few different approaches have been employed by investigators to predict records. One general approach has been to utilize models that are made up of a deterministic term to account for the trend and a stochastic term to account for the variation. See, for example, Blest (1996) for details and a list of models considered for both deterministic and stochastic components. It has been noted that in general, application of this approach cannot produce meaningful performance estimates for the distant future. This is due to the fact that in most sports, in addition to records themselves, the amount by which they have been improved has also decreased with time. To address this, Noubary (1994) considered innovative models comprised of an envelope functions and a stationary stochastic processes in multiplicative forms.

Next, noting that records are extreme values, a popular approach has been to apply the extreme value theory and fit one of the three limiting distributions to, for example, best record of each year. See for example Krzysztof and Mero (2013) and Stephenson and Tawn (2013) for application of this approach and DeHann and Ferreira (2006), Cole (2001), Ahsanullah and Kirmani (2008), and Beirlant et al (2004) for recent developments of extreme value theory. Of three extreme value distributions, the type I (Gumbel distribution) often provided a better fit despite the fact that it has no lower bound. On the other hand, the type III distribution with a lower bound has often led to unreliable results and in some cases estimates of a lower bound greater than those that have already occurred. This is not the only problem with application of extreme value theory. Other problems include:

1. The estimating procedures are complicated.

2. Their application requires a moderate or large sample.

3. Since only the best record of each period (e.g., a year) is used, information contained, for example, in the second best record of that period is not utilized. This point is particularly important when the time span of the available data is short.

In a more innovative approach the probabilities of future performances are calculated using models for the lower tail of the distribution for performance measures. Here, one assumes that the tail belongs to a given parametric family and carries out the inference using the times below some predetermined value y_o. In most applications these differences are treated as independent random variables. Like most asymptotic results, application of this approach is not free of

problems. The obvious problems are the choice of a parametric family, determination of the threshold value y_o and problems related to the intractable likelihood equations. Also, as in application of extreme value theory, estimate of ultimate record are often unacceptable.

Considering these difficulties, this article attempts to apply an approach based on the theory of outstanding values or records. Despite its appeal this theory has not been applied to most sports.

2. Theory of Records

The theory of records deals with values that are strictly greater than or less than all previous values. Usually the first value is counted as a record. Then a value is a record (lower record or record low) if it is less than all the previous values. The study of record values, their frequencies, times of their occurrences, their distances from each other, etc. constitutes the theory of records. See Ahsanullah (1995), Arnold et al (1998), Glick (1978), and Gulati and Padgett (2003) for details. Formally, the theory deals with four main random variables:

(1) The number of records in a sequence of *n* observations.

(2) The record times (times records occur).

(3) The waiting time between the records.

(4) The record values.

The theory of records has not been widely used to address questions regarding sports records. This is because:

(1) the theory is developed for independent and identically distributed sequences and, as such, is not directly applicable to sports records, and

(2) in most sports, records have occurred more frequently than what the theory predicts.

For example, consider the men's 100 meter data for the period 1912 – 2012. The data includes 20 records of which half of them (10) were set since 1990. As is shown below (result (b)), to produce 20 records, more than one-hundred million independent and identically distributed official attempts are needed. Also, unlike the theoretical expectation (result (d) below), the waiting times between the records have decreased significantly with time especially in the recent past. See Figure 1 for the history of 100 meter.

To account for these and other contributing factors such as diet, shoes, and track type, etc., we treat the problem as an independent and identically distributed one, but make up for the increase in probability and frequency of records by inflating the number of attempts**.**

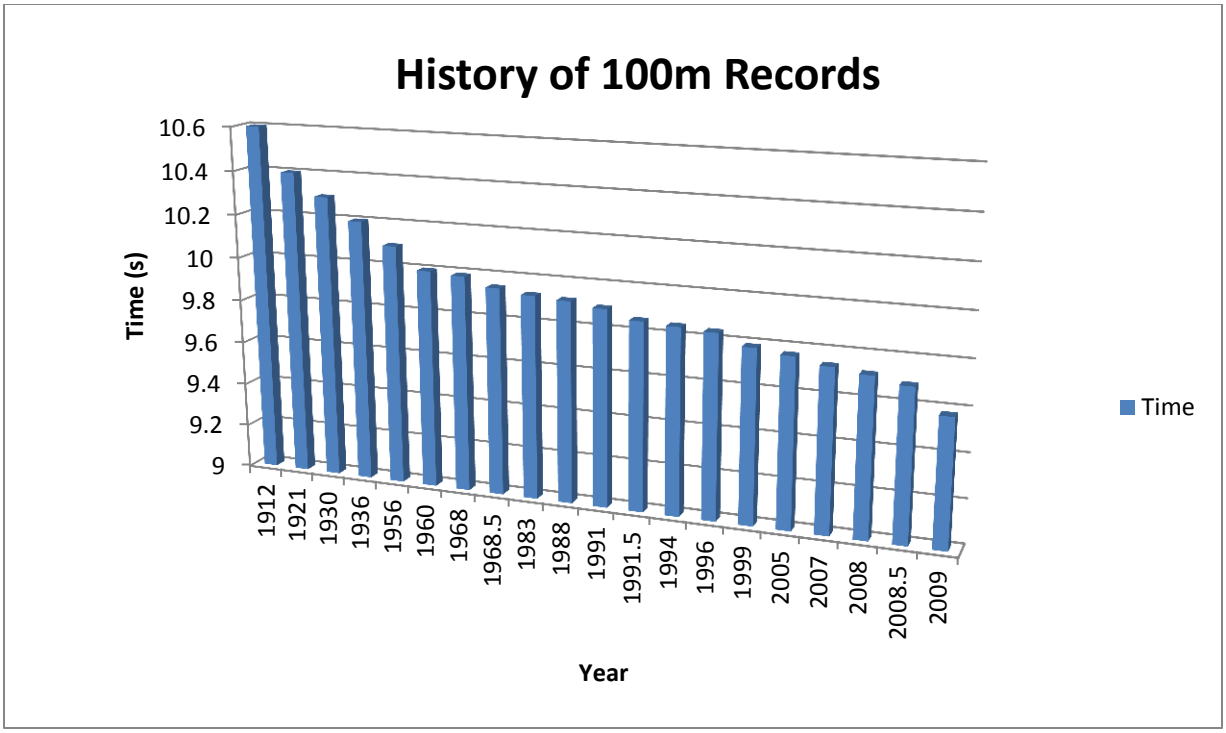

Figure 1. History of 100 Meter Dash Progression

We start by stating results of the theory of records for independent and identically distributed observations that we plan to use:

(a) If there is an initial sequence of n_1 observations and a batch of n_2 future observations, then the probability that the additional batch contains a new record is $n_2/(n_1+n_2)$.

(b) For large n, $P_{r,n}$, the probability that a series of length n contains exactly r records, is given by

$$P_{r,n} \sim \frac{1}{(r-1)!n}[\ln(n)+\gamma]^{r-1} \qquad (1)$$

where $\gamma = 0.5772$ is Euler's constant (see appendix).

(c) As sample size $n \to \infty$, the frequency of the records among observations indexed by $a_n < i < b_n$ tends to a Poisson count with mean $\ln(b/a)$.

(d) The median of W_r, the waiting time between the $(r\text{-}1)^{th}$ and r^{th} records, are:

Table 1. Medians of Waiting Times Between Successive Records and Their Ratios

Record Number r	2	3	4	5	6	7	8
Median (W_r)	4	10	26	69	183	490	1316
$\text{Med}(W_r)/\text{Med}(W_{r-1})$		2.50	2.60	2.65	2.65	2.68	2.69

Moreover,

$$\frac{\text{Median}(W_{r+1})}{\text{Median}(W_r)} \approx e = 2.718...$$

and

$$\ln(W_r)/r \to 1.$$

Also, $\ln(W_r)$ is approximately equivalent to the arrival time sequence of a Poisson process.

Two points should be noted. First, we worked with the median since the expected value of W_r is infinite even for $r = 2$. Second, it is possible to model $\ln(W_r)$ as a non-homogeneous Poisson process noting that sports records may have different patterns and are more frequent than records for independent and identically distributed sequences.

To illustrate the suggested approach, let us apply the above results to 100 meter dash data. We start with result (b). Using the maximum likelihood method and maximizing $P_{r,n}$ with respect to n we find n = 100,212,150. This is an estimate for the number of independent and identically distributed attempts that is required to produce 20 records. Next, we need to distribute these attempts over the period of 100 years. According to result (d), the median number of attempts required to arrive at a new record is $e = 2.718\ldots$ times the median number of attempts that was required to arrive at the present record. This suggests a geometric increase with rate e. If we assume that one unit of attempt was needed to arrive at the second record, the total number of attempts to arrive at record number 20 may be calculated as

$$1 + e + e^2 + \ldots + e^{18} = 103{,}872{,}541.$$

Note this is a slight overestimation since for early records the ratios are less than e.

Now rather than this, we can distribute the attempts over the years it has taken to arrive at record number 20. Suppose for example that i is the annual geometric rate of increase in number of attempts. This means we are assuming that the number of attempts in any given year has been i times the number of attempts during the year before it. With this assumption the value of i can be found by solving the equation

$$1 + i + i^2 + \cdots + i^{100} = 100{,}212{,}150,$$

where i^j represents the number of attempts in year j. Here we get i = 1.179888, which means there has been 17.9888% more attempts per year. To predict records for the future one and five years (in this case the year 2013 and the period 2013 – 2018) we replace n_1 = 100,212,150 and

n_2 = $(1.179888)^{101}$ = 18026823 for the next year, and n_1 = 100,212,150 and n_2 = $(1.179888)^{101}$ + ... + $(1.179888)^{105}$ = 128939191 for the next 5 years in result (a) above. This leads to probability estimates of 0.152461 and 0.562681 for a new record in the next one and five years, respectively. The same probability for the next 10 years is 0.8087529.
It is interesting to note that Berry (2002) has used the increase in the male population of the world as an adjusting factor for participation. We note his model for the growth of the world's male population, namely,

$$\text{Population in Year } t = 1.6\ exp\ [0.0088(t - 1900)] \qquad (2)$$

implies a geometric increase with annual rate of *exp* (0.0088) = 1.0088388. It should be mentioned that for 100 meter dash data this rate is clearly unrealistic.

Next, using result (c), the frequency of the records among observations 100, 212,150 and 100,212,150 + 128939191 = 229151341 has approximately a Poisson distribution with mean $\lambda = \ln(229151341/100212150)$ = 0.8270932. Using this, the probabilities of none and one record in the period 2013- 2018 are 0.4373 and 0.3617 respectively. Note that 0.4373 is also the 5 years survival probability of the present record.

Finally, looking at the data we see a clear change in the rate of records and their waiting times starting the year1990. In fact, after 1990:

1. 10 records were set compared to 10 records in the previous 78 years (1912-1990).

2. Record improvements are more notable, from 9.92 to 9.58 compared to the previous 78 years, from 10.6 to 9.92. Also, the waiting times between records were significantly shorter.

Considering this, we decided to use only the data since 1990. This makes sense as this data carries more relevant information about the recent records and the future ones. Using the data for this period and following the same procedure we found i = 1.3615, n_1 = 4550 and n_2 = 1646 for the next year, and n_1 = 4550 and n_2 = $(1.3615)^{24}$ + $(1.3615)^{25}$ + $(1.3615)^{26}$ + $(1.3615)^{27}$ + $(1.3615)^{28}$ = 16748.23 for the next 5 years. This led to probability estimates of 0.2657 and 0.78673 for a new record in the next one and five years, respectively. We think these are more realistic considering the outcomes of the recent competitions of this event.

We end this section by noting that one would expect even better results if the geometric increase could be replaced by increase in population of participants or number of attempts. For this we could, for example, consider models such as Logistic or Gompertz or more generally a model of the form

$$y_{n+1} - y_n = H(y_n) = i*f(y_n)\,(1- g(y_n)),$$

where y_n denotes the number of participants or number of attempts at year n (generation n). One of the simplest and frequently used models that contains a formulation which avoids indefinite

growth and represent effects of overcrowding is when i is a linear function of the previous year's participation. This choice of i leads to a model of type

$$y_{n+1} - y_n = i^* y_n (1 - y_n / h) = H(y_n),$$

known as Logistic equation. Here, i^* represents the rate of growth and h represents the carrying capacity. For the 100 meter dash, h may be the maximum number of individuals who qualify to participate in an event such as the Olympics. This type of model is reasonable for sports where usually rapid initial improvements are followed by much slower advances.

Noubary (2005) considered the following simpler model instead that exhibits the same behavior as the Logistic equation

$$y_{n+1} = y_n exp[r^* (1 - y_n / h)].$$

For example, the number of attempts in the future 1 and 10 years are 412 and 4876 if we use y_0 = 100, r^* = 0.04 and h = 50. The corresponding numbers using the logistic equation are 402 and 4742. These numbers result in smaller probability estimates compared to the geometric increase of 4%.

3. Attempts as Non-homogeneous Poisson Process

In the previous sections we assumed an increasing but fixed number of attempts per year. It is also possible to assume that the number of attempts to break a record is random and is governed by a non-homogeneous Poisson process. The following is a brief description of this approach.

Let $R > 0$ and $S > 0$ be two random variables with respective distribution functions $F_R(.)$ and $F_S(.)$. Suppose R, the record in a given sport, is subject to a set of events (attempts) S occurring according to a point process $\boldsymbol{P}$. Then the record breaks if the value of S exceeds (succeeds) R. The value of S is a function of the type of sport, number of participants, prize, training, environmental factors such as temperature, altitude, etc., and factors important to the athletes and the public. The value of R depends on factors such as the type and popularity of the sport, amount of rewards or prizes, number of formal competitions, etc. The probability p of breaking a record in a single attempt is then,

$$p = P(S > R) = 1 - \int_0^\infty F_S(x) dF_R(x).$$

When applying this model, one is frequently interested in the probability of breaking a record in a specified interval, say (0, t], where 0 represents the beginning of the period. Assuming that T represents the survival time, the probability of the record being broken in the time interval (0, t], denoted by $F_T(t)$, can be obtained as

$$F_T(t) = P(T \le t) = 1 - P(T > t) = 1 - L_T(t),$$

where $L_T(t) = P(T > t)$, $L_T(0) = 1$ is the survival function. If R is a record subject to a sequence of attempts $S_1, S_2, \ldots, S_n$, then $L_T(t)$ is given by

$$L_T(t) = \sum_{r=0}^{\infty} P(N(t) = r)\bar{P}(r) \tag{3}$$

where $\{N(t), t \geq 0\}$ is a general counting process of attempts occurring randomly in time and $\bar{P}(r) = P(\max(S_1, S_2, \ldots, S_r) < R)$, $r = 1, 2, \ldots, n$, with $\bar{P}(0) = 1$. Note that here $\bar{P}(r)$ represents the probability of surviving the first r attempts. When attempt is governed by a homogenous Poisson process with rate λ, then from Equation (3) we obtain

$$L_T(t) = \sum_{r=0}^{\infty} [\, e^{-\lambda t}(\lambda t)^r / r! \,]\, \bar{P}(r).$$

If a further assumption is made that the attempts are independent and identically distributed random variables, then this reduces to

$$L_T(t) = \sum e^{-\lambda t}(\lambda t)^r / r!)(1-p)^r = \exp(-\lambda t p).$$

Thus, if the mean rate of attempts and the period of interest are given, then $L_T(t)$ can be calculated for any p. Hence, the main problem for the situation described above is that of estimating the p, i.e. the probability of breaking a record in a single attempt.

Suppose now that $\boldsymbol{P}$ is Poisson with time-dependent rate $\lambda(t) > 0$ and let

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

If $F_R(.)$ denotes the distribution function of R and $F_S(.)$ the distribution function of $\{S_n, n = 1,2,\ldots\}$, then since $(T > t)$ if and only if $\max(S_1, S_2, \ldots, S_n) < R$, the required probability $P(T > t)$ is given by

$$P(T > t) = \int_0^{\infty} \sum_{n=0}^{\infty} e^{-\Lambda(t)} \frac{(\Lambda(t))^n}{n!} [F_S(x)]^n dF_R(x). \tag{4}$$

Note that $[F_S(.)]^n$ is the distribution function of $max(S_1, S_2, \ldots, S_n)$. Expression (4) can also be written as

$$P(T > t) = \int_0^{\infty} \exp[-\Lambda(t)(1 - F_S(x)]dF_R(x) \tag{5}$$

If $R = R_0$ is given (e.g. R_0 is the present record), then

$$P(T > t \mid R = R_0) = \exp[-\Lambda(t)(1 - F_S(R_0))]$$

In general, application of Equation (5) requires knowledge of both $F_R(.)$ and $F_S(.)$ which may not be available. However, there is an important case discussed below where calculations can be carried out with less information and without any numerical integration. Taking the viewpoint that the strength or importance of a record in a given sport may be measured by the number of attempts it requires to break it, this case and the assumptions made seem reasonable.

Suppose that ***P*** has been observed throughout the time interval $(-\tau, 0]$, where 0 represents the present time. Suppose also that the largest value in this interval is the present record and is used as a reference for determining further records. Then,

$$F_R(x) = P(R < x) = \sum_{n=0}^{\infty} P[\max(S_1, S_2, ..., S_n) < x \mid N(\tau) = n)]P(N(\tau) = n)$$

$$= \sum_{n=0}^{\infty} e^{-\Lambda(\tau)} \frac{(\Lambda(\tau))^n}{n!} (F_S(x))^n = \exp[-\Lambda(\tau)(1 - F_S(x))]$$

and application of Equation (3) yields

$$P(T > t) = \Lambda(\tau)[1 - \exp(-(\Lambda(\tau) + \Lambda(t)))]/[\Lambda(\tau) + \Lambda(t)]. \quad (6)$$

With confidence given by the right-hand side of Equation (6), there will be no new maximum in (0,t] greater than the one in $(-\tau,0]$. Thus, in this case the survival probability depends only on the rate of attempts.

Note that for large values of $\Lambda(\tau) + \Lambda(t)$, we have approximately

$$P(T > t) = \Lambda(\tau)/[\Lambda(\tau) + \Lambda(t)]. \quad (7)$$

This results in answers similar to those discussed in earlier sections. For example, for a geometric increase in number of attempts the survival probability of the 100 meter record in the next 5 and 10 years are 0.4373 and 0.1912 respectively. Investigation of other forms of increase in number of attempts is the goal of our future research.

We end this section by making a point regarding the limit of human abilities as it relates to the idea of a possible ultimate record. In terms of what is discussed here, the ultimate record is the one that will survive forever, i.e. its survival probability is 1. Since it is generally believed that every record will eventually be broken it is more practical to think of a survival time such as 50 years and a survival probability such as 90% for a record to be considered an ultimate record. To demonstrate suppose that $\Lambda(\tau) + \Lambda(t)$ is large. Then, we could use the following

$$P(T > 50) = \Lambda(\tau)/[\Lambda(\tau) + \Lambda(50)] > 90\%$$

and estimate τ for a given $\lambda(t) > 0$.

4. Conclusion

The100m dash defines the fastest man on Earth. In well over 100 years there have been only 25 men who have attained this title. This makes one wonder whether records will be lowered further and indefinitely or there is an ultimate record. Clearly, it is more reasonable to assume that there is a limit, and if so whether it is possible to estimate it. This article presented a statistical method as a step towards answering the question regarding the possible limit. Since sport records are not generated by an independent and identically distributed random variables, the approach suggested here makes a change to the data and applies the results of the theory of records for independent and identically distributed data. The result shows that such approach can produce acceptable answer.

REFERENCES

Ahsanullah, M. (1995). Record Statistics, Commack: Nova Science Publishers, Inc.

Ahsanullah, M. and Kirmani, S.N.U.A. (2008). Topics in Extreme Values: Nora Science Publishers. Inc.

Andel, J. (2001). Mathematics of Chance. Weily, New York.

Arnold, C.A., Balakrishnan, N, and Nagaraja, H.N. (1998). Records, New York: Wiley.

Barrow, J.D. (2012). How Usain Bolt can run faster-effortlessly. *Significant.* 9.2. 9-12.

Beirlant, J, Geogebeur, Y., Segers, J., and Teugels, J. (2004). Statistics of Extremes: Theory and Applications: Wiley.

Bennett, J. (1998) *Statistics in Sports*, New York, Arnold Publication.

Berry, S.M. (2002). A statistician reads the sports pages, *Chance.* 15, 49-53.

Blest, D. C. (1996). Focus on sport lower bounds for athletic performance. *The Statistician*, 45, 2, 243-253.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values: Springer.

Davis, R., and Resnick, S. (1984). Tail estimates motivated by extreme-value theory, *Ann. Statist.* 12, 1467-1487.

De Hann, L. and Ferreira, A. (2006). Extreme Value Theory: An Introduction: Springer Series in Operations Research.

Glick, N. (1978). Breaking records and breaking boards, *Amer. Math Monthly*, 85, 2-26.

Gulati, S. and Padgett, W.J. (2003). Parametric and Nonparametric Inference for Record-breaking Data, London: Springer-Verlag.

Krzysztof, M. and Mero, A. (2013). A kinematics analysis of three best 100 m performances ever. Journal of Human Kinetics. 36, 149-160.

Noubary, R. (2005). A procedure for prediction of sports records, *J. of Quantitative Analysis in Sports*, 1, 1-12.

Noubary, R. (1994). An Envelope Function Model for Forecasting Athletic Records. *J. of Forecasting.* 13, 11-20.

Solow, A.R. and Smith, W. (2005). How surprising is a new record. *The Amer. Statistician*, 59, 153-155.

Stephenson, A.G. and Tawn, J.A. (2013). Determining the Best Track Performances of All Time Using a Conceptual Model for Athletics Records. *Journal of Quantitative Analysis in Sports*, 9, 1, 67- 76.

Appendix: Maximum Likelihood Estimate of Number of Attempts

Let $P_{r,n}$ denote the probability that in a series of length *n* there will be exactly *r* records. Then,

$$P_{n,n} = \frac{1}{n!} \quad , \qquad P_{1,n} = \frac{1}{n}$$

$$P_{r,n} = \frac{n-1}{n} P_{r,n-1} + \frac{1}{n} P_{r-1,n-1},$$

$$P_{1,1} = 1,$$

$$P_{r,0} = 0, \qquad r \le n.$$

Also, successive substitution for *n*, results in

$$P_{r,n} = \frac{1}{n} \sum_{j=r-1}^{n-1} P_{r-1,j} \ .$$

For example,

$$P_{2,n} = \frac{1}{n} \sum_{j=1}^{n-1} P_{1,j} = \frac{1}{n} \sum_{j=1}^{n-1} 1/\ j \quad \approx \quad \frac{1}{n} (\ln(n-1) + \gamma).$$

Using the properties of the Stirling numbers it can be shown that (Andel 2001) as $n \to \infty$

$$P_{r,n} \sim \frac{1}{(r-1)!\, n} (ln(n) + \gamma)^{r-1} \tag{8}$$

Table 2 provides maximum likelihood estimate for *n* for *r* = 1, 2…, 12 together with the maximum value of the $P_{r,n}$ itself. As can be seen the value of *n* increases rapidly. For *r*-values greater than 12 one could apply the following observation. Noting that 2/1=2, 8/2=4, 25/8=3.125, 73/25=2.92, 204/73=2.795, 565/204=2.77, 1557/565=2.756, 4275/1557=2.746, 11710/4275=2.739, 32022/11710=2.735, 87464/32022=2.731, we conjecture that the ratio is tending to *e = 2.718.* This is also evident from the approximate expressions (8), for large *n*.

When n is large its maximizing value is $exp(r\text{-}1\text{-}\gamma)$. Thus, for example, a good approximation for n when r =13 *is* (87464)(2.718)= 237727. In addition to maximum likelihood values of n using exact expression for $P_{r,n}$ (column 2), Table 2 presents maximum likelihood values of n denoted $\boldsymbol{n_1}$, using the approximate expression (8) for large n.

Table 2. Maximum Likelihood Values of n

r	***n***	**Max. Prob.**
1	1	1.0
2	2	0.5
3	5	0.325694
4	12	0.253788
5	31	0.214182
6	84	0.188597
7	227	0.170410
8	616	0.156648
9	1674	0.145767
10	4550	0.136886
11	12368	0.129456
12	33618	0.123122