BMS Publishing House

MULTIDISCIPLINARY JOURNAL OF
EDUCATIONAL
RESEARCH INNOVATION AND DEVELOPMENT

VOL. 3 | ISSUE 2
JUL–DEC 2026
ISSN 3048-4685
www.bmsgroup.in

RESEARCH ARTICLE

DOI-10.5281/zenodo.21916102

A Method of Data Analysis to Determine the Preparation Level of Students using *Marks-versus-Percentile* Data

Sudipto Roy

Department of Physics, St. Xavier's College,
Kolkata, West Bengal, India
Email: roy.sudipto@sxccal.edu
ORCID URL: https://orcid.org/0000-0002-8811-2511

Abstract

This study presents a simple method for analysing marks-versus-percentile data from an examination. The main purpose is to obtain a measure of the overall preparation level of the examinees. The method is developed using marks-versus-percentile data for the Joint Entrance Examination–Main [JEE (Main)] held in April 2024. An empirical mathematical function is used to describe the relation between marks and percentile score. From this function, a probability function for obtaining different marks is obtained. This probability function is then used to find the most probable marks, the average marks, and the standard deviation of marks. The probability of obtaining marks within a specified range and the probability of obtaining marks equal to or greater than a specified value can also be calculated. For the data used in this study, the fitted function gives a coefficient of determination of 0.99894. The most probable marks are found to be 19, while the expectation value and standard deviation are 47.26 and 49.74 marks, respectively. The sum of the calculated probabilities is very close to unity. The method is simple and can be applied to other examinations. Its accuracy can be improved by using a larger marks-versus-percentile dataset.

Keywords: *Marks-versus-percentile, Percentile score, JEE (Main), Probability distribution, Most probable marks, Expectation value, Standard deviation, Examination data, Average preparation level, Empirical model.*

1. Introduction

A percentile score shows how well a candidate has performed compared with the other candidates in an examination. Percentile scores are also used in other areas to compare performance. They are used, for example, to evaluate journals, books, research papers, and researchers in different ways [1–4]. In an examination, marks alone may not always be enough to understand a candidate's position among all the candidates. A candidate may also need to know the percentile score corresponding to the marks obtained, especially when applying for

CITATION - Roy, S. (2026). A Method of Data Analysis to Determine the Preparation Level of Students using Marks-versus-Percentile Data. Multidisciplinary Journal of Educational Research Innovation and Development, 3(2), 1–9. https://doi.org/10.5281/zenodo.21916102

admission to different institutions. Therefore, it is useful to have a mathematical method for finding the percentile score when the marks are known. A mathematical function for obtaining the percentile score from the marks obtained in an examination was derived in an earlier study [5].

In this study, we propose a simple way to analyse marks-versus-percentile data using an empirical function. Besides finding the percentile score, it is often useful to know the chance of getting marks above a certain value. For this, we need a function that can give the probability of obtaining different marks. We have used marks-versus-percentile data to develop such a function. From the way the percentile score changes with marks, we have chosen an empirical formula for the percentile score as a function of marks. The values of the constants in this formula should be found from a sufficiently large set of marks-versus-percentile data from earlier results. In this study, we have used a small set of JEE (Main) data. Information about JEE (Main) is available from the National Testing Agency (NTA) [6,7], while the marks-versus-percentile data used here were obtained from an online source [8]. We have found the values of the constants that give the best fit to the data. From the fitted function, we have obtained a probability function for the marks. This function gives the most probable marks and can also be used to find the probability of getting marks equal to or greater than any chosen value. We have also shown how the average marks and standard deviation can be calculated. The method is simple, and its accuracy can be improved by using a larger set of data from previous examinations.

2. Model Formulation

Let us assume that there exists a function $P(m)$ representing the percentile score of a candidate who has scored m marks in an examination. The definition of percentile score is as follows [5].

$$P(m) = \frac{number\ of\ candidates\ with\ marks \leq m}{total\ number\ of\ candidates} \times 100 \tag{1}$$

Let $f(m)$ be the fraction of examinees who have scored m marks each. This fraction can be expressed as,

$$f(m) = \frac{number\ of\ candidates\ with\ marks = m}{total\ number\ of\ candidates} \tag{2}$$

Based on the definition of $P(m)$, represented by equation (1), $f(m)$ can be expressed as,

$$f(m) = \frac{P(m) - P(m-\beta)}{100} \tag{3}$$

In writing equation (3), we have assumed that the variable m changes in steps of β. Thus, $(m-\beta)$ is the possible score just below m. Equation (3) cannot be the definition of $f(m)$ when m is the lowest score obtainable in the examination. In that case, we have, $f(m) = P(m)/100$, which is valid according to equations (1) and (2). If the number of candidates taking the examination is sufficiently large, the fraction of candidates, each having m marks, can approximately be regarded as the probability of scoring that marks in the examination. In the present study, we have used the function $f(m)$ as the probability that an examinee scores m marks.

The probability of obtaining marks between any two values of m, such as q and r (i.e., $q \leq m \leq r$), can be expressed as [9-12],

$$P(q,r) = \sum_{m=q}^{r} f(m) \tag{4}$$

Let $g(m)$ be the probability that a candidate scores marks $\geq m$ in the examination. It can be expressed as [9-12],

$$g(m) = \sum_{x=m}^{m_{max}} f(x) \tag{5}$$

In equation (5), m_{max} is the highest marks obtainable in the examination.
The expectation value of marks, denoted by m_{ex} here, is expressed as [9-12],

$$m_{ex} = \sum_{m=m_{min}}^{m_{max}} m\, f(m) \tag{6}$$

In equation (6), m_{min} stands for the lowest marks obtainable in the examination.
The standard deviation of marks, denoted here by m_{sd}, is given by [9-12],

$$m_{sd} = \left[\sum_{m=m_{min}}^{m_{max}} (m - m_{ex})^2\, f(m)\right]^{\frac{1}{2}} \tag{7}$$

A simpler formula for m_{sd} is given below [9-12].

$$m_{sd} = \left[\left(\sum_{m=m_{min}}^{m_{max}} m^2\, f(m)\right) - (m_{ex})^2\right]^{\frac{1}{2}} \tag{8}$$

3. Data Analysis

For the present study, we have used a set of *marks vs. percentile* data of the examination named JEE (Main), held in April 2024, obtained from a certain website [8]. Table-1 of this article contains that dataset.

Table 1. *Marks vs. Percentile* Data of April 2024 session of JEE (Main)

Marks	Percentile	Marks	Percentile
82	80	158	96.5
96	85	163	97
113	90	170	97.5
119	91	177	98
124	92	188	98.5
130	93	199	99
137	94	220	99.5
145	95	226	99.6
149	95.5	234	99.7

153	96	271	99.8

According to this table, the percentile score increases by 13 for an increase in marks from 82 to 130. For almost the same change in marks, i.e., from 130 to 177, the percentile score increases by 5. Based on this nature of dependence of the percentile score upon marks, we have assumed the following empirical expression to represent the percentile score $P(m)$ as a function of marks (m).

$$P(m) = A[1 - Exp\{-b(m + c)\}]^n \tag{9}$$

The JEE (Main) question paper has 75 questions carrying 4 marks each, the full marks being 300. Each wrong answer causes a deduction of 1 mark. Therefore, the value of the parameter β (of eqn. 3) is 1.

The value of $P(m)$ should be positive and it should be increasing with marks (m), as per equation (1). Due to negative marking, the lowest value of $P(m)$ (i.e., zero) should correspond to a negative value of marks (m). For these conditions to be satisfied, we must have $A, b, c, n > 0$, in equation (9).

Since the maximum marks obtainable in the JEE (Mains) is 300 (the *full marks*), equation (9) must yield $P(m) = 100$ (the highest percentile score) for $m = 300$. This fact leads to the following value for the constant parameter A which belongs to the expression for $P(m)$ (eqn. 9).

$$A = \frac{100}{[1 - Exp\{-b(300 + c)\}]^n} \tag{10}$$

As per equation (9), $P(m) = 0$ for $m = -c$. It means that, there is no examinee for whom $m \leq -c$. Thus, the lowest marks scored is, $m_{min} = -c + 1$, since m changes in steps of 1. For any analysis using equation (9), one should consider the values of m lying in the range: $-c + 1 \leq m \leq 300$.

The constant parameter b in equation (9) determines how rapidly $P(m)$ approaches its maximum value (i.e., 100, the highest percentile score) as m increases. It has been observed that, as the difficulty level rises, one gets a greater percentile score for the same marks secured in the examination. This is due to a fall in the number of candidates scoring higher marks. The value of $P(m)$, for a certain m, decreases as the parameter n increases. Thus, both b & n need to be changed, as the difficulty level changes.

Using equation (10) in equation (9), one gets,

$$P(m) = 100 \frac{[1 - Exp\{-b(m + c)\}]^n}{[1 - Exp\{-b(300 + c)\}]^n} \tag{11}$$

Using equation (11) in equation (3), we can write,

$$f(m) = \frac{[1 - Exp\{-b(m + c)\}]^n - [1 - Exp\{-b(m - 1 + c)\}]^n}{[1 - Exp\{-b(300 + c)\}]^n} \tag{12}$$

The number of examinees for each session of JEE (Main) is close to 1 million. Due this large sample size, the function $f(m)$ can be regarded as the probability that an examinee obtains m marks in the examination. Here, m can be treated as a discrete random variable, with values in the range: $-c + 1 \leq m \leq 300$.

Using equation (12) in equation (4), we get the following expression for $P(q, r)$,

$$P(q,r) = \sum_{m=q}^{r} \frac{[1 - Exp\{-b(m+c)\}]^n - [1 - Exp\{-b(m-1+c)\}]^n}{[1 - Exp\{-b(300+c)\}]^n} \tag{13}$$

Using equation (12) in equation (5), we get the following expression for $g(m)$,

$$g(m) = \sum_{x=m}^{300} \frac{[1 - Exp\{-b(x+c)\}]^n - [1 - Exp\{-b(x-1+c)\}]^n}{[1 - Exp\{-b(300+c)\}]^n} \tag{14}$$

Using equation (12) in equation (6), we get the following expression for m_{ex},

$$m_{ex} = \sum_{m=-c+1}^{300} m \left[\frac{[1 - Exp\{-b(m+c)\}]^n - [1 - Exp\{-b(m-1+c)\}]^n}{[1 - Exp\{-b(300+c)\}]^n}\right] \tag{15}$$

Using equations (12) and (15) in equation (8), one gets the following expression for m_{sd},

$$m_{sd} = \left[\left(\sum_{m=-c+1}^{300} m^2 \, \frac{[1 - Exp\{-b(m+c)\}]^n - [1 - Exp\{-b(m-1+c)\}]^n}{[1 - Exp\{-b(300+c)\}]^n}\right) - \left(\sum_{m=-c+1}^{300} m \left[\frac{[1 - Exp\{-b(m+c)\}]^n - [1 - Exp\{-b(m-1+c)\}]^n}{[1 - Exp\{-b(300+c)\}]^n}\right]\right)^2\right]^{\frac{1}{2}} \tag{16}$$

4. Results and Discussion

Fitting equation (11) to the *marks vs. percentile* dataset of Table-1, we have obtained, $b = 0.02344$, $c = 46$ and $n = 4.44651$. For this curve fitting, the coefficient of determination is, $R^2 = 0.99894$. Using the values of b, c and n, one obtains $A = 100.134$ from equation (10). For $m = -c \equiv -46$, we have $P(m) = 0$ (as per eqn. 11), which implies that no examinee has obtained marks ≤ -46. Thus, $m_{min} = -c + 1 = -45$. For no analysis, one should consider the values of m for which $m \leq m_{min} - 1$ or, $m \leq -46$.

Figure 1 shows the best fit of the function $P(m)$ (eqn. 11) to the dataset of Table-1. Figure 2 shows the same $P(m)$-versus-m curve for the entire range of values of m.

Figure 3 shows the variation of the probability $f(m)$ as a function of marks (m), based on equation (12). It shows that a candidate has the highest probability of scoring 19 marks and the corresponding probability is close to 0.00981.

Figure 4 shows the variation of the cumulative probability $g(m)$ as a function of marks (m), based on equation (14). It seems to be falling at a faster rate beyond $m = 250$. For the clarity of data, we have chosen logarithmic scales for showing the values of both $f(m)$ and $g(m)$, in Figures 3 and 4 respectively.

Using equation (13), the probability of scoring marks between 100 & 200 (including these two values) comes out to be $P(100, 200) = 0.126371$, which means that 12.6371% of the examinees are likely to score marks in this range.

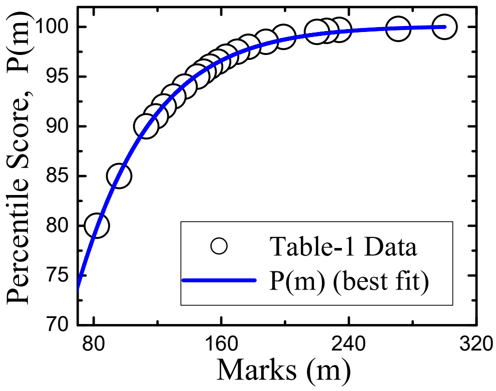

Figure 1. Percentile score versus marks. The blue curve is the best fit to the *Table-1* data (circles), based on the function $P(m)$ of equation (11).

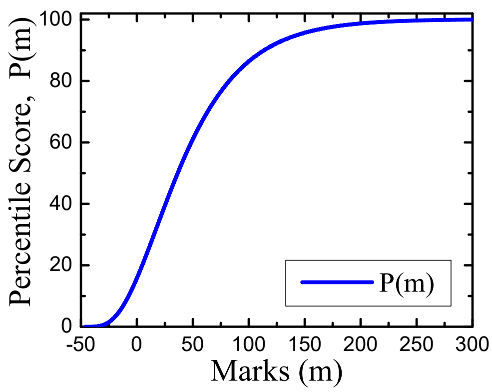

Figure 2. Percentile score versus marks, using equation (11), over the entire range of m, for the parameter values required for the best fit in Fig. 1.

Using equations (15) and (16), respectively, we have obtained $m_{ex} = 47.26$ and $m_{SD} = 49.74$. If the function $f(m)$ (of eqn. 12) has to serve as the probability of scoring m marks in the examination, its sum over all values of m should be unity. Our calculation shows, $\sum_{m=-45}^{300} f(m) = 1.000002$. The deviation of the sum from unity may be due to the smallness of the dataset (Table-1) used in this study. If we had used a sufficiently large dataset (i.e., a set of *marks vs. percentile* data having almost all possible values of m), we could probably have obtained a sum closer to unity.

The expectation value, the standard deviation and the most probable value of m, constitute a set of numbers which may collectively be regarded as an index for the average preparation level of the examinees who took the JEE (Main) in the session of April 2024.

One can use equations (11) to (16) for making predictions for any session of the JEE (Main) with an accuracy which depends upon the values of the constant parameters (b, c & n) associated with $P(m)$. To determine their values properly, one must use a reasonably large data of *marks vs. percentile*.

The coefficient of variation (or CV, which is, *standard deviation / mean*) can be regarded as an index of the overall performance of the examinees. The smaller the value of CV, the better would be the collective performance of the candidates who have taken the examination. For the JEE (Main) of April 2024, the mean and standard deviation of marks are respectively 47.26 and 49.74. Therefore, $CV = \frac{49.74}{47.26} = 1.052$ or 105.2%, which is quite large, indicating a poor collective performance by the examinees.

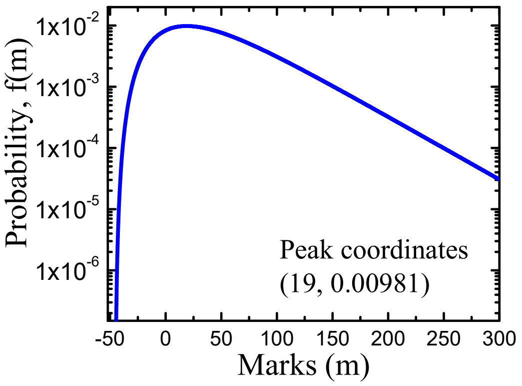

Figure 3. Probability versus marks. The highest probability corresponds to $m = 19$. The scale along the vertical axis is logarithmic.

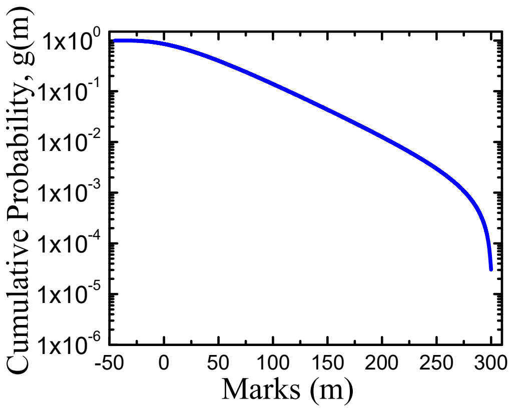

Figure 4. Cumulative probability versus marks. $g(m)$ is the probability of scoring marks $\geq m$. The scale along the vertical axis is logarithmic.

5. Conclusion

A simple mathematical method has been presented for analysing marks-versus-percentile data from an examination. The method converts the percentile information into a probability function for the marks obtained by the examinees. From this function, useful quantities such as the most probable marks, average marks, standard deviation, and different score probabilities can be obtained. For the JEE (Main) data studied here, the most probable marks are 19, while the average marks and standard deviation are 47.26 and 49.74, respectively. These quantities can together provide a simple measure of the overall preparation level of the examinees.

The present analysis is based on a relatively small set of marks-versus-percentile data. Therefore, the accuracy of the results and of any prediction based on the method depends on the amount and quality of the available data. A larger dataset covering more marks would be expected to give more reliable parameter values and predictions. The method is not limited to JEE (Main) and can be applied to other examinations if suitable marks-versus-percentile data are available. The empirical function used here may also be replaced by another suitable function in future studies. The results obtained using different functions can then be compared.

References

1. Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for information Science and Technology*, *64*(2), 372-379. Available at: https://doi.org/10.1002/asi.22775
2. Bornmann, L., & Williams, R. (2020). An evaluation of percentile measures of citation impact, and a proposal for making them better. *Scientometrics*, *124*(2), 1457-1478. Available at: https://doi.org/10.1007/s11192-020-03512-7
3. Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of informetrics*, *7*(1), 158-165. Available at: https://doi.org/10.1016/j.joi.2012.10.001
4. Yamamoto, K., & Yasunaga, T. (2022). A percentile rank score of group productivity: an evaluation of publication productivity for researchers from various

fields. *Scientometrics*, *127*(4), 1737-1754. Available at: https://doi.org/10.1007/s11192-022-04278-w

5. Roy, S. (2023). Estimation of percentile score based on marks obtained in an examination: A simple mathematical model. *International Journal of Physics and Mathematics*, *5*(1), 25-32. Available at: https://doi.org/10.33545/26648636.2023.v5.i1a.48
6. The website (homepage) of the National Testing Agency (NTA): https://nta.ac.in/
7. The webpage of the NTA regarding JEE (Main): https://jeemain.nta.ac.in/
8. The data (*marks vs. percentile*), used for the present study, have been obtained from the following website: https://collegedunia.com/exams/jee-main/rank-vs-marks
9. Boas, M. L. (2006). *Mathematical methods in the physical sciences*. John Wiley & Sons.
10. Arfken, G. B., Weber, H. J., & Harris, F. E. (2011). *Mathematical methods for physicists: a comprehensive guide*. Academic press.
11. Riley, K. F., Hobson, M. P., & Bence, S. J. (2006). *Mathematical methods for physics and engineering: a comprehensive guide*. Cambridge university press.
12. Ross, S. (1976). *A first course in probability*. New York: Macmillan.