OPEN ACCESS

Ind. Jr. of Mod. Res. and Rev | *PEER-REVIEWED JOURNAL* | *Volume 4 Issue 8 [Aug] Year 2026*

Indian Journal of Modern Research and Reviews

This Journal is a member of the *'Committee on Publication Ethics'*

Online ISSN:2584-184X

Research Article

Marks-versus-Percentile Data: A Simple Mathematical Analysis Based on a Hypothetical Dataset

Sudipto Roy

Department of Physics, St. Xavier's College, Kolkata, West Bengal, India

Corresponding Author: *Sudipto Roy

DOI: https://doi.org/10.5281/zenodo.21979142

Abstract

This study presents a simple mathematical method for analysing examination-related data and describing the overall performance of a group. A hypothetical dataset of scores and percentile values is used to demonstrate the method. An empirical formula is fitted to the percentile data. From this formula, a probability distribution is obtained. This distribution is then used to find the most probable score, the average score, the spread of scores, and the probability of obtaining a score within a given range or above a given level. The method also shows how percentile values can be estimated from scores when suitable model parameters are available. It provides a simple way to study the distribution of scores and to obtain useful measures of the performance of a group. The method can also be applied to data from an actual examination. However, the results presented in this study are only illustrative because the dataset is hypothetical and is based on a limited set of data. The results should therefore not be taken as an exact description of any particular examination. More reliable results would require a sufficiently large and representative dataset from previous examinations. The accuracy of the estimates will depend on the amount and quality of the data used to determine the model parameters.

Manuscript Information

- **ISSN No:** 2584-184X
- **Received:** 07-07-2026
- **Accepted:** 12-07-2026
- **Published:** 17-08-2026
- **MRR:**4(8); 2026: 107-113
- **©2026, All Rights Reserved**
- **Plagiarism Checked:** Yes
- **Peer Review Process:** Yes

How to Cite this Article

Roy S. *Marks-versus*-Percentile Data: A Simple Mathematical Analysis Based on a Hypothetical Dataset. Indian J Mod Res Rev. 2026;4(8):107-113.

Access this Article Online

www.mrrjournal.in

KEYWORDS: Percentile Score, Most Probable Marks, Marks-versus-Percentile Record, Probability Distribution, Statistical Analysis, Mathematical Modelling.

1. INTRODUCTION

Percentile score is a measure of one's performance relative to the rest of the candidates who have taken a certain examination. Percentile scores are also calculated to evaluate journals, books, research papers and researchers, in various ways, to obtain relative measures of performance [1-4]. The merit of a publication can be judged by finding the percentage of publications (in the same field) which have received equal and less number of citations. Apart from the marks scored in an examination, one often needs to know the percentile score to find his/her chance to get admission to different institutions. There should be a mathematical way to determine the percentile score when one already knows the marks obtained by an examinee. In a study carried out in 2023, a function was mathematically derived (based on probabilistic considerations), using which the percentile score of an examinee can be calculated from marks obtained in an examination [5].

In the present study, we have attempted to find a much simpler way to analyze *marks vs. percentile* data with the help of an empirical function. Apart from calculating the percentile score, it might often be necessary for a candidate to estimate the chance of scoring marks beyond a certain percentage of the full marks for an examination. It can be done only if there is a function available to them, using which, one can calculate the probability of scoring any marks. To formulate a function of that kind we have used a set of *marks vs. percentile* data. Based on the behavior of how the percentile score increases with marks, we have proposed here an empirical formula which is meant to serve as an expression for the percentile score as a function of marks. For this formula to work properly in a particular year, the constant parameters of the formula have to be determined using a reasonably large dataset of marks-versus-percentile data from the previous year. We have used a small hypothetical dataset of marks-versus-percentile data, constructed by considering the general trend seen in publicly available information related to JEE (Main) preparation and examination analysis [6-8]. We have determined the parameter values required for a best fit of our empirical formula to the dataset. Based on this formula, a function $f(m)$ has been derived, which is the probability of obtaining m marks in the examination. The values of $f(m)$ as a function of m has given us the most probable marks (m) and the corresponding probability. Using $f(m)$, one can estimate the probability of getting marks equal to or greater than a certain value of m. A useful check is that the sum of $f(m)$, over all possible values of m, has come out to be very close to unity, supporting its interpretation as a probability distribution, although we have used a small dataset to formulate this function. We have shown how the expectation value and standard deviation can be calculated from *marks vs. percentile* data. The method of data analysis is very simple and it can be used to make predictions with sufficient accuracy if one uses a large volume of data corresponding to previous years' results of the examination.

2. METHODOLOGY

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],

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

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],

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

In equation (6), m_{min} stands for the lowest marks obtainable in the examination where $m_{min} = \lfloor -c \rfloor + 1$.
The standard deviation of marks, denoted here by m_{sd}, is given by [9-12],

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

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

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

3. ANALYSIS OF DATA

For the present study, a hypothetical set of marks-versus-percentile data has been considered. The data have been constructed by considering the general trend seen in marks-versus-percentile information available from sources related to JEE (Main) preparation and examination analysis [6-8]. The data are not intended to represent the actual results of any particular examination. They are used only to demonstrate the proposed method. Table 1 contains the hypothetical dataset used in the analysis.

Table 1. Hypothetical marks-versus-percentile data

Marks	Percentile	Marks	Percentile
83	80.15	157	96.58
95	84.85	166	97.12
114	90.21	171	97.33
119	91.26	184	98.21
123	92.01	190	98.74
130	93.34	213	99.08
135	93.53	230	99.42
144	95.03	245	99.72
147	95.25	255	99.74
154	96.16	264	99.97

According to this table, the percentile score increases by about 13.19 for an increase in marks from 83 to 130. For a comparable increase in marks, from 130 to 184, the percentile score increases by about 4.87. Thus, the percentile score increases rapidly at lower marks but more slowly at higher marks. Based on this behaviour, we have assumed the following empirical expression to represent the percentile score $P(m)$ as a function of marks m.

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

For the present hypothetical dataset, the marks are considered to change in steps of 1. Therefore, the value of the parameter β in Eq. (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 considered in the present hypothetical dataset is 300, 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 in equation (9).

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

As per equation (9), $P(m) = 0$ for $m = -c$. Since m is an integer-valued variable changing in steps of 1, the lowest possible value of m satisfying $m > -c$ is $m_{min} = \lfloor -c \rfloor + 1$. For the present analysis, $c = 43.0640$, giving $m_{min} = -43$. Thus, the values of m considered in the analysis lie in the range $-43 \leq m \leq 300$.

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,

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

The empirical expression in equation (11) is considered for $m > -c$; for $m \leq -c$, $P(m) = 0$.
Using equation (11) in equation (3), we can write,

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

For the present analysis, $f(m)$ is treated as the probability of obtaining m marks, as described in the methodology. Here, m can be treated as a discrete random variable, with values in the range: $-43 \leq m \leq 300$.

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},

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

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

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

4. RESULTS AND DISCUSSION

Fitting equation (11) to the hypothetical marks-versus-percentile data in Table 1 gives $b = 0.0236246$, $c = 43.0640$, and $n = 4.25168$. The coefficient of determination is $R^2 = 0.999165$. Thus, the empirical function gives a very good fit to the 20 data points used in the present study.

Using these values of b, c, and n, equation (10) gives $A = 100.129$. According to equation (9), $P(m) = 0$ for $m = -c = -43.0640$. Therefore, since m changes in steps of 1, the lowest value of m considered in the analysis is $m = -43$, and the range of marks used in the analysis is $-43 \leq m \leq 300$.

Figure 1 shows the fitted percentile function together with the data points in Table 1. The curve follows the data closely over the range covered by the table.

Figure 2 shows the same function over the complete range of marks used in the analysis. The percentile score increases with marks and approaches 100 as the marks approach 300.

Figure 3 shows the probability $f(m)$ of obtaining m marks. The probability reaches its maximum at $m = 19$. Thus, according to the proposed model, 19 marks is the most probable score, and the corresponding probability is approximately 0.00989 (0.989%).

Figure 4 shows that the probability of obtaining at least m marks decreases as m increases. The decrease becomes particularly rapid at high marks, especially beyond about $m = 250$. This means that, according to the model, the probability of obtaining very high marks is relatively small.

Using equation (13), the probability of obtaining marks from 100 to 200, including both 100 and 200, is $P(100, 200) = 0.126654$. Thus, according to the proposed model, about 12.67% of the examinees are expected to score between 100 and 200 marks.

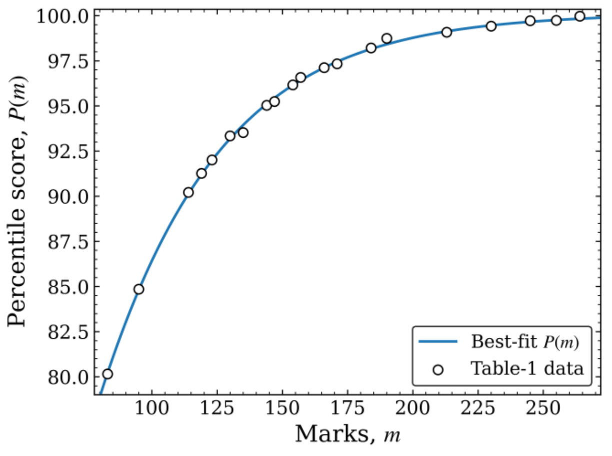

Figure 1. Percentile score $P(m)$ versus marks m. The circles show the data points in Table 1, and the blue curve shows the best fit obtained using equation (11).

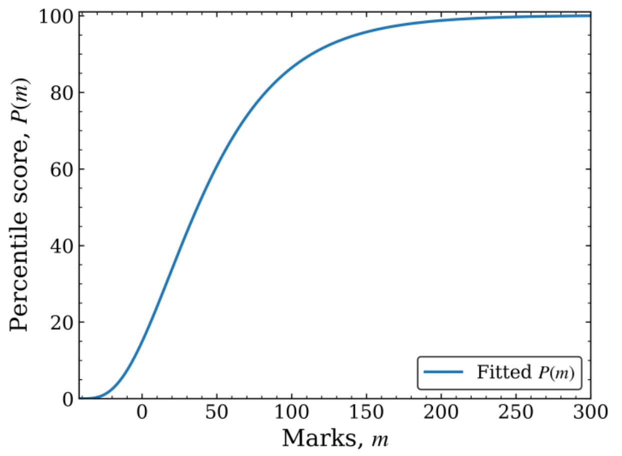

Figure 2. Percentile score $P(m)$ versus marks m over the complete range of marks ($-43 \leq m \leq 300$) used in the analysis, using the fitted parameter values.

Using equations (15) and (16), the mean marks and standard deviation are 47.54 and 49.61, respectively. The mean is higher than the most probable value of 19 marks. This difference occurs because the probability distribution is not symmetric and extends towards higher marks.

The sum of $f(m)$ over the complete range of marks is 1.000000, confirming that the calculated probability distribution is normalized. Thus, the function provides a normalized probability distribution over the range of marks considered in the analysis.

The mean marks, standard deviation and most probable marks provide useful measures of the overall marks distribution. These measures may be used together as indicators of the preparation level of the group represented by the hypothetical dataset.

Equations (11) to (16) can also be used to make estimates for other examinations. The accuracy of these estimates depends on the values of the parameters b, c, and n. These parameters should be determined from a sufficiently large set of marks-versus-percentile data from previous years.

The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, provides a measure of the spread of the marks relative to the mean.

For the hypothetical dataset, the mean marks are 47.54 and the standard deviation is 49.61. Therefore, $CV = 104.35\%$. This relatively large value indicates that the marks are widely spread relative to the mean.

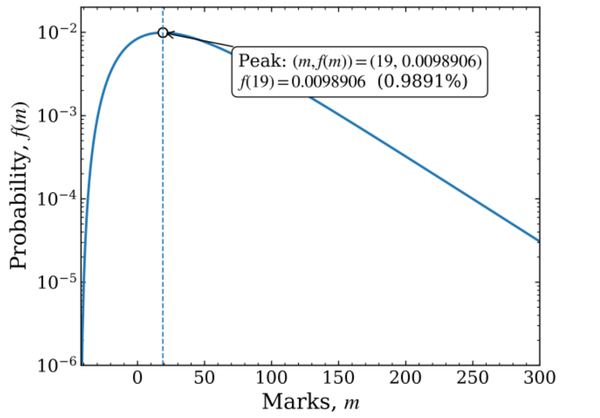

Figure 3. Probability $f(m)$ of obtaining m marks. The maximum occurs at $m = 19$, with $f(19) = 0.00989$. The vertical axis has a logarithmic scale.

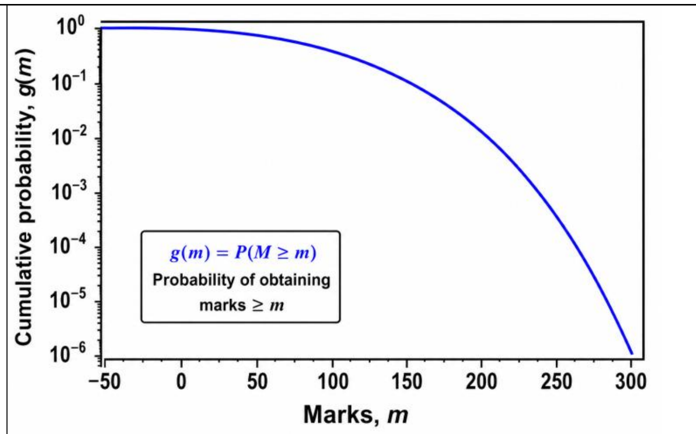

Figure 4. Cumulative probability $g(m) = P(M \geq m)$ as a function of marks m, showing the probability of obtaining marks greater than or equal to m.

The results obtained in this study are based on the hypothetical marks-versus-percentile data used for the analysis. Only 20 data points were available for fitting the empirical function. Therefore, the values of the most probable marks, mean marks, standard deviation and calculated probabilities may change if a larger and more complete dataset is used. The present results demonstrate the use of the proposed method and should not be taken as an exact description of the marks distribution of any particular examination.

5. CONCLUDING REMARKS

In the present study, we have formulated a new method to determine the average preparation level of a group, based on marks-versus-percentile data. The accuracy of this determination depends upon the volume of data used for this purpose. The present study uses a hypothetical dataset to demonstrate the method. The same analysis can be carried out using a sufficiently large dataset from an actual examination, presumably with greater accuracy. A candidate, who has somehow calculated the marks to be obtained, can estimate the percentile score using the expression for $P(m)$, if the values of the parameters b, c and n have already been determined based on a sufficiently large volume of data obtained from previous years' results. The empirical expression for $P(m)$, chosen for the present study (eqn. 9), can be replaced by a new function. One can do the entire analysis using that function and compare the findings with those of the present study.

REFERENCES

1. Waltman L, Schreiber M. On the calculation of percentile-based bibliometric indicators. J Am Soc Inf Sci Technol. 2013;64(2):372-379. doi:10.1002/asi.22775.
2. Bornmann L, Williams R. An evaluation of percentile measures of citation impact, and a proposal for making them better. Scientometrics. 2020;124(2):1457-1478. doi:10.1007/s11192-020-03512-7.
3. Bornmann L, Leydesdorff L, Mutz R. The use of percentiles and percentile rank classes in the analysis of bibliometric data: opportunities and limits. J Informetr. 2013;7(1):158-165. doi:10.1016/j.joi.2012.10.001.
4. Yamamoto K, Yasunaga T. A percentile rank score of group productivity: an evaluation of publication productivity for researchers from various fields. Scientometrics. 2022;127(4):1737-1754. doi:10.1007/s11192-022-04278-w.
5. Roy S. Estimation of percentile score based on marks obtained in an examination: a simple mathematical model. Int J Phys Math. 2023;5(1):25-32. doi:10.33545/26648636.2023.v5.i1a.48.
6. National Testing Agency. Information bulletin for Joint Entrance Examination (Main) 2025. New Delhi: National Testing Agency; 2025.
7. National Testing Agency. Declaration of the result/NTA scores for the Joint Entrance Examination [JEE (Main)-2025] of Paper 1 (B.E./B.Tech.). New Delhi: National Testing Agency; 2025.
8. Careers360. JEE Main 2024 marks vs percentile. Careers360; 2024.
9. Boas ML. Mathematical methods in the physical sciences. 3rd ed. Hoboken (NJ): John Wiley & Sons; 2006.
10. Arfken GB, Weber HJ, Harris FE. Mathematical methods for physicists: a comprehensive guide. 7th ed. Oxford: Academic Press; 2011.

11. Riley KF, Hobson MP, Bence SJ. Mathematical methods for physics and engineering: a comprehensive guide. 3rd ed. Cambridge: Cambridge University Press; 2006.
12. Ross S. A first course in probability. New York: Macmillan; 1976.

About the Author

Dr. Sudipto Roy is a faculty member in the Department of Physics at St. Xavier's College, Kolkata, West Bengal, India. He has been involved in teaching physics at the graduate and post-graduate levels for many years. He has nearly twenty- five years of teaching experience combining his tenures at different institutions.

He has carried out some research in the fields of cosmology, mathematical biology, mathematical/statistical modelling of various phenomena, organic light-emitting devices etc., and published articles in national and international journals over the last twenty-five years. His publications can be accessed through the link: https://www.researchgate.net/profile/Sudipto-Roy-5/research