Estimation of Percentile Score Based on Marks Obtained in an Examination: A Simple Mathematical Model

Sudipto Roy
Email: roy.sudipto@sxccal.edu

Assistant Professor, Department of Physics, St. Xavier's College, 30 Mother Teresa Sarani (Park Street), Kolkata-700016, West Bengal, India

Abstract
In order to find whether a student is eligible for admission to higher academic institutions, it is often necessary to estimate the percentile score of the student, based on the marks obtained in a competitive examination. In the present article, we have discussed a very simple mathematical model to calculate the percentile score based on marks. For this purpose, we have defined a function representing the probability that a certain fraction of the syllabus has been studied by a candidate before appearing for the examination. Another function has been derived, in terms of that fraction, representing the probable percentage of marks obtained by the candidate. Using these functions, we have derived expressions for the expected percentile score and the rank of the candidate in terms of the percentage of marks. To determine the values of the constant parameters involved in the present model, one is supposed to use the marks-versus-percentile-versus-rank data obtained from the results of previous years. Due to the unavailability of these data, we have used different combinations of values of the parameters to show graphically how the percentile score and rank of a candidate vary as functions of the percentage of marks obtained in the examination.

Keywords: Negative marking, Marks vs Percentile, Marks vs Rank, JEE (Main), CUET, NTA

1. Introduction

The percentile score of a candidate in an examination is actually a measure of where the candidate stands, in terms of merit, compared to all examinees. There may be more than one definition of the percentile score [1]. According to a document regarding the Common University Entrance Test (CUET), conducted by the National Testing Agency (NTA) [2], the percentile score of a student appearing for the test is the percentage of candidates who have secured marks equal to or less than the marks obtained by the student [3]. The raw score of each candidate in each subject is normalised by the NTA by using the equipercentile method [4]. Based on their scheme, the percentile score of a candidate is first calculated using the marks of all candidates appearing for a subject on the same shift. Using the equipercentile method, the percentile scores of the candidates are converted into normalised marks taking into consideration the levels of difficulty of multiple sessions [5]. The same method is used for the examinations conducted by the Institute of Banking Personnel Selection (IBPS) [6]. The percentile score is also calculated for the examinees of the Joint Entrance Examination (JEE) Main, conducted by the NTA. JEE (Main) is conducted on multiple days and sessions and there can be different levels of difficulty of the question papers of these examinations. To maintain uniformity in the assessment process, a method is used by the NTA to obtain the normalized percentile scores for each subject and also for the aggregate. The rank (All India Rank) is assigned to a student on the basis of his or her percentile score obtained in the examination [7].

In each session of the UGC NET exam, conducted by the NTA, there are more than one set of question paper. The abbreviation, UGC, stands for the University Grants Commission (of India) and NET stands or the National Eligibility Test. Despite the best efforts, made by the NTA, to maintain uniformity, there can be different degrees of difficulty of these question papers, affecting the fairness of evaluation. In order to establish fairness in the examination, normalisation of marks in the UGC NET is carried out on the basis of the percentile scores [8]. For this purpose, the marks of the examinees, appearing for the same examination in different sessions, are compared.

The examinees generally estimate their marks for the JEE (Main) from the Final Answer Key released by the NTA. Before the percentile scores and ranks are released by the NTA, they search for the expected percentile scores and ranks, through internet, in the tables containing the marks-*versus*-percentile-*versus*-rank data prepared by different agencies (which provide guidance for such examinations) on the basis of information obtained from previous year's results.

Through this article, we present a simple mathematical model regarding the estimation of the expected percentile score and rank of a student on the basis of marks obtained in an examination. To formulate a model of this kind, one needs quantitative information regarding the level of preparation of the students before they appear for the examination. We have considered the syllabus to be divided into a certain number of parts, which are equal in terms of length and difficulty. The level of preparation of a particular student for the examination may be quantitatively judged or specified, in probably the simplest way, by counting the number of parts studied completely by the student. The probability that a student has studied a certain number of parts is expected to decrease as the number increases, as per common observations in the society. Based on this idea, we have defined a probability function which has been used to derive an expression for the percentile score a candidate, by calculating simply the percentage of candidates who are equally or less prepared (for the examination) compared to the candidate under consideration. The aspect of negative marking, which is a part of the marking scheme for many competitive examinations, has been taken into account to calculate the marks obtained by an examinee. We have derived mathematical expressions for the percentile score and rank in terms of the percentage of marks obtained by a candidate. For different sets of values of the unknown parameters connected to this model, we have shown graphically the dependence of the percentile score and rank upon the percentage of marks.

2. Model Formulation

Let us consider the syllabus of an examination to be divided into N equal parts, where the *equality* of any two parts is in terms of their states based on a combination of factors like *importance*, *length*, *difficulty* etc. The probability that a student has completely studied a certain number of these parts, say x, may be defined in the following way.

$$F(x) = \frac{a}{b^x} \ (x = 0,1,2,\dots,N) \qquad (1)$$

Here x is a discrete random variable. To satisfy the requirement that, $0 < F(x) < 1$, for all values of x, the parameters a and b must satisfy the conditions: $0 < a < 1$ and $b > 1$.

The definition of $F(x)$, as represented by equation (1), is based on a common observation that, the larger the value of x, the smaller would be the likelihood that this number of parts have been studied by a candidate before he or she appears for the examination.

Since the sum of probabilities for all possible values of the random variable x is unity, the probability function $F(x)$ of equation (1) must satisfy the following condition [9, 10, 11].

$$\sum_{x=0}^{N} F(x) = 1 \tag{2}$$

Substituting equation (1) into equation (2) we get,

$$a\left(1 + \frac{1}{b} + \frac{1}{b^2} + \frac{1}{b^3} + \cdots + \frac{1}{b^N}\right) = 1 \tag{3}$$

On the left-hand side of equation (3), we have a geometric series (with $N + 1$ number of terms) multiplied by the parameter a.
For the geometric series, $c + ct + ct^2 + \cdots + ct^{n-1}$, where the 1st term is c, the common ratio is t and the number of terms is n, it can be shown that the sum of all terms is $\frac{c(t^n-1)}{t-1}$ [12]. Using this formula to calculate the sum of the series in equation (3), we get the following relation between the parameters a and b.

$$a = \frac{b^{N+1}-b^N}{b^{N+1}-1} \tag{4}$$

It is evident from equation (4) that, for an extremely large value of N (for which we can neglect 1 in the denominator), we can write, $a \approx 1 - \frac{1}{b}$.
Substituting equation (4) into equation (1), we get,

$$F(x) = \frac{b^{N+1}-b^N}{b^{N+1}-1} b^{-x} \ \ (x = 0,1,2, \ldots, N) \tag{5}$$

If Y be the total number of candidates who have appeared for the examination, the number of those who have studied x out of N parts of the syllabus can be expressed as,

$$y(x) = Y\, F(x) \tag{6}$$

To derive the expression for the percentile score, in terms of x, we have used the following definition in the present model [3].

$$percentile\ score\ of\ a\ student = 100 \times \frac{\begin{matrix} number\ of\ students\ with\ marks \\ equal\ or\ less\ than\ that\ of\ the\ student \end{matrix}}{\begin{matrix} total\ number\ of\ students\ who\ have \\ appeared\ for\ the\ examination \end{matrix}} \tag{7}$$

Let m be the marks obtained by a student. It is natural to assume that m increases as x increases. For simplicity, let us consider m to be a single valued function of x. The percentile score of a student, who has studied x number of parts, is therefore equal to the percentage of students who have studied x or a smaller number of parts of the syllabus. Thus, the formula for calculating the percentile score (based on eqns. 6 & 7) can be expressed as,

$$P = \frac{\sum_0^x y(x)}{Y} \times 100 = 100 \times \sum_0^x F(x) \tag{8}$$

Substituting equation (5) into equation (8) we get,

$$P = 100 \times \frac{b^{N+1}-b^{N}}{b^{N+1}-1} \sum_{0}^{x} b^{-x} = 100 \times \frac{b^{N+1}-b^{N}}{b^{N+1}-1} \frac{b-b^{-x}}{b-1} \tag{9}$$

Here $\sum_{0}^{x} b^{-x} = \frac{b-b^{-x}}{b-1}$, according to the formula for the sum of a geometric series [12]. For an extremely large value of N (for which $b^{N+1} \gg 1$) , P approaches $100(1 - b^{-x-1})$.
Let us define a parameter, named *preparation index* (K), which can be expressed as $K = x/N$. In terms of this parameter, the expression for the percentile score can be written as,

$$P = 100 \times \frac{b^{N+1}-b^{N}}{b^{N+1}-1} \frac{b-b^{-KN}}{b-1} \tag{10}$$

Equation (10) has been obtained by substituting $x = KN$ into equation (9). For an extremely large value of N, P approaches $100(1 - b^{-KN-1})$.

Let S be the total number of questions to be answered in the examination. Since K is the fraction of the syllabus studied by a certain number of students, the product KS can be regarded as the simplest estimate of the average number of questions attempted by a member of that group. The number of questions that has been answered correctly can be expressed as, gKS, where $0 < g \leq 1$. The number of questions answered incorrectly is therefore, $(1-g)KS$. The parameter g is such that $1/g$ can be regarded as a measure of the degree of difficulty of the question paper. The easier the questions, the closer would be the value of g to unity. Let q be the marks obtained by a candidate for each correct answer and r be the marks deducted for an incorrect answer (*negative marking*). Thus qS is the total marks. Based on these values, the average marks secured by a student, whose preparation index is K, is given by,

$$m = qgKS - r(1-g)KS \tag{11}$$

The percentage of marks obtained by a student can be expressed as,

$$M = 100 \times \frac{m}{qS} = 100 \times K\left[g - \frac{r}{q}(1-g)\right] \tag{12}$$

Eliminating K from equations (10) and (12), we get,

$$P = 100 \times \frac{b^{N+1}-b^{N}}{b^{N+1}-1} \frac{1}{b-1}\left[b - b^{-\frac{NM}{100\left\{g-\frac{r}{q}(1-g)\right\}}}\right] \tag{13}$$

Equation (13) is an expression for the percentile score (P) in terms of the percentage of marks (M) obtained by a student.
The number of candidates who have secured marks smaller than or equal to that of a certain student, who has studied x parts of the syllabus, is $YP/100$, according to the definition of the percentile score (eqn. 7). Therefore, the expected rank of the student is given by,

$$R = Y - \frac{YP}{100} = \frac{Y}{100}[100 - P] \tag{14}$$

Substituting the expression for the percentile score (P) from equation (13) into equation (14) we get,

$$R = Y\left[1 - \frac{b^{N+1}-b^{N}}{b^{N+1}-1}\frac{1}{b-1}\left[b - b^{-\frac{NM}{100\left\{g-\frac{r}{q}(1-g)\right\}}}\right]\right] \tag{15}$$

Equation (15) is an expression for the expected rank (R) of a candidate in terms of the percentage of marks (M). To assess the performance of a particular candidate properly, relative to other candidates, it would be more meaningful to calculate the ratio R/Y than R. Using equation (15), R/Y is given by,

$$\frac{R}{Y} = 1 - \frac{b^{N+1}-b^{N}}{b^{N+1}-1}\frac{1}{b-1}\left[b - b^{-\frac{NM}{100\left\{g-\frac{r}{q}(1-g)\right\}}}\right] \tag{16}$$

A relation between R/Y and P can be obtained from equation (14), which is $R/Y = 1 - P/100$. Compared to the prediction of the percentile score, the rank of a candidate cannot be predicted with much accuracy because, there can be more than one candidate securing the same marks and they are not to be assigned the same rank.

Using equations (13) and (16), one can determine, in principle, the values of the parameters N, b and g with the help of the marks-*versus*-percentile-*versus*-rank data based on the results of the examinations held in the past.

The mean values for x and x^2 are given by, $\mu(x) = \sum_0^N x\, F(x)$ and $\mu(x^2) = \sum_0^N x^2\, F(x)$, respectively [9, 11]. The standard deviation for x can be expressed as, $\sigma(x) = \sqrt{\sum_0^N F(x)(x-\mu)^2} = \sqrt{\mu(x^2) - [\mu(x)]^2}$ [9, 11]. Substituting for $F(x)$ in these expressions from equation (5) we get,

$$\mu(x) = \frac{b^{N+1}-b^{N}}{b^{N+1}-1}\sum_0^N \frac{x}{b^x} = \frac{b^{N+1}-b^{N}}{b^{N+1}-1}\left(\frac{1}{b} + \frac{2}{b^2} + \frac{3}{b^3} + \cdots + \frac{N}{b^N}\right) \tag{17}$$

$$\sigma(x) = \left[\frac{b^{N+1}-b^{N}}{b^{N+1}-1}\left(\frac{1}{b} + \frac{4}{b^2} + \frac{9}{b^3} + \cdots + \frac{N^2}{b^N}\right) - \left(\frac{b^{N+1}-b^{N}}{b^{N+1}-1}\left(\frac{1}{b} + \frac{2}{b^2} + \frac{3}{b^3} + \cdots + \frac{N}{b^N}\right)\right)^2\right]^{1/2} \tag{18}$$

The series sum, for each of the two equations above, can be calculated numerically.

Based on the expression $K = x/N$, we have $\mu(K) = \mu(x)/N$ and $\sigma(K) = \sigma(x)/N$. Using these two relations in equation (12) we obtain,

$$\mu(M) = 100 \times \frac{\mu(x)}{N}\left[g - \frac{r}{q}(1-g)\right] \tag{19}$$

$$\sigma(M) = 100 \times \frac{\sigma(x)}{N}\left[g - \frac{r}{q}(1-g)\right] \tag{20}$$

Equations (19) and (20) are respectively the expressions for the mean and standard deviation of the percentage of marks obtained by the examinees. The values of $\mu(x)$ and $\sigma(x)$ in these expressions have to be calculated using equations (17) and (18) respectively.

3. Results and Discussions

Based on the mathematical expressions derived in the previous section, we have shown graphically the behaviours of the functions $F(x)$, P, R/Y in the Figures 1-8. The effect of variation of the parameters N, b and g, upon their behaviours, have been shown in these plots, which are self-explanatory. In the plots showing the variation of the percentile score (P), it is found to increase (as a function of K or M) with a gradually decreasing slope to reach its highest value, i.e., 100. In the plots of R/Y, it decreases with M, at a gradually slower rate. To predict the percentile score of a candidate from marks (using eqn. 13), with sufficient accuracy, the values of the parameters N, b and g have to be determined correctly by analysing the data (marks-*versus*-percentile) based on the results of examinations held in the past. Using these values, $\mu(K)$, $\sigma(K)$, $\mu(M)$ and $\sigma(M)$ can be determined with appreciable accuracy. The values of $\mu(K)$ and $\sigma(K)$ may be regarded as representing quantitatively the overall academic ability or proficiency of the entire community of students studying for a certain examination. In the same way, the values of $\mu(M)$ and $\sigma(M)$ represent a measure of their collective performance in the examination. Due to the unavailability of data, we have chosen different combinations of values of these parameters for our plots.

4. Concluding Remarks

The purpose of constructing the present model is to derive simple mathematical expressions to estimate the expected rank and the percentile score of a student based on the marks obtained in an examination. While defining the probability function $F(x)$, we assumed for simplicity that its value decreases monotonically with x, which means that $F(x)$ has its highest value (i.e., a) at $x = 0$. In reality, the level of preparation (or academic strength) of the students can be such that the probability function $F(x)$ has its peak at $x = x_0$ with $x_0 > 0$. This behaviour of the function indicates that a student is most likely to have studied x_0 number of parts of the syllabus which is divided into N parts in the present model. For the improvement of this model, one may choose several functional forms of $F(x)$ for which the highest value is not at $x = 0$. One of the forms, showing this behaviour, is $F(x) = ab^{-l(x-x_0)^2}$ (with $0 < a < 1$, $b > 1$, $l > 0$, $x_0 > 0$) which has its peak at $x = x_0$. The expression for m (eqn. 11) has been derived on the basis of an assumption that a student won't generally attempt questions from unfamiliar areas. But, contrary to this assumption, there can be cases where a student might be tempted to attempt questions from the parts of the syllabus not studied at all. There must be several factors governing the choice of questions attempted by a student. To improve the present model we need to derive a new expression for m, taking into account different possibilities regarding the choice of questions to be attempted by a candidate. The limitation of the present study is that we have not been able to validate our model with the help of the marks-*versus*-percentile (or marks-*versus*-rank) data based on the results of previous years, because these data are not available through internet from the agencies that organise the examinations. Using the simple mathematical scheme discussed in the present study, the readers of this article, who have access to the data of previous years, will be able to determine the values of the parameters N, b and g correctly so that predictions can be made using the expressions for P and R/Y (eqns. 13 & 16 respectively) with sufficient accuracy.

Acknowledgement

The author of this article is thankful to all colleagues, friends, students and family members for their encouragement, support and constructive criticism.

FIGURES

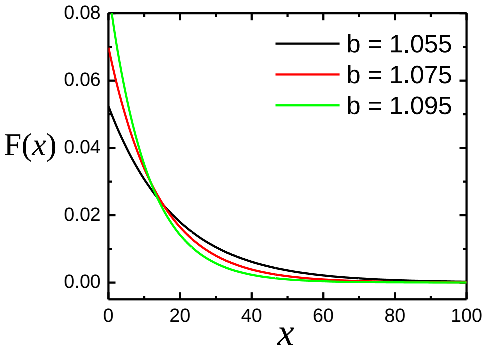

Fig 1: Plots of probability function $F(x)$ *versus* x, for three values of b. Here $N = 100$, $g = 1$, $q = 4$ and $r = 1$.

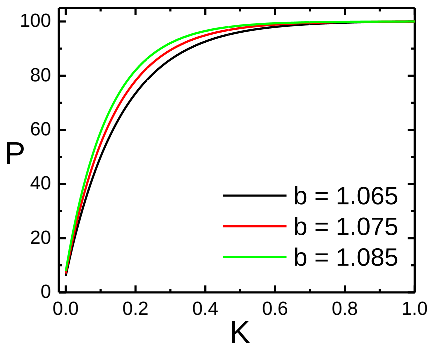

Fig 2: Plots of percentile score (P) *versus* preparation index (K), for three values of b. Here $N = 100$, $g = 1$, $q = 4$ and $r = 1$.

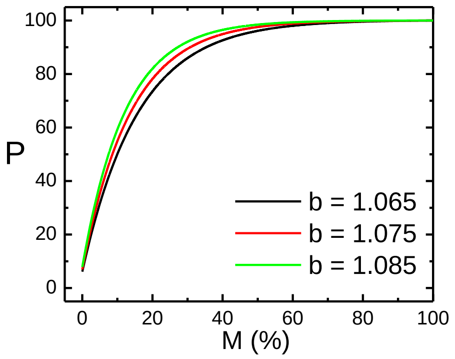

Fig 3: Plots of percentile score (P) *versus* percentage of marks (M), for three values of b. Here $N = 100$, $g = 1$, $q = 4$ and $r = 1$.

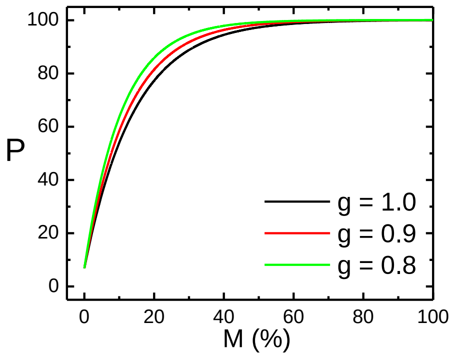

Fig 4: Plots of percentile score (P) *versus* percentage of marks (M), for three values of g. Here $N = 100$, $b = 1.075$, $q = 4$ and $r = 1$.

FIGURES

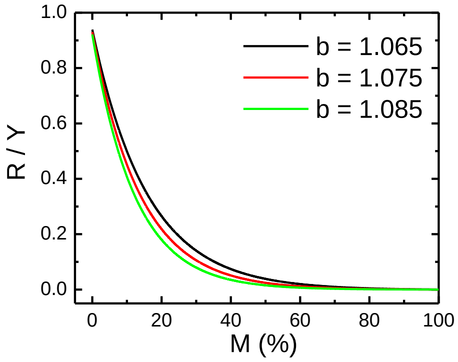

Fig 5: Plots of *Rank/Number of candidates* (R/Y) *versus* percentage of marks (M), for three values of b. Here $N = 100$, $g = 1$, $q = 4$ and $r = 1$.

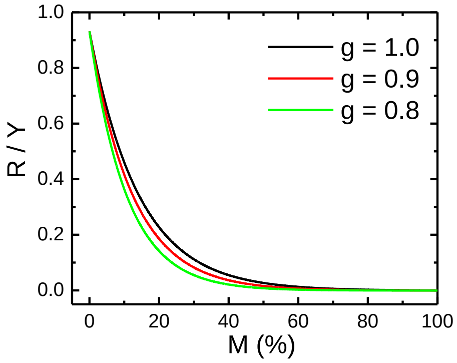

Fig 6: Plots of *Rank/Number of candidates* (R/Y) *versus* percentage of marks (M), for three values of g. Here $N = 100$, $b = 1.075$, $q = 4$ and $r = 1$.

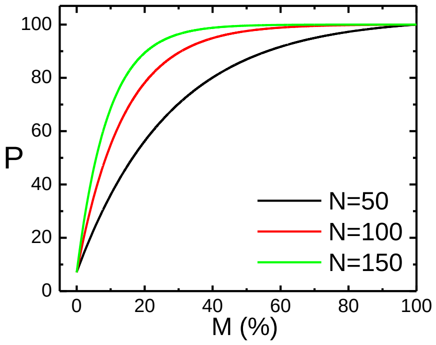

Fig 7: Plots of percentile score (P) *versus* percentage of marks (M), for three values of N. Here $b = 1.075$, $g = 1$, $q = 4$ and $r = 1$.

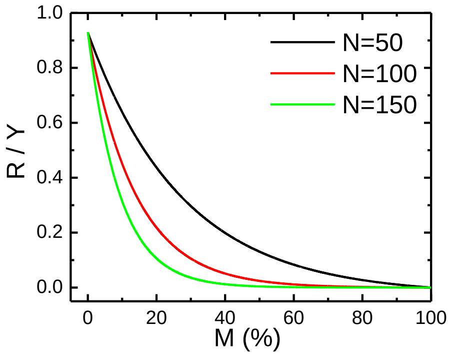

Fig 8: Plots of *Rank/Number of candidates* (R/Y) *versus* percentage of marks (M), for three values of N. Here $b = 1.075$, $g = 1$, $q = 4$ and $r = 1$.

References

1. Lutz M and McGill M. Percentile Rank. In: Salkind N, editors. Encyclopaedia of Research Design. Sage Publications; c2010. p. 1027.
   Available from: https://methods.sagepub.com/reference/encyc-of-research-design/n310.xml
   Chapter DOI: https://doi.org/10.4135/9781412961288

2. Official website of the National Testing Agency (NTA), Government of India.
   Available from: https://nta.ac.in/Home

3. Procedure to be adopted for compilation of Normalized Scores for multisession Papers in Common University Entrance Test (Undergraduate): CUET (UG) – 2022.
   Available from: https://nta.ac.in/Download/Notice/Notice_20220920220719.pdf

4. Lindsay CA and Prichard MA. An Analytical Procedure for the Equipercentile Method of Equating Tests, Journal of Educational Measurement. 1971;8(3):203-207.

5. Sharma M. CUET UG 2022 result: Equipercentile method explained. News / Education Today / India Today. 2022.
   Available from: https://www.indiatoday.in/education-today/news/story/cuet-ug-2022-result-equipercentile-method-explained-2001174-2022-09-16

6. IBPS Equipercentile Method, IBPS Marks Calculation Method, Byju's Exam Prep. Available from: https://byjus.com/bank-exam/ibps-equipercentile-method/

7. JEE Mains Result 2023 - How Percentile Score is Calculated, shiksha.com.
   Available from: https://www.shiksha.com/engineering/jee-main-exam-results

8. UGC NET Normalisation of Marks 2023: Result Preparation Criteria, shiksha.com.
   Available from: https://www.shiksha.com/sarkari-exams/teaching/articles/ugc-net-normalisation-of-marks-result-preparation-criteria-blogId-26373

9. Boas ML. Mathematical Methods in the Physical Sciences. 3rd ed. Wiley-India; c2009.

10. Arfken GB and Weber HJ. Mathematical Methods for Physicists. 4th ed. Academic Press, Inc.; c1995.

11. Riley KF, Hobson MP and Bence SJ. Mathematical Methods for Physics and Engineering. 3rd ed. Cambridge University Press; c2006.

12. Jordan D and Smith P. Mathematical Techniques. 2nd ed. Oxford University Press; c1997.