WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Mathematical Model for Percentile and Rank Estimation in Competitive Examinations and its Analogy with Statistical Mechanics

Swastika Bhadra[1], Sumana Paul Mondal[2], Judith Ghosh[3], Sudipto Roy[4]

[1,2,3]UG Student (2022-2025), Department of Physics, St. Xavier's College, Kolkata, West Bengal, India.

[4]Faculty Member, Department of Physics, St. Xavier's College, Kolkata, West Bengal, India.

Emails : [1]swastikabhadra07@gmail.com,[2]rims2004.nov@gmail.com,[3]ghoshjudith@gmail.com,[4]roy.sudipto@sxccal.edu

DOI : https://doi.org/10.5281/zenodo.15852802

ABSTRACT

Assessing student performance and making decisions regarding college admissions heavily rely on percentile estimation. In this work, a basic mathematical model for evaluating the probability of syllabus coverage and corresponding marks has been developed. Based on the students' preparation level, we have defined empirical expressions for estimating the percentile scores and ranks of the students. Furthermore, we derive equations for the mean, standard deviation, and most probable marks. To account for the role played by negative marking in such examinations, we have modified our model to incorporate its effect. A crucial component of this model is the difficulty parameter whose effect is profound in the entire model. The findings have been discussed and then compared with real data graphically. Additionally, we graphically compare our model to a statistical mechanics-based formulation, establishing their equivalence. Our results demonstrate that percentile growth is non-linear, with small change in marks leading to large percentile shifts. External factors like paper difficulty and normalization techniques are indicated by deviations of real data from theoretical expectations. These insights can improve the accuracy of percentile predictions and guide preparation techniques for competitive exams.

Keywords: Percentile Estimation, Rank Estimation, Syllabus Coverage, Negative Marking, Difficulty Parameter, Statistical Mechanics, Maxwell-Boltzmann Distribution, JEE Mains, Normalization Process, Preparation Index

1. Introduction

All over the country, there are several examinations that use the normalization method to convert the raw score of individuals to percentiles. The percentiles obtained in these examinations by students determine the fates of the students in getting into a particular college for a particular course of study. Percentile ascertains the position of a student relative to the performance of other students appearing for these evaluations. Examinations like JEE Mains, JEE Advanced, CUET UG, UGC NET, etc. conduct exams in various sessions and over a period of time instead of a single day, for which different papers are prepared. No matter what steps are taken, there still lies an ambiguity of the difficulty of exam across different sessions which might lead to skewed results. Here, the abbreviation UGC stands for University Grant Commission and NET stands for National Eligibility Test. The students who attempt difficult paper sets might end up getting lower scores as compared to the students who are writing easier set of questions, which is unfair to the motive of various testing agencies conducting the exams. In order to overcome such difficulties, "Normalization process" is conducted based on Percentile Scores. The procedure of Normalization is a standardized process for comparing the marks obtained by the students across different sessions and the same practice is followed for a large number of exams throughout.

An apt example of how the normalization works can be seen in the Joint Entrance Examination (JEE) Mains conducted by National Testing Agency (NTA). As stated by NTA in its official website, the percentile score of a particular student is the percentage of students who have obtained marks equal to or less than the marks secured by the student [1]. CUET UG (Central University Entrance Test Undergraduate) has an entire section in its information bulletin published on its official website which explains this Equi-Percentile method in great detail. Then instead of using the raw scores, the percentile scores are used to rank the students, based on which the students are given entry into various colleges all across the country [2]. This kind of ranking scheme is also used for assessments carried out by Institute of Banking Personnel Selection (IBPS) for banking positions [3].

Although the procedure for formulation of the percentile scores has been revealed on various websites but it still stands a very complex procedure. The simple reason being that the raw scores and percentile data is not released by the testing agencies and there is no way for the students to get hold of their raw scores. Moreover, there are innumerable examinees taking such assessments and, unless it is dealt with professionally, it is very difficult for general public to handle such huge amount of data even if the data had been handy for us, which is again a hypothetical situation. Therefore, it is crucial to comprehend and analyze these normalization processes since they enable us to properly understand how educators can manage the student body and forecast student performance based on preparation level. This paper's analysis can also provide feedback on the degree of difficulty of examination papers and its quantification, helping those who are involved in setting question papers.

The motivation behind the work has been taken from the papers written by one of the authors of the present article, Sudipto Roy. In one of his articles, a straightforward mathematical model has been established that can be used to calculate students' percentile and, consequently, their rank based on the

marks/grades they receive in an exam [4]. The entire syllabus has been broken up into sections of comparable complexity and preparation time. As a result, the marks have been developed as a percentage function. The percentile has been provided as a probability function of the percentage of marks based on widely accepted observations. Given the significant role, played by negative marking in all these examinations, it has also been taken into account in our scheme. Additionally, after negative marking was added to the model, the percentile and ranks of the students were represented mathematically as the functions of marks they received.

In another paper by Sudipto Roy, he has considered the students' marks to be known, and has developed a method to estimate the percentile score [5]. Calculating the most probable marks and their associated probabilities has been made simpler by using a probability function $f(m)$. The likelihood of receiving marks within a certain required range (specific to examinations) may therefore be determined with the aid of a cumulative probability $g(m)$ representing the likelihood of obtaining marks greater than a certain cutoff set by an institution for its admission process. We will represent our model graphically by changing various parameters like difficulty level, etc. to understand the implications of these parameters on our model.

This study is driven by the need to develop a model which is based on statistical mechanics so that the principles from the same can be utilized to make more precise percentile estimation. This knowledge can be used to a more effective preparation strategy and help the academic instructors to ensure fairness. In order to do so, the Maxwell Boltzmann distribution can be used to study the distribution of marks, considering the similarity that can be drawn between the distribution of particles in different energy levels and the distribution of students across different levels/slabs of marks in the entire range; and between the students and the particles in an ensemble. In such a model, the percentile can be taken as a cumulative probability, and finally, ranks can also be expressed in terms of marks. Since NTA and other testing organizations do not provide any of their raw data, we have used the data readily available from the marks-rank-percentile statistics shown in the websites such as Careers360, Shiksha, and BYJUs [3, 12, 13]. This allows the graphs produced by our theoretical model to be compared with those based on the real-world data.

2. Model Formulation

In this model, the case of JEE Mains has been chosen for simplicity. The JEE Mains paper consists of 75 questions. Each correct answer fetches 4 marks and an incorrect answer results in a deduction of 1 mark for a student, thus the maximum obtainable marks is 300 and the minimum is −75. NTA releases an answer key which enables the students to estimate their raw scores but, since the percentile calculation is a very complex procedure, the students look into various websites to estimate their percentiles and ranks based on the scores. This paper focuses on estimating the percentile scores of students based on marks, as per the definition of percentile score by the NTA in its site. Moreover, a direct correlation can be established between the fraction of the syllabus completed and the estimated percentile score.

2.1 Syllabus Coverage

Let the entire syllabus for the examination be divided into N equal parts where equality of the parts is determined by different factors like importance, length, difficulty, etc. The probability that a student covers x parts of the syllabus is defined by a function $f(x)$ such that,

$$f(x) = \frac{a}{b^x} \tag{01}$$

where x is a discrete random variable that runs from 0 to N

Since $f(x)$ is a probability distribution function, it will always have value from 0 to 1 (both inclusive).

To meet this requirement, we must have, $0 < a < 1$, and, $b > 1$. The choice of the functional form of $f(x)$ is based on the observation that, as the number of syllabus potions increases, the chances of a student covering those portions should decrease. In other words, as x increases, $f(x)$ is likely to decrease.

Since the sum of probabilities for all possible values of x is 1,

$$\sum_{x=0}^{N} f(x) = 1 \tag{02}$$

Substituting from equation (01) into equation (02), we get,

$$\sum_{x=0}^{N} \frac{a}{b^x} = 1 \tag{03}$$

Equation (03) leads to the following value of a [6-8].

$$a = \frac{b^{N+1} - b^N}{b^{N+1} - 1} \tag{04}$$

As per equation (04), for an extremely large values of N, $a \approx 1 - \frac{1}{b}$

Using equation (04) in equation (01), we get,

$$f(x) = \frac{b^{N+1} - b^N}{b^{N+1} - 1} b^{-x} \tag{05}$$

where $x = 0,1,2, \ldots, N$

For larger values of b, $f(x)$ decreases faster as x increases.

2.2 Percentile Evaluation

Let the total number of students who have appeared for the examination be, Y.

Number of students who have prepared x out of N parts of the syllabus can be represented by $y(x)$ as,

$$y(x) = Y f(x) \quad (06)$$

Marks of a student can be considered as a function of portions of the syllabus covered. If other parameters like the performance of students on the day of exam, the time limit, the difficulty level of the exam are ignored, then we can say that, as more parts of syllabus is covered, more questions would be attempted and therefore the marks of the student would increase.
We have used the percentile definition from the NTA website which states that,

$$Percentile\ of\ a\ student = \frac{Number\ of\ students\ who\ have\ scored\ less\ than\ or\ equal\ to\ the\ marks\ obtained\ by\ the\ student}{Total\ number\ of\ students} \times 100$$

Based on this definition we can infer that the percentile of a student, who has covered x potions of the syllabus, can be taken to be the ratio of the number of students covering x, or less portions to the total number of students taking the examination. For simplicity, we consider marks (obtained by a student) to be proportional to the number of portions covered by the student.
Therefore, the percentile score P is,

$$P = \frac{\sum_0^x y(x)}{Y} \times 100 \quad (07)$$

Using equation (06) in equation (07), we get,

$$P = \frac{\sum_0^x Y f(x)}{Y} \times 100 = 100 \times \sum_0^x f(x) \quad (08)$$

Using equation (05) in equation (08), we get,

$$P = 100 \times \sum_0^x \frac{b^{N+1} - b^N}{b^{N+1} - 1} b^{-x} \quad (09)$$

Using the formula for sum of geometric series on the term b^{-x} [6-8], the expression for percentile (P) becomes,

$$P = 100 \times \left(\frac{b^{N+1} - b^N}{b^{N+1} - 1}\right)\left(\frac{b - b^{-x}}{b - 1}\right) \quad (10)$$

For large values of N, $P \approx 100 \times (1 - b^{-x-1})$ (11)

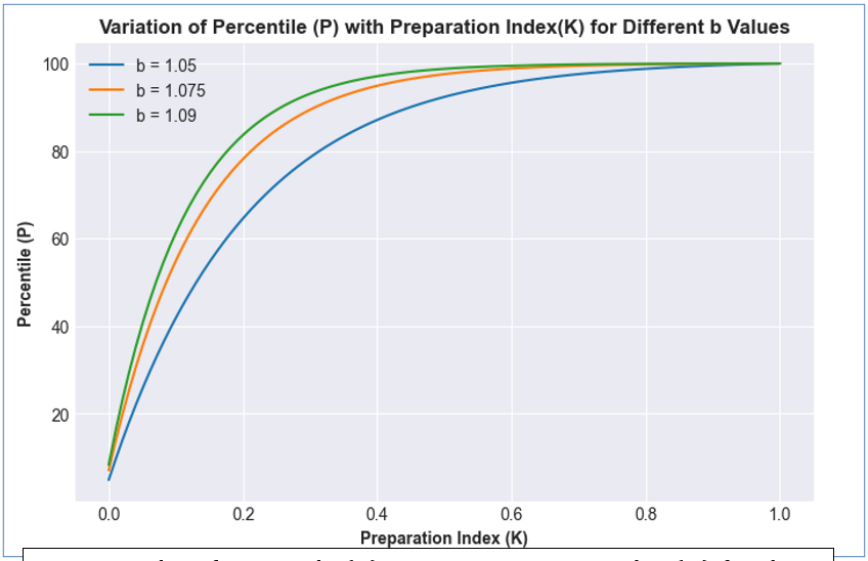

Figure 1: Plot of Percentile (P) versus Preparation Index (K) for three values of b, $N = 100$

2.3 Preparation Index

Let us now introduce another parameter, called the *preparation index K*, which can be defined as the ratio of the number of parts studied by a student (let x) to the total number of parts present in the syllabus, i.e., $K = \frac{x}{N}$
In terms of preparation index, we can rewrite equation (10) as,

$$P = 100 \times \left(\frac{b^{N+1} - b^{N}}{b^{N+1} - 1}\right)\left(\frac{b - b^{-KN}}{b - 1}\right) \quad (12)$$

Similarly for extremely large values of N when $b^{N+1} \gg 1$, we can write equation (12) as,

$$P \approx 100 \times (1 - b^{-KN-1}) \quad (13)$$

In Figure 1, we have plotted the percentile (P) of the student against its preparation index ($K = x/N$) using equation (12). As the preparation index increases, the percentile increases at a gradually slower rate. As b increases, the percentile rises faster with K, getting closer to 100 more rapidly.

2.4 Role of Negative Marking

In most of the competitive examinations that are held on an All-India basis, negative marking scheme has a major role to play. They penalize the students for every wrong answer thus discouraging students from guessing answers randomly.

Let the total number of questions be S. We know that K is the fraction of the syllabus covered by a student; thus, KS is can be regarded as approximately the number of questions attempted by the student ignoring other factors like the mental state of a student in the exam hall, the time constraint, and other complex parameters. Now, let us introduce another parameter g such that $1/g$ is the difficulty level of the exam. Easier the paper, the parameter g gets closer to 1. Therefore, we can write gKS be the total number of questions that has been attempted correctly by any student with $0 < g \leq 1$. The number of questions that has not been answered correctly by the student will then become $(1 - g)KS$. Let us assume that a student gets $+p$ for every correct answer and $-q$ for every wrong answer (with $p, q > 0$). Total marks (or $full\ marks$) of the paper is thus pS.

Marks (m) secured by a student with preparation index $K(= x/N)$ is therefore given by,

$$m = pgKS - q(1 - g)KS \quad (14)$$

Percentage of marks obtained by the student is,

$$M = 100 \times \frac{pgKS - q(1-g)KS}{pS} = 100 \times \mathrm{K}\left[g - \frac{q}{p}(1 - g)\right] \quad (15)$$

Using equation (15) we write,

$$K = \frac{M}{100 \times \left[g - \frac{q}{p}(1-g)\right]} \quad (16)$$

Using equation (16) in equation (12) we get,

$$P = 100 \times \left(\frac{b^{N+1} - b^{N}}{b^{N+1} - 1}\right)\left(\frac{1}{b-1}\right)\left(b - b^{-\frac{NM}{100 \times \left[g - \frac{q}{p}(1-g)\right]}}\right) \quad (17)$$

Equation (17) is an expression of percentile (P) in terms of the percentage of marks M.

We have plotted equation (17) in the following graphs which basically plots the percentile of a student against his percentage of marks. Thus, one can study the role of parameters like N, b and g, in governing the percentile score.

Figure 2 studies the effect of change of b. For higher values of b, the value of P increases faster as a function of M (marks percentage).

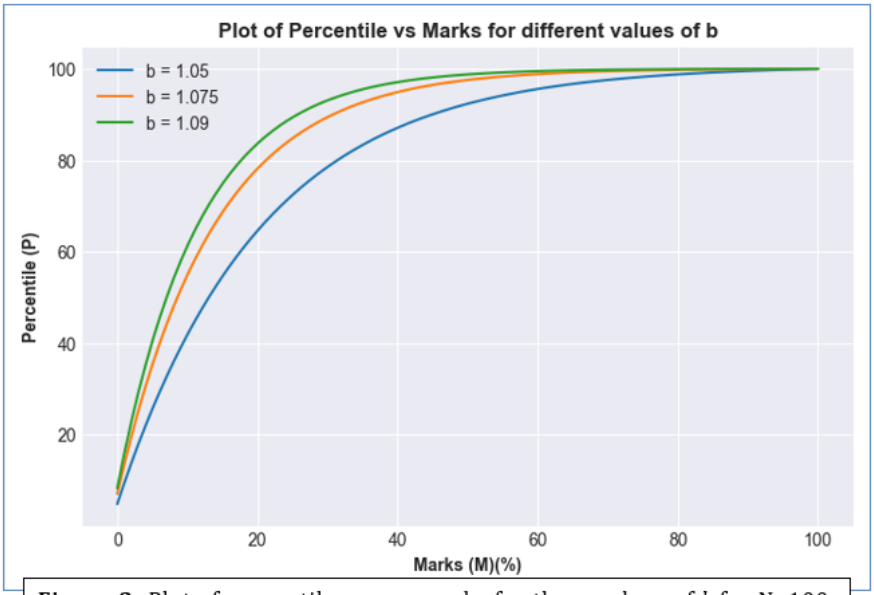

Figure 2: Plot of percentile versus marks for three values of b for N=100, g=1, p=4, q=1.

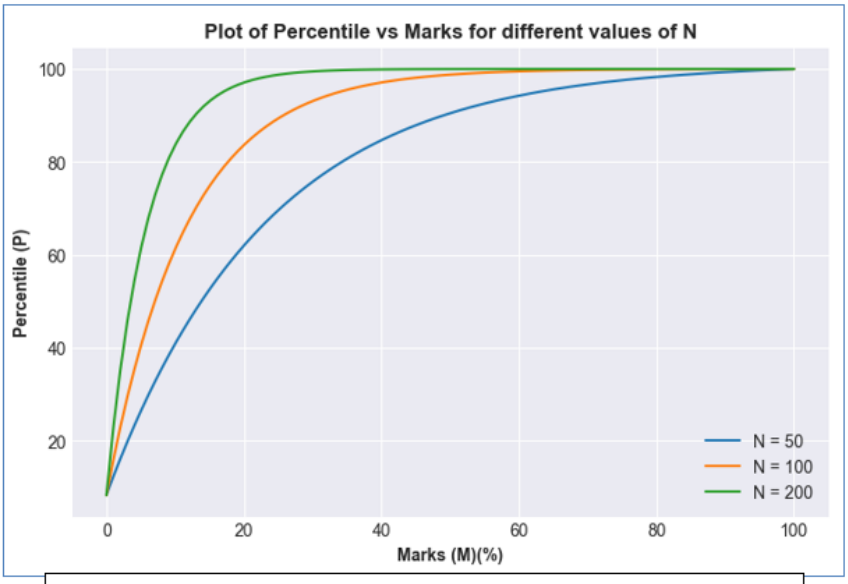

Figure 3: Plot of percentile (P) versus marks (M) for three values of N, b=1.075, p=4, q=1, g=1

Figure 3 highlights the plot of percentile versus marks percentage (M) for different values of N to understand the effect of change of N on the percentile trend. The percentile plots become steeper as N rises.

2.5 Rank Calculation

The entire scheme for percentile estimation has been modelled so as to understand how the ranking scheme of such examination works on the basis of which students get themselves enrolled in different institutes.
The number of candidates who have scored marks less than or equal to the student whose percentile score is P, can be written as $YP/100$. The number of students who have secured marks greater than the student is therefore, $Y - YP/100$.
Thus, expected rank (R) of the student is given by,

$$R = Y\left(1 - \frac{P}{100}\right) = \frac{Y}{100}(100 - P) \tag{18}$$

Using equation (17) in (18), we get,

$$R = Y\left(1 - \left(\frac{b^{N+1} - b^{N}}{b^{N+1} - 1}\right)\left(\frac{1}{b-1}\right)\left(b - b^{-\frac{NM}{100 \times \left[g - \frac{q}{p}(1-g)\right]}}\right)\right) \tag{19}$$

Equation (19) is an expression for rank of a student (R) in terms of the percentage of marks (M).
Instead of using rank directly for evaluation of the performance of students, we can take the factor R/Y as a more accurate parameter to assess the performance of the students. Since R/Y gives the achievement of the student relative to the other students, it is more meaningful to use this parameter instead of using the rank directly.

$$\frac{R}{Y} = \left(1 - \left(\frac{b^{N+1} - b^{N}}{b^{N+1} - 1}\right)\left(\frac{1}{b-1}\right)\left(b - b^{-\frac{NM}{100 \times \left[g - \frac{q}{p}(1-g)\right]}}\right)\right) \tag{20}$$

Figure 4 shows plots of R/Y as a function of marks percentage (M) for three values of N. As M increases, R/Y decreases for all N. As N increases, the value of N/Y falls faster with M.
Figure 5 shows plots of R/Y as a function of marks percentage (M) for three values of b. As M increases, R/Y decreases for all b. As b increases, the value of N/Y falls faster with M.

Using the expressions of the percentile and rank, we can calculate the unknown parameters in the model N, b, g based on the marks, rank and percentile that has been collected from the data of the previous examinations.
The mean value of x and x^2 are given by $\mu(x) = \sum_{x=o}^{N} xf(x)$ and $\mu(x^2) = \sum_{x=o}^{N} x^2 f(x)$. The standard deviation of x is given as $\sigma(x) = \sum_{x=0}^{N} f(x)(x-\mu)^2 = \sqrt{\mu(x^2) - [\mu(x)]^2}$[6-8].

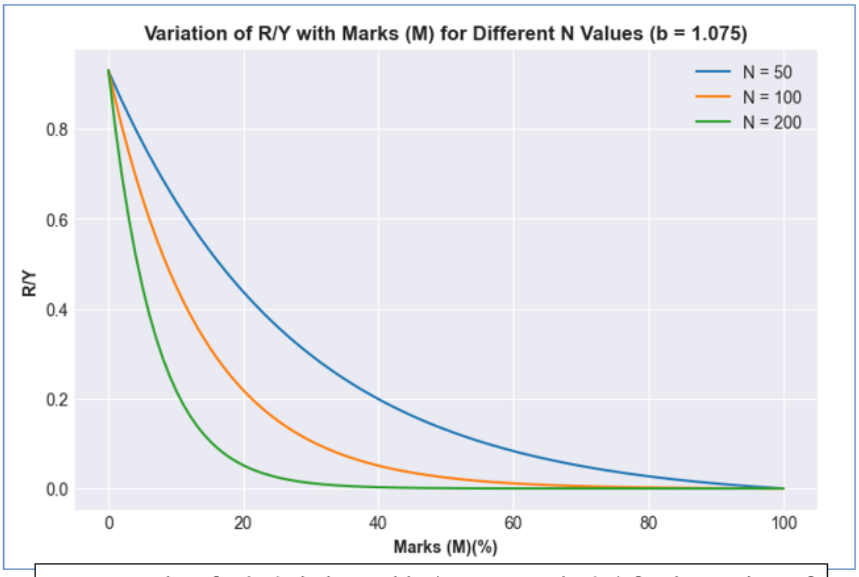

Figure 4: Plot of R/Y (relative ranking) versus Marks (M) for three values of N, b=1.075, p=4, q=1, g=1

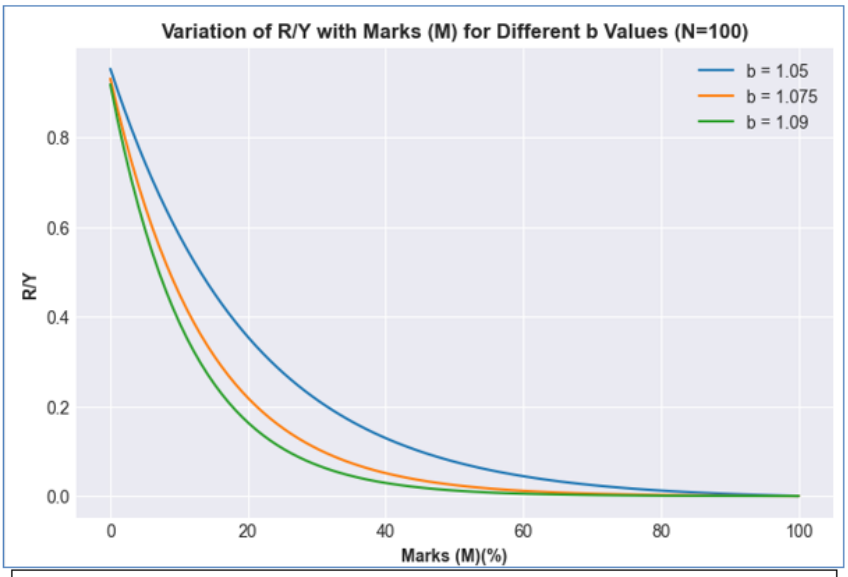

Figure 5: Plot of Rank/Total number of students (R/Y) versus Marks (M) for three different values of b, n=100, p=4, q=1, g=1

Based on equation (05), we can write,

$$\mu(x) = \left(\frac{b^{N+1}-b^{N}}{b^{N+1}-1}\right)\left(\frac{1}{b}+\frac{2}{b^{2}}+\frac{3}{b^{3}}+\ldots+\frac{N}{b^{N}}\right) \tag{21}$$

$$\sigma(x) = \left[\begin{array}{l}\left(\frac{b^{N+1}-b^{N}}{b^{N+1}-1}\right)\left(\frac{1}{b}+\frac{4}{b^{2}}+\frac{9}{b^{3}}+\cdots+\frac{N^{2}}{b^{N}}\right) \\ -\left\{\left(\frac{b^{N+1}-b^{N}}{b^{N+1}-1}\right)\left(\frac{1}{b}+\frac{4}{b^{2}}+\frac{9}{b^{3}}+\ldots+\frac{N^{2}}{b^{N}}\right)\right\}^{2}\end{array}\right]^{1/2} \tag{22}$$

Since we have preparation index $K = x/N$, we can write $\mu(K) = \mu(x)/N$ and $\sigma(K) = \sigma(x)/N$
The marks percentage is, $M = 100 \times K[g-(q/p)(1-g)]$
The mean and standard deviation of M, denoted by $\mu(M)$ and $\sigma(M)$, respectively, can be expressed as,

$$\mu(M) = 100 \times \mu(K)\left[g-\frac{q}{p}(1-g)\right] \tag{23}$$

$$\mu(M) = 100 \times \frac{\mu(x)}{N}\left[g-\frac{q}{p}(1-g)\right] \tag{24}$$

$$\sigma(M) = 100 \times (\sigma(x))/N\,[g - q/p\ (1-g)] \quad (25)$$

Thus, we have successfully established a model which can estimate the percentile and rank of the student if the marks are known. The marks can be calculated from the answer key that NTA releases before it declares its results. The unknown parameters associated with the model can be calculated if only enough data can be obtained which at the moment, is not available to us.

2.6 Calculations Based on Known Percentile

If we observe the general trend of competitive examinations, we will find that the percentile score increases with increase in the marks, and the rate of its change (with marks) decreases as marks increase. Using this observation, we propose, in this model, an empirical expression for Percentile Score P,

$$P(m) = a[1 - exp\{-b(m+c)\}]^n \quad (26)$$

where $a, b, c, n > 0$. Here, the symbols, a & b, denote parameters which are different from those denoted by a & b in the sections from 2.1 to 2.5.
The reason why such a function is chosen is because of the facts that we have observed, $P(m)$ increases with m, approaching the value of a asymptotically. For larger values of n, the rate of increase of $P(m)$, with m, decreases slowly.
Since for simplification we are specifically considering the case of JEE Mains, we can take the total marks of the entire paper as 300. The corresponding percentile $P(m)$ should be 100 (highest possible percentile score) for $m = 300$. Using this, we can write equation (26) as,
$100 = a[1 - exp\{-b(300+c)\}]^n$, which leads to,

$$a = \frac{100}{[1-exp\{-b(300+c)\}]^n} \quad (27)$$

Using equation (27) in equation (26), we get,

$$P(m) = \frac{100[1-exp\{-b(m+c)\}]^n}{[1-exp\{-b(300+c)\}]^n} \quad (28)$$

Nobody scores marks for which the percentile score is zero. The lowest marks that can be obtained is $m = -c + 1$, as zero percentile is obtained when $m = -c$ from equation (26), since marks increases by 1. Finding a certain combination of values for the constant parameters (b, c, and n) is necessary to guarantee the accuracy of the results produced by this formula (equation 28). While doing any analysis based on equation (28), one should use those values $m$that falls in the range given by, $-c + 1 < m \leq 300$, to avoid unacceptable results or behaviour for $P(m)$.
By definition, percentile score of a student is the percentage of students getting marks less than or equal to that of the student. Thus, the percentage of students getting exactly m marks will be calculated as $P(m) - P(m-1)$. Hence, the fraction of students getting exactly m marks can be expressed as:

$$F(m) = \frac{P(m) - P(m-1)}{100} \quad (29)$$

Using equation (28) in (29), we get,

$$F(m) = \frac{[1-exp\{-b(m+c)\}]^n - [1-exp\{-b(m-1+c)\}]^n}{[1-exp\{-b(300+c)\}]^n} \quad (30)$$

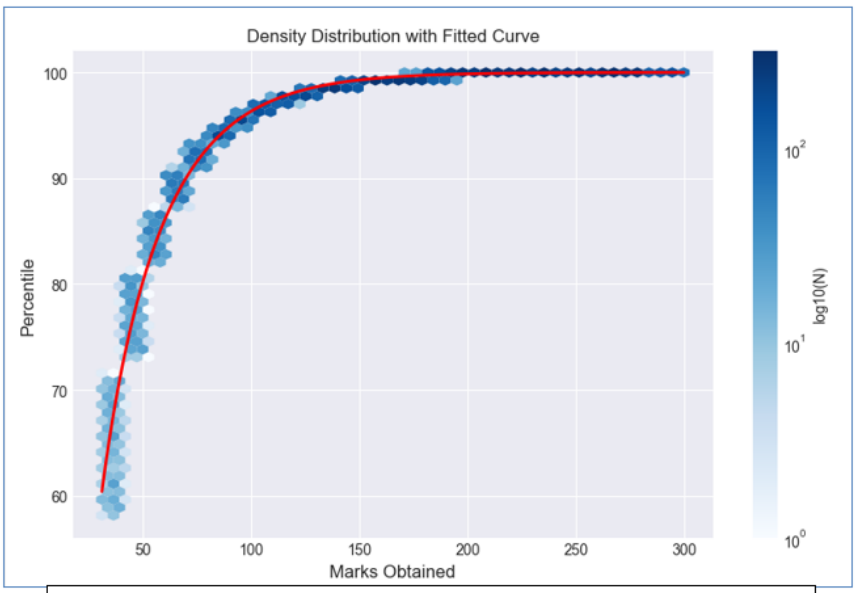

Figure 6: Plot of Percentile versus Marks for b = 0.032502, c = -13.530458, n = 1.602279. This graph highlights the real data as the blue points and the red line is the fitted curve

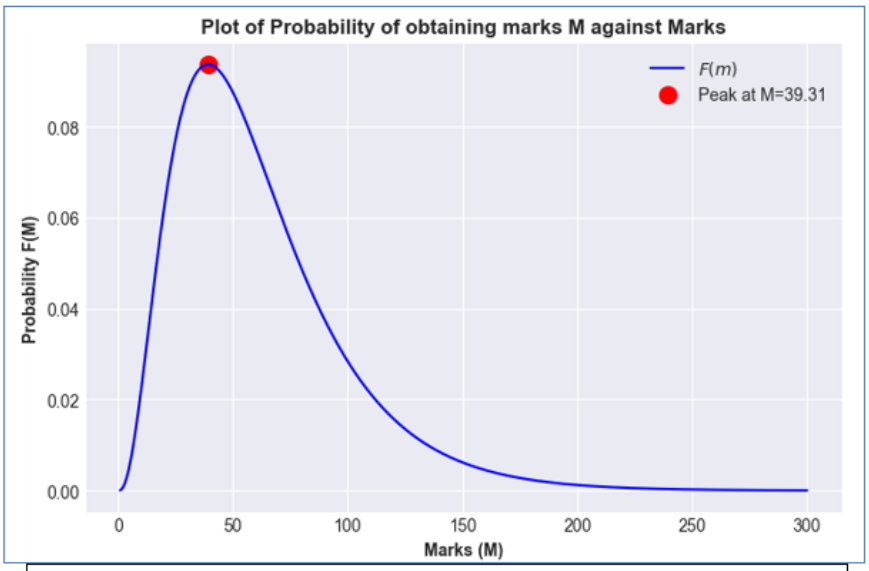

Figure 7: Plot of Probability of obtaining Marks F(m) versus Marks(m).

Figure 6 shows the percentile versus marks plot for the data that has been collected from real sources. The different shades of blue data points signify the density of percentile across the marks. This has been done because the data points are huge in number and, on plotting them, it becomes very dense to actually interpret. The red curve is the best fit curve of equation (28) to the data. The parameter values obtained by curve fitting are, $b = 0.032502$, $c = -13.530458$, $n = 1.602279$.

Figure 7 shows the probability of obtaining marks m versus marks as given by equation (30). This shows that most of the students score marks in this range whereas the graph towards the two extremes converges to zero showing that there is a very small percentage of students in the two extreme ranges. Thus, the graph follows an asymmetric bell-shaped curve which we will see as the kind of behavior for real data. The peak value of marks is at 39.3, which is the most probable marks. The number of students obtaining that score can be calculated as follows: if the number of students taking part in the exam is 10^6, then, based on $F(m = 39.31) = 0.093684$, number of students scoring 39.31 is $0.093684 \times 10^6 = 93{,}684$.

We can consider $F(m)$ as the probability of the students obtaining m marks because of the large sample size. Here, m can be taken as a random variable in the range $-c + 1 < m \leq 300$. By using equation (30), it is found that for $n \leq 1$, $F(m)$ is a monotonically decreasing function of m, with the lowest possible marks, $m = -c + 1$, corresponding to the highest value of $F(m)$. When $n > 1$, $F(m)$ peaks at a certain value of m, which could be thought of as the most likely score for the test takers. The lowest possible scores cannot actually be the most likely grades since no applicant is supposed to go into the test knowing nothing or feeling compelled to try questions for which they have no clue to the correct answer, which would result in a low score. Consequently, the range of values $n \leq 1$ is not acceptable for the parameter n. It is evident that as n rises, the most probable value of m increases.

3. Analogy with Statistical Mechanics

3.1 Portion of the Syllabus Covered

We may consider a system of students appearing for JEE Mains (for simplicity, we are specifically considering the case) to be compared to a system of particles. In that case the marks that can be obtained by the students within the range from -75 to 300 can be taken as a microstate. Then the rank and the percentile calculation can be considered to be macro-states since these properties define the entire system and can be only calculated if we take the entire system into consideration. Now since the students of the system are not interacting with each other we can consider the system to be made of ideal gas molecules instead of real gas. Using the property of ideal gas, when the molecules collide with each other, the collisions are purely elastic, i.e., the energy is conserved. In our model there is no interaction between the different systems.

Now let us assume that the entire syllabus is divided into equal parts (the equality being decided upon by factors like difficulty, number of questions in that part, time of preparation). Just like we discussed earlier, it is not possible for all students to complete the entire syllabus. Let us define a parameter $p(x)$ as the probability of students who can complete x parts of the syllabus. We can then write $p(x)$ as [9-11],

$$p(x) = \frac{1}{Z} e^{-\beta E(x)} \tag{31}$$

Here $E(x)$ is the *effort* energy required to prepare x portions of the syllabus. Since the effort energy increases with increase in the portions that have to be covered, we can say, by simple considerations, that, $E(x) = \alpha x$, where α is a positive constant. Thus, equation (31) becomes,

$$p(x) = \frac{1}{Z} e^{-\beta \alpha x} \tag{32}$$

The partition function is defined here as Z =$\sum_N e^{-\beta E(x)}$ [9-11]. This is similar to the partition function used by Maxwell Boltzmann in defining his model that explains the number of particles present in different microstates of same energy. Since $\sum_N p(x)$=1, we introduce the term$\sum_N e^{-\beta E(x)}$ which ensures this requirement. Thus $\sum_N e^{-\beta E(x)}$ helps us to normalize the function.

We also know that the difficulty of the paper is a very important parameter in determining the percentile of the students. Same amount of preparation can also lead to obtaining different percentiles considering the fact that the more difficult the exam is, lesser will be the number of students who can attempt greater number of sections of paper and hence, score higher marks. This is taken care of by the parameter β which stands for the *difficulty level* of examination. Greater the value of β, students are more likely to score lower marks as compared to for a lower value of β. Temperature plays an equivalent role in distributing the particles in different energy states. Lesser number of particles exists in higher energy states at higher temperatures.

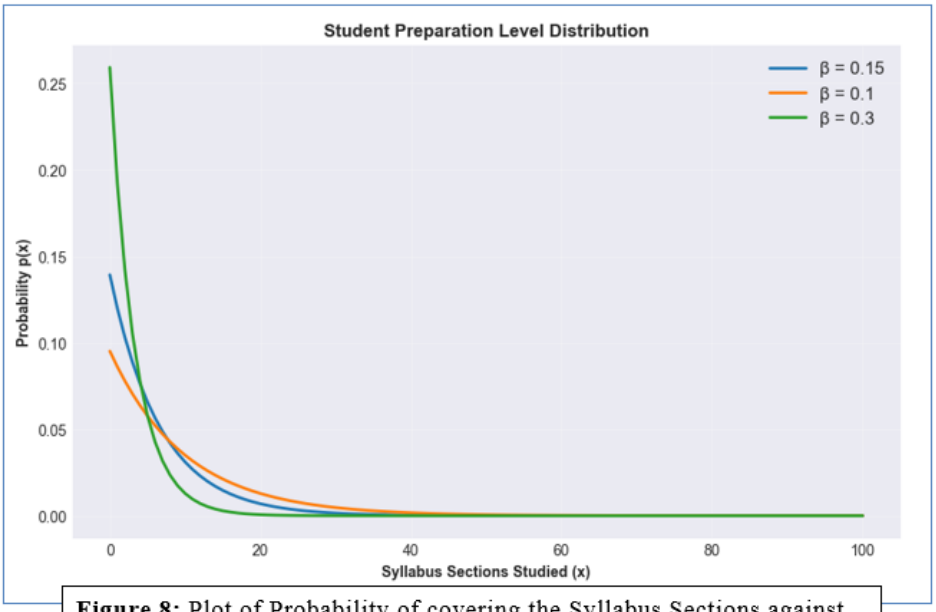

Figure 8: Plot of Probability of covering the Syllabus Sections against Syllabus Sections for different values of β

Figure 8 shows the probability of coverage of x parts of the syllabus, as a function of the variable x, based on equation (32). We analyze the effect of different values of β on the graph. As the value of β increases, there is a steeper fall of $p(x)$ with x.

Hence, students behave like particles and the amount of syllabus covered can be considered analogous to different energy levels. The distribution of students across the various values of *x* (portions of syllabus) is similar to that of the distribution of particles across different energy levels. Thus, the distribution function follows a Boltzmann curve where β is the *difficulty level* of the exam and has inverse relation to temperature. Just like only a very few particles occupy the higher energy levels, a very few students cover the entire syllabus. Most of the students cover a moderate amount of syllabus while few students cover almost nothing. This helps us to draw the similarity between the Boltzmann curve and the behavior of our model. For high β (the *difficulty level*), very few students cover the entire syllabus but for low value of β, students prepare syllabus more uniformly.

3.2 Marks Distribution

A student scores marks depending on his performance on the day of assessment. Assuming that the distribution is evenly distributed would mean that every student has received equal marks, which is not possible. The ones who have prepared seriously tend to obtain more marks than those who have prepared less extensively, provided that other factors like panic, confidence are not affecting their state to perform in the exams. Moreover, if we compare a real data graph [shown in Figure 9], we will find that the graph is bell-shaped and skewed with a long tail towards the high marks' region. Figure 9 depicts the probability $P(m)$ versus marks m based on the data for JEE (Mains) obtained from certain sources [12, 13]. This shows that number of students obtaining lesser marks is more prominent than the number of students who have obtained higher marks. There is a certain mark or several certain marks that will have definitely the most probability of occurrence. There is another fact that can be seen from the real data graph – the distribution does not start to increase directly at zero. There is a certain probability of obtaining negative marks but students do not obtain -75 in any of the data that we could get our hands on. This can be representative of the fact that neither any student goes unprepared to the exam nor attempts all questions wrong.

As the difficulty level of an examinations increases, fewer students score high marks but when the exam is easier, the distribution is more evenly spread meaning that there is more uniformity in the distribution of marks.

Keeping in mind the above observations, we define the probability of getting marks *M* as,

$$F(M) = \frac{|M-M_0|^{\frac{1}{2}} e^{-\beta_{neg}|M-M_0|}\theta(M_0-M) + |M-M_0|^{\frac{1}{2}} e^{-\beta_{pos}|M-M_0|}\theta(M-M_0)}{Z} \quad (33)$$

This function can be seen analogous to the Maxwell Boltzmann distribution of energy. The number of particles that can be distributed among different energy levels with energy E to $E + dE$can be considered to be directly proportional to $E^{1/2}e^{-E/kT}/z$.

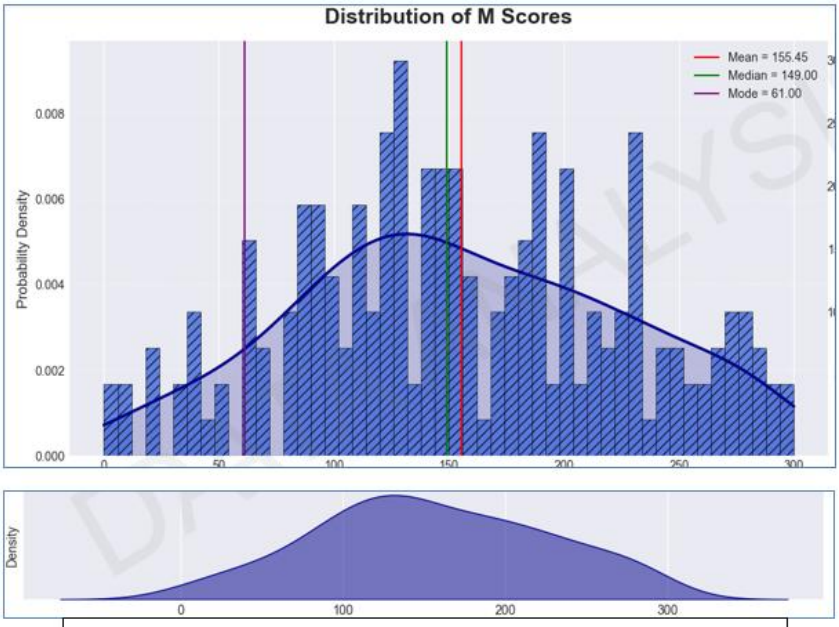

Figure 9: Marks Distribution (Probability of obtaining M marks) along with the display of mode, median and mean.

If we simplify equation (33), we will get,

$$F(M) = \frac{|M-M_0|^{1/2}e^{-\beta|M-M_0|}}{Z} \tag{34}$$

Equation (34) can be considered to be very similar to the Maxwell Boltzmann distribution of speed. From the plot of the real data (Figure 9), we clearly see the similarities between the two.

The reason why we have substituted energy in the Boltzmann energy distribution is because marks and energy seem to show analogous behavior and can also be shown so using the graphical representations of the two. The $|M-M_0|^{1/2}$ term has been considered as the correction factor and can be physically interpreted as the correction introduced for density of states. This is done so that the probability density does not increase linearly with marks for $M > 0$. This is reflective of the observations made in real life data:

- Low Marks: A small percentage of students receive extremely low (or negative) grades as a result of guesswork or incomplete information.
- Mid-Range Marks: Because they tackle questions for which they are only marginally prepared, a greater percentage of students receive mid-range marks.
- High Marks: Because it takes almost flawless preparation, fewer students receive extremely high grades.

Moreover, if we do not take the term $|M-M_0|^{1/2}$, then at M=0, the function would have a finite value, suggesting a non-negligible chance of scoring precisely zero. This is frequently unattainable. But since, we are ensuring that the term $M^{1/2}$ is being used in this function, the function becomes zero as the M $\rightarrow 0^+$, thus making a more realistic approach.

$|M-M_0|$ takes care of the negative marking because if we do not take the absolute value, considering the case of negative marking, we will get imaginary numbers which is not possible.

Here, Z is the partition function which can be written as:

$$Z=\sum_{M=-75}^{300}|M-M_0|^{1/2}e^{-\beta|M-M_0|} \tag{35}$$

Though according to the data collected no student can get the lowest possible marks, but the probability of getting negative marks is not negligible, this is the reason why instead of directly using $|M|^{1/2}$, we have introduced the term M_0 which ensures the shift in the probability distribution function. For values of M_0>0, the function will shift towards the right. Similarly, if we take the value of $M_0 < 0$ then the function will shift towards the negative region. But if want the function to start plotting at M=0, we take the value of $M_0 = 0$.

The factor β has been taken as $\beta_{pos}$and $\beta_{neg}$for the positive and negative regions respectively. This has been done to ensure the asymmetric decay in the positive and negative regions. The negative region will show a greater decay than the positive region and it goes to zero after a certain negative value. Thus, two different β factors have been considered.

Another important part of the function that has been mentioned in equation (33) is the Heaviside step function $\Theta(x)$. The Heaviside step function is a mathematical function used to switch the behavior of a function based on the value of x. Therefore, we can write,

$$\theta(x) = 1; x \geq 0 \; and \; \theta(x) = 0; x < 0 \tag{36}$$

Based on this definition we can say that $\Theta(M - M_0)$ ensures that only the first term of equation (33) is applicable for M $\geq M_0$ whereas $\Theta(M_0 - M)$ ensures only the first term of equation (33) is applicable for M $< M_0$. Thus, we are able to formulate the function that represents the real data with sufficient accuracy.

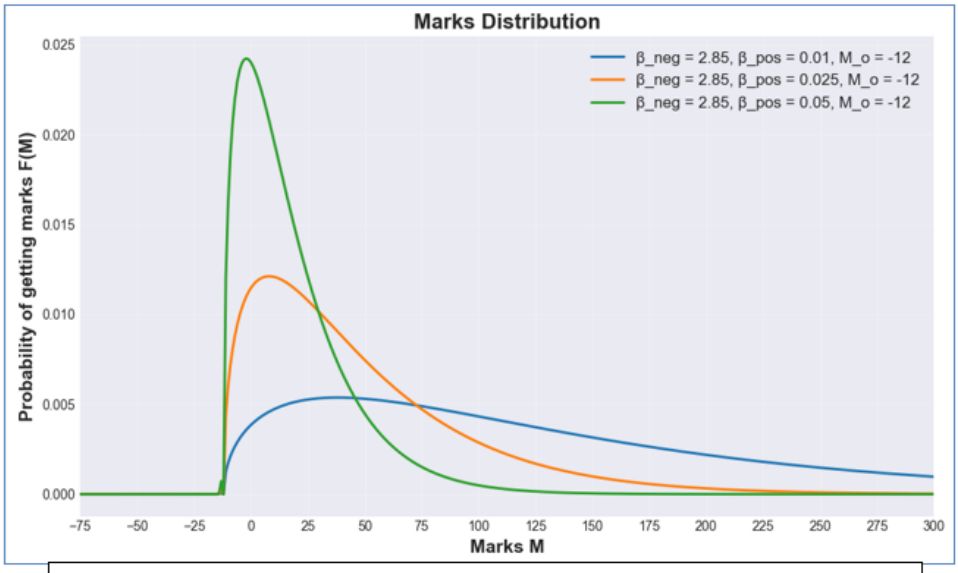

Figure 10: Marks distribution at M_0 =-12 and β_{neg}=2.85 but for different values of β_{pos}

Figure 10 has been plotted to understand the nature of marks distribution for different values of β. The real exam data (Figure 9) has been closely found to resemble the bell-shaped curve that we have obtained in this graph. Thus, Maxwell Boltzmann distribution law can be applied to real marks distribution very well especially for large scale exams. The Boltzmann factor $e^{-\beta|M-M_0|}$ is the one determining that at higher scores, the percentage of students will decrease. From the comparison of β, we can say that if β is greater, that is for difficult examination, the graph is narrower and a higher peak is achieved. This implies that for difficult exam very small number of students achieve good marks and the ones who are really well prepared gets higher scores. This behavior of marks distribution against various difficulty levels can be compared to the way particles behave at various temperatures. The probability of particles existing in higher energy states becomes less as we increase the temperature. The calibration of the y scale can be understood as follows: if 10 lakhs students are participating then 0.025 on the scale would mean 25,000.
Using the function of marks thus defined, we can calculate various aspects of the same. The average or expected marks can be calculated as follows:

$$\mu = \frac{1}{Z}\sum_{M=-75}^{300} M\,F(M) = \frac{1}{Z}\sum_{M=-75}^{300} M|M-M_0|^{\frac{1}{2}}e^{-\beta_{neg}|M-M_0|}\Theta(M_0 - M) + |M-M_0|^{\frac{1}{2}}e^{-\beta_{pos}|M-M_0|}\Theta(M - M_0) \quad (37)$$

The expectation value of marks of the students is very crucial because it helps us in understanding the overall performance of the students and thus, enables us to judge our performance relative to the accomplishments of all other students.
The expectation value of M and M^2 are $\frac{1}{Z}\sum_{M=-75}^{300} M|M-M_0|^{1/2}e^{-\beta|M-M_0|}$ and $\frac{1}{Z}\sum_{M=-75}^{300} M^2|M-M_0|^{1/2}e^{-\beta|M-M_0|}$ respectively.

3.3 Percentile Calculation

As we have already mentioned, the percentile of a particular student depends upon the ratio of the number of students who have scored marks less than or equal to that of the student to the total number of students who have appeared for the examination. According to this definition,

$$P(M) = \sum_{m'=-75}^{M} F(m') = \sum_{m'=-75}^{M} \frac{|M-M_0|^{1/2}e^{-\beta|M-M_0|}}{Z} \quad (38)$$

The above equation has been obtained after substituting the value of *F(m)* from equation (34). Taking the value of $Z=\sum_{M=-75}^{300}|M-M_0|^{1/2}e^{-\beta|M-M_0|}$, we get equation (36) as:

$$P(M) = \frac{\sum_{m'=-75}^{M}|M-M_0|^{1/2}e^{-\beta|M-M_0|}}{\sum_{m=-75}^{300}|M-M_0|^{1/2}e^{-\beta|M-M_0|}} \quad (39)$$

The percentile function here, is interpreted as a ratio of cumulative probability divided by the normalization factor.

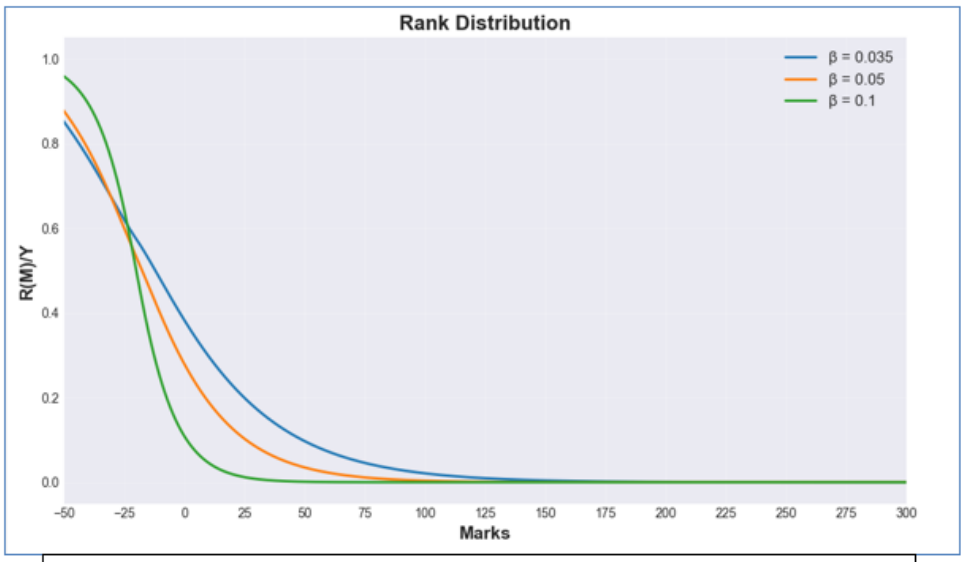

Figure 11: Plot of Relative Ranking R(m)/Y versus Marks (m) for different values of β, $M_0 = -12$

3.4 Rank Estimation

Let the rank of a student with marks M be $R(M)$, the percentile score be $P(M)$, and the total number of examinees be Y.
Rank of the student with marks M is given as:

$$R(M) = Y[1 - P(M)/100]) \tag{40}$$

Using equation (39) in (40), one gets,

$$\frac{R(M)}{Y} = 1 - \frac{\sum_{m'=-75}^{M} |M-M_0|^{\frac{1}{2}} e^{-\beta|M-M_0|}}{\sum_{m=-75}^{300} |M-M_0|^{\frac{1}{2}} e^{-\beta|M-M_0|}} / 100 \tag{41}$$

Figure 11 shows that, at smaller marks, the rank drops steeply. At the extreme values of marks, there is saturation achieved. The graph approaches 1 at low marks meaning most students are clustered at lower marks. There is a steep decline in the region of marks from -50 to 100, approximately. A very small change in percentile will cause a huge change in the rank if the student scores marks in that range. The curves approach very small values at higher marks, which means that very few students are able to reach that level.

4. Real Data Simulation

Though the Scores-versus-Percentile-versus-Rank data from the NTA would have been the most authentic information to carry out our study, but unfortunately, it does not make the results public. The data used in this analysis was taken from JEE Mains datasets available through some websites [12, 13]. The method of their collection of data is not known but it can be inferred that since they have access to the results of many students, they can create a general trend from those scores anonymously.
Figure 9 is based on this dataset. This figure shows the probability of students (along y axis) getting certain marks as a function of marks itself (along x axis). This gives us an insight to the various aspects of an examination model. As can be seen from the graph, the mode i.e., the most frequently occurring marks of the distribution is 61.00 while the average of the distribution is 155.45. This helps us in understanding the marks that the majority of the students obtain and teach the students accordingly. If we follow the general trend of the graph, then we can see that most of the marks are clustered in the mid-range, implying that maximum students score marks in that range. The graph tapers at the two extreme ends – lesser population in higher scores and similarly, a smaller number of people obtain very low scores. The graph does resemble a bell-shape which supports the theory that we have deduced on the basis of statistical mechanics.
Figure 12 shows the plot of percentile versus marks based on the dataset that has been obtained from internet [12, 13]. This plot is very similar to the ones that we have obtained from our models. At lower marks, the curve rises sharply, becoming less steep as marks increases. A slight rise in marks at the lower end causes large increase in percentile score.
Although there are many data points available, the study would have been more informative if we could properly distinguish between the different shifts of the examination so that the difficulty level parameter and other factors around it could be studied. The study cannot be properly made because of the lack of data available. Differences between the actual data and the theoretically generated data point to the presence of other variables that may be included in subsequent improvements of the model, including changes in paper difficulty-level, adaptive normalization, and trends in student preparation. This model could further be extended to other competitive exams and its dataset can also be studied in a similar way.

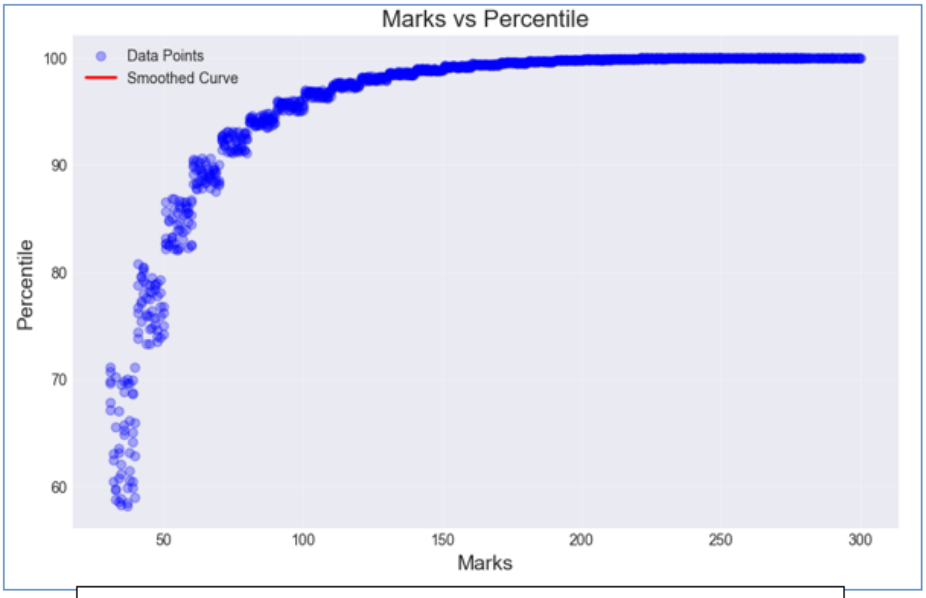

Figure 12: Plot of Percentile versus Marks for actual data.

5. Conclusion

From the above discussion, we can infer that the probability of a student scoring certain marks can be plotted using the Maxwell Boltzmann distribution. Since there can be a similarity drawn between the students in an examination and the number of particles in a system, we can define a function for examinees similar to the statistical distribution. This scheme can be very effective in predicting the percentile and the rank of a student and also their chances of getting admission to various courses, based on previous years' trends. This model also calculates the average preparation level and the most probable marks which will help the institutes to train students more effectively based on their relative stand in the crowd. The difficulty level parameter (β), which has been introduced here, helps the academic educators to analyze a paper and groom their students accordingly.

Because of the lack of availability of data, the model has several shortcomings. But the model could be made more efficient if the government sources (i.e., the examination organizing agencies) release their data. This model that has been formulated as a generalized theory, implying that it can be extended to other national and international level examinations like CUET UG, NEET, SAT, etc. Because of its universality, it can also be applied to various school and college examinations. If the teachers and the professors have access to the marks of the students across various examinations, then they can use them to determine the parameters associated with various functions defined in our models. The results obtained can be used to analyze the performance of the students depending on the examination difficulty level. The method of study, involving various graphs mentioned in this article, can prove to be of assistance all across the field of education, when it is applied to real data. This model could be used for a better analysis of performance by understanding the effect of other parameters like time constraint and psychological state of the student on the day of examination. In future, if the shortcomings of the model are overcome by its improvement, it would indeed be enormously useful to various sections of the society.

References

1. National Testing Agency. (n.d.). *Home*. Retrieved July 8, 2025, from https://nta.ac.in/Home
2. National Testing Agency. (n.d.). *Common University Entrance Test (CUET-UG)*. Retrieved July 8, 2025, from https://cuet.nta.nic.in/
3. Byju's Exam Prep. (n.d.). *IBPS equi-percentile method, IBPS marks calculation method*. Retrieved July 8, 2025, from https://byjus.com/bank-exam/ibps-equipercentile-method/
4. 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. https://doi.org/10.33545/26648636.2023.v5.i1a.48
5. Roy, S. (2024). Marks vs. percentile data of the JEE (Main): A study to assess the preparation level of the examinees. *World Journal of Advanced Research and Reviews, 23*(2), 735–743. https://doi.org/10.30574/wjarr.2024.23.2.2384
6. Boas, M. L. (2009). *Mathematical methods in the physical sciences* (3rd ed.). Wiley-India.
7. Arfken, G. B., & Weber, H. J. (1995). *Mathematical methods for physicists* (4th ed.). Academic Press.
8. Riley, K. F., Hobson, M. P., & Bence, S. J. (2006). *Mathematical methods for physics and engineering* (3rd ed.). Cambridge University Press.
9. Gupta, A. B., & Roy, H. P. (2020). *Thermal physics* (5th ed.). Books and Allied (P) Ltd.
10. Reif, F. (2009). *Fundamentals of statistical and thermal physics*. Waveland Press.

11. Sears, F. W., & Salinger, G. L. (2013). *Thermodynamics, kinetic theory, and statistical thermodynamics* (3rd ed.). Narosa Publishing House.
12. Careers360. (n.d.). *Home*. Retrieved July 8, 2025, from https://engineering.careers360.com/articles/jee-main-marks-vs-percentile
13. Allen. (n.d.). *Home*. Retrieved July 8, 2025, from https://allen.in/jee-main/marks-vs-rank