Correcting Percentile Rank Formula:
A Brief Revision of the property 'Reversibility'

Nava Teja

Mushamnavam9530@gmail.com

Abstract

This paper gives a detailed description of a hidden little flaw in the formula to calculate the Percentile Rank. This paper is written in a problem-solver's perspective rather than in a replicator's approach. Along with the discovery of error, two possible solutions of the same are also provided after comprehensive research on both the errors themselves and their root causes. Although the flaw itself does not have significant impact on the result yet it is worth correcting the formula as it went unnoticed for many decades by so many mathematicians, statisticians, and data scientists.

1. Introduction

The foundation for this theoretical research has been led by a simple property of a function called 'Reversibility.' Although, both the error in the Percentile Rank[1] Formula and the two solutions provided show less impact or significance in the data science field, the major role is played by the property 'Reversibility' itself. Later, in the paper, you will find how this simple property detected the mistake in the formula and helped to correct the same.

[1] Percentile Rank is often abbreviated to PR. To avoid unnecessary confusion, this paper refers to it in its extended form only.

Before moving further into the details of the formula for calculating Percentile Rank (or simply called Rank sometimes), it is important to know the significance of the Percentile and the origin of its formula[2]. Knowing the differences between the Percentage and the Percentile clarifies the doubts regarding the need of Percentile as well as the limitation of the Percentage.

2. Limitation of the Percentage

The Percentage (derived from Latin words 'per centum' which mean 'by hundred') is used to change the denominator of a fraction, either proper or improper to 100 so that it will be easier to compare two or more numbers by just comparing their numerators only.

This can be achieved by simply multiplying and dividing the decimal or fraction by 100. In other words, 'a Percentage is a fraction multiplied by 100 on both the numerator and the denominator.' Therefore, the formula becomes:

$$Percentage = \frac{Numerator \times 100}{Denominator \times 100}$$

$$\therefore Percentage = \left(\frac{Numerator \times 100}{Denominaotr}\right) \%$$

As we know, the Quantitative (Numerical) Data is categorized into 2 types namely:

- Continuous Data
- Discrete Data

The important point to be noted here is that a fraction or a decimal is used to represent a number in the range of continuous data only. They cannot be used to denote the numbers in the range of discrete data which will ultimately affect the Percentage concept to be limited to the continuous data itself. Thus created the need of a new concept dealing with the discrete data.

[2] The Percentile formula and the Percentile Rank formula are not same and are discussed later in the paper in detail.

3. Origin of Percentile

To overcome the limitation of the Percentage and the need for denoting the number in the range of discrete data invented the Percentile (derived form 'Percentage' and coined by Francis Galton in 1885). To calculate the Percentile, one must order the discrete data in ascending fashion. The number denoting the specific value or data point, say X, which is called as 'Rank' or 'Percentile Rank,' is used to specify the relative position of the data point in the data set. The number of data points before X is divided by the total number of data points leaving with a fraction, whose denominator is then converted to 100 just like in the case of Percentage. This leaves us with the formula:

$$Percentile = \frac{Number\ of\ values\ existing\ before\ X\ \times 100}{Total\ number\ of\ values\ \times 100}$$

$$\therefore Percentile = \left(\frac{Number\ of\ values\ existing\ before\ X\ \times 100}{Total\ number\ of\ values}\right)\ \%$$

The important point to be noted here is that unlike Percentage ($\in W$), Percentile exists in the range $[0, 1]$ because the numerator can never exceed the denominator. This logic is essential as it will help in the later part in the paper.

4. Percentile Function is Reversible

Unlike many cryptographic hash functions, which are irreversible, the Percentile function is a reversible function. That means, a Rank can be converted into a Percentile and the Percentile can be converted back to the Percentile Rank. The formula for calculating the corresponding Percentile Rank is given by:

$$Percentile\ Rank = Percentile \times\ (n + 1)$$

Here the points to be noted are:

- 'n' is the Total number of values.
- 'Percentile' implicitly includes the '%' or '100 in the denominator.'
- As the formula returns a decimal, it is rounded off to the nearest integer as the Rank $\in$ N.

But the problem here is that the formula does not seem to be derived directly from the original formula of the Percentile. This statement comes straight from the fact that the 'n + 1' term in the right-hand side of the equation is never mentioned in the original Percentile formula.

5. Given Percentile Rank Formula is incorrect

To verify the truth value of the formula, let us consider a dataset ranging from 1 to 500 (both inclusive). And let us convert Ranks of some random numbers[3] to their corresponding Percentiles and vice-versa. This idea should work because the function is reversible and should result same output.

Case-1: When Percentile is above 50

Let the data point be 301.

The Percentile of 301 in the dataset would be:

$$Percentile = \left(\frac{Number\ of\ values\ existing\ before\ 301\ \times 100}{Total\ number\ of\ values}\right)\ \%$$

$$Percentile = \left(\frac{300\ \times 100}{500}\right)\ \%$$

$$\therefore Percentile = 60\%$$

To verify the Percentile Rank formula, let us substitute 60% in it.

$$Percentile\ Rank = Percentile\ \times (n + 1)$$

$$Percentile\ Rank = \left(\frac{60}{100}\right) \times (500 + 1)$$

$$\therefore Percentile\ Rank = 300.6\ \approx 301$$

In this case, the formula worked charmingly.

---

[3] As the dataset is starting from 1, and their ranks start from 1, we can safely use them interchangeably.

Case-2: When Percentile is below 50

Let the data point be 201.

The Percentile of 201 in the dataset would be:

$$Percentile = \left(\frac{Number\ of\ values\ existing\ before\ 201\ \times 100}{Total\ number\ of\ values}\right)\ \%$$

$$Percentile = \left(\frac{200\ \times 100}{500}\right)\ \%$$

$$\therefore Percentile = 40\%$$

To verify the Percentile Rank formula again, let us substitute 40% in it.

$$Percentile\ Rank = Percentile\ \times (n + 1)$$

$$Percentile\ Rank = \left(\frac{40}{100}\right) \times (500 + 1)$$

$$\therefore Percentile\ Rank = 200.4\ \approx 200$$

In this case, the formula turned out be incorrect. Although the error is just 1 data value[4], it should not be possible for a reversible function. If you repeat the procedure for every data point in the dataset, you will realise that the formula withstands for only half of the values or in other words Percentiles in the range (50, 100].[5]

Not only in this example of dataset, but for any dataset range, the formula works only half-way. This directly suggests that the flaw has nothing to do with either the examples or the dataset, but with the formula itself.

[4] As we are rounding off to the nearest integer, the maximum error is 1 data value, otherwise it would be 0.5 data value. This will help in the later part of the paper.

[5] The 50th Percentile can satisfy the formula if the Rank is an odd number and cannot in case of an even Rank.

6. Cause of the Flaw

Using afore mentioned some important concepts, let us examine the root cause of the flaw.

$$Percentile = Percentile \times (n + 1)$$

By distributive property, we can write the equation as below:

$$Percentile\ Rank = (n \times Percentile) + (Percentile)$$

In the right-hand side of the equation, there are 2 terms. Let us call them 1st term and 2nd term respectively. If you recall the original Percentile formula, you will realize that there is no possible error in the 1st term. And if you remember that the Percentile ranges from 0 to 1, the 2nd term appears to be in the range [0, 1].

This should make sense because the part where we round off the Rank comes from the 2nd term. As we are facing the problem only when we round off, we can safely conclude that the flaw exists in the 2nd term and it is causing the error in the formula.

In other words, if the value of the 2nd term is:

- Greater than 0.5, then we round it off to 1.
- Less than 0.5, then we round it off to 0.
- Equal to 0.5, then we round it off to:
    - 1 if 1st term is odd.
    - 0 if 1st term is even.

7. Correct Formula of Percentile Rank

1st Solution: Using Least Integer Function

As you have already figured out, the common solution for this kind of problem is that 'Instead of rounding off to the nearest integer, round if off to the least integer greater than itself.' In other words, use Least Integer Function or Ceiling Function. Then the corrected formula looks like:

$$\boxed{Percentile\ Rank = \lceil\ Percentile \times (n + 1)\ \rceil}$$

2nd Solution: Actual Formula

This is the true formula if derived straight from the original Percentile formula.

$$Percentile = (n \times Percentile) + 1$$

8. Derivation of the Actual Formula

The key role played in this solution is afore mentioned function property called 'Reversibility,' As we discussed earlier, the Percentile is a Reversible Function. That means, by rearranging the terms, we can get the formula we desired.

$$Percentile = \frac{Number\ of\ values\ existing\ before\ Rank}{Total\ number\ of\ values\ (n)} \times 100\%$$

We know that,

$$Number\ of\ values\ existing\ before\ Rank = Rank - 1.$$

By substituting it in the formula, we get:

$$Percentile = \frac{Rank - 1}{n} \times 100\%$$

As 100% is equal to 1, let us eliminate it for simplicity.

$$Percentile = \frac{Rank - 1}{n}$$

$$n \times Percentile = Rank - 1$$

$$\Rightarrow Rank = n \times Percentile + 1$$

$$\therefore Rank = (Percentile \times n) + 1$$

9. Actual Reason for the flaw

By now, you might have realized that the 2nd solution is almost like the formula we dealt with.

$$Percentile\ Rank = Percentile \times (n+1) \longrightarrow \text{Incorrect Formula}$$

$$Percentile\ Rank = (Percentile \times n) + 1 \longrightarrow \text{Derived Formula}$$

This is very disappointing as the flaw was originated due to the misplacement of the parentheses. You might be thinking that this was a simple mistake. But remember, the fact of simplicity or complexity will be known only after research!

10.Conclusion

The problems we worked upon explicitly are:

- Finding that the Percentile Rank formula is incorrect.
- Finding the possible solutions for the same.
- Finding that the flaw was originated by a typo.

But the concepts we implicitly dealt with are:

- Distributive Property.
- Reversibility Property.
- Limiting Range of a Percentile.
- Rounding off to which value.

Without these concepts, it would be impossible to conduct such a disappointing yet conclusive research.