# Artificial Intelligence and Machine Learning with Moment Generating Functions to Enhance Biological Count Data Analysis

PAUL D. GLENN II[1], NANCY L. GLENN GRIESINGER[2], DEMETRIOS KAZAKOS[3]
[1]Department of Animal Science, McGill University, Montreal, Quebec, CANADA
[2]Department of Mathematical Sciences, Texas Southern University, Houston, Texas, U.S.A
[3]Department of Mathematical Sciences, Texas Southern University, Houston, Texas, U.S.A

*Abstract:* - This research presents artificial intelligence with supervised machine learning to determine the median count given training data from biological laboratory experimentation. After determining the central moment of the median, machine learning is then used to predict the population median value. When computer programs learn they access available data, then autonomously analyze the information to make informed, data-driven decisions. Moment generating function restrictions identify a parameter of interest by restricting the expected value of the moment generating function. This research proposes a theoretical foundation for achieving such predictions. A summary of the key elements underlying the statistical power of the Mann– Whitney test, a nonparametric hypothesis test for the median of count data, is also presented. A nonparametric approach allows for accurate predictions while relaxing distributional assumptions such as normality.

*Key-words:* - Artificial Intelligence, Machine Learning, Moment Generating Function, Nonparametric Statistics, Mann-Whitney and Statistical Power.

Received: March 12, 2024. Revised: Agust 4, 2024. Accepted: September 9, 2024. Published: October 8, 2024.

## 1. Introduction

The work of biologists extends beyond experimental design and conducting laboratory experiments. It also involves data collection, statistical analysis, and computing. Selecting appropriate statistical tests for any given data set can be challenging. Using the wrong statistical test not only jeopardizes the experiment's validity, but also increases the likelihood of drawing incorrect conclusions. In biological research, the misuse of parametric methods, such as the Student's t–test and analysis of variance (ANOVA), is quite common [1], [2], [3], [4].

In biology, laboratory experiments often yield statistical data types that describe countable quantities that take on only nonnegative integers or counting numbers [5]. These data are known as count data. Parametric methods are often misapplied to count data, which are not continuous data. Obtaining any data from biological laboratory experimental results often takes month or years. The inappropriate use of statistical tests greatly increases the chances of both type 1 and type 2 errors, thus rendering the painstaking biological research process futile.

Type 1 error refers to obtaining a false positive result, while type 2 error refers to obtaining a false negative result [6]. Statistical tests can be broadly classified into parametric and nonparametric tests, and the choice of test depends on the assumptions made about the data set [7]. For a parametric test such as ANOVA, six assumptions must be met: normal distribution, approximately equal variances, two or more independent groups, independent observations, continuous data, and the absence of significant outliers. Nonparametric tests are more flexible by relaxing distributional assumptions. Furthermore, nonparametric tests can be employed for ordinal and count data. The Kruskal–Wallis test serves as a nonparametric equivalent of ANOVA and the Mann-Whitney test as a nonparametric alternative to the Student's t–test. Of note, ANOVA is a generalization of the two-sample Student's t–test.

To compare sample sizes needed for comparable levels of statistical power when the Student's t–test is inappropriately applied to count data, Monte–Carlo simulations are carried out comparing the statistical power of the Student's t–test to the Mann–Whitney test. Power is a function of level of significance, effect size, and sample size. Hence, one can increase power by increasing sample size. As this research illustrates, increasing the sample size in biological experiments is often a painstaking process. We instead propose obtaining an initial sample from biological experiments as training data, then employing artificial intelligence and machine learning to predict the median counts based on the initial training data obtained from the laboratory.

This research is based on a theoretical and methodological foundation for determining the central moment of the median. Artificial intelligence and machine learning are then used to predict the population median value. Next, estimating conditional moment generating function restrictions is then used to learn with moment restrictions. When computer programs learn they access available data, then autonomously analyze the information to make informed, data-driven decisions. Moment generating function restrictions identify a parameter of interest by restricting the expected value of moment generating functions.

# 2. Biological Experiments and Count Data

In the realm of biological research, the ability to analyze count data is a crucial skill that can help researchers uncover insightful findings beyond experiments. Count data, a type of data which has a one-to-one correspondence with the set of nonnegative integers, are generated by a variety of biological experiments. Its analysis requires specialized statistical techniques.

Count data analysis necessitates a thorough understanding of statistical methods, alongside a deep comprehension of underlying biological principles. It is therefore essential to approach the analysis of count data with rigor, precision, and accuracy. By doing so, researchers can ensure that their findings are not only reliable but also contribute to the advancement of scientific knowledge. This section details procedures for obtaining data from biological experimentation which often takes month or years.

## 2.1 Preparatory Biological Procedures

The preparation processes in biological data generation can be demanding, often taking longer than the actual experiment itself. Comparing the number of neurons serves as a prime example. In the study by Passini, et al., there were multiple steps, including oligonucleotide synthesis, animal breeding, dissection, and microscopy [8]. Oligonucleotide synthesis is an intricate chemical process that encompasses the creation of short RNA or DNA sequences. In their study, the researchers designed an oligonucleotide, ASO-10-27, to target a specific gene or mRNA and evaluated its effectiveness as a treatment for spinal muscular atrophy (SMA) by testing its effectiveness on mice experiencing SMA.

Animal procedures were conducted using a mouse model of SMA. The mice strains used were bred in the lab by mating parents with specific genotypes to produce neonatal mice with the desired genotypes. This experimental procedure highlights the dedication of researchers in their quest for scientific knowledge and demonstrates the comprehensive techniques employed to generate data.

## 2.2 Count Data from Biological Experiments

After successfully producing the oligonucleotide treatment and generating the necessary mice strains, the mice were administered the oligonucleotide treatment shortly after being born. They were then placed in a carefully controlled environment where their health was regularly monitored, and specific dietary and environmental conditions were maintained. To further study the effects of the treatment, the mice were ultimately euthanized, and their spinal cords were carefully extracted. The spinal cords were then meticulously sliced at specific intervals to isolate tissue slides representing the cervical, thoracic, and lumbar spinal regions.

These tissue sections were then stained using Anti-Choline Acetyltransferase (ChAT) immune-staining, allowing for visualization and quantification of neurons through microscopy. The resulting neuronal counts represented valuable count data for determining efficacy of the oligonucleotide treatment. To establish a basis for comparison, several groups of untreated mice underwent the same procedures, enabling a comprehensive analysis of the outcomes between the treated and untreated groups. Since the outcomes are count data, a nonparametric test such as the Mann–Whitney test is a correct statistical test for analysis, not a t–test or ANOVA.

# 3. Results

This section presents the statistical power results when comparing the Student's t–test to the Mann-Whitney test when analyzing count data such as the data obtained from experiments detailed in the study by Passini, et al. [8].

## 3.1 Analysis of Variance and Student's t–Test

The one–way analysis of variance (ANOVA) is a widely used statistical hypothesis testing procedure for comparing three or more population means. ANOVA is a generalization of the two–sample, independent samples Student's t–test which compares the means of two populations.

Both ANOVA and the Student's t–test are parameteric tests that require assumptions to be satisfied. One such assumption is that the population is normally distributed and that that the data are continuous. When these as well as other assumptions are not satisfied, a proper alternative is to employ a non–parametric hypothesis test such as the Mann–Whitney test.

## 3.2 Mann–Whitney Alternative to Student's t–Test

The Mann-Whitney test is a nonparametric alternative to a Student's t–test [9]. While the Student's t–test compares the population means, the Mann–Whitney test compares another measure of central tendency, the population median.

The advantage of a Mann-Whitney test is that it relaxes assumptions such as normality, and it does not require the data to be continuous. Hence, it can be utilized with count data.

## 3.3 Statistical Power Results

To investigate the effects of the incorrect application of the Student's t–test when an appropriate test is a Mann–Whitney, this research carried out Monte-Carlo simulations using the statistical package R [10] to perform an empirical power analysis using the following algorithm:

```
Step 1:  Choose specific distribution of data
Step 2:  Determine level of significance $\alpha$, effect size
    Step 3:  Simulate data: $n_1 = n_2 = n$; $H_a$ true
    Step 4:  Apply hypothesis test to data
    Step 5:  Store result of hypothesis test
Step 6:  Repeat Steps 3 to 5 for $100,000$ simulations
Step 7:  Calculate empirical power
```

The empirical power is determined by dividing the number of times $H_0$ rejected by the number of simulations.

Corresponding minimum sample size determined from Monte Carlo simulation results for levels of power $0.70$, $0.80$ and $0.90$ are contained in Table I.

TABLE I
MINIMUM SAMPLE SIZE $n_i$ FOR STATISTICAL POWER $i$

| **Hypothesis Test**[a] | **Statistical Power** | | |
|---|---|---|---|
| | ***0.70*** | ***0.80*** | ***0.90*** |
| Student's t–Test | $n_{0.70} = 1775$ | $n_{0.80} = 2300$ | $n_{0.90} = 3050$ |
| Mann–Whitney | $n_{0.70} = 3$ | $n_{0.80} = 4$ | $n_{0.90} = 5$ |

[a] Based on level of significance $\alpha = 0.05$.

# 4. Artificial Intelligence and Statistical Moment Generating Functions

This section presents a machine learning approach to predict the median count based on training laboratory data. The proposed approach is aimed at providing accurate and reliable predictions, which can be of great value in various applications.

## 4.1 Expressing Median as Moments of a Distribution

In the field of statistics, a random variable is a function whose domain is the sample space and range is the real line. The distribution of the random variable $X$ tells which values the random variable takes on and how often it takes on these values. One way to uniquely characterize the distribution of a random variable is through a moment generating function, if the function exists. The moment generating function of a random variable can be expressed as the expected value,

$$M_X(t) = E[e^{tX}], \tag{1}$$

for some positive number $h$ such that $-h < t < h$. Furthermore, the mean, variance and median may be expressed in terms of a moment generating function, respectively:

- Mean is first moment of $X$

$$M'_X(0) = E[X] = \mu \tag{2}$$

- Variance is second central moment of $X$

$$M''_X(0) - [M'_X(0)]^2 \tag{3}$$

The median can be characterized in terms of moment generating function as well [11]. If $\tilde{x}$ indicates the sample median and $h(x)$ is the asymptotic distribution of $\tilde{x}$, then the $k$ moments $\tilde{\mu}_k$ of $\tilde{x}$ are defined as

$$\tilde{\mu}_k = E_h(x - \tilde{\mu}_1)^k. \tag{4}$$

The connections among the moments of the median, machine learning, and artificial intelligence is that robust machine learning can be expressed as conditional moment restrictions that constrain the conditional expectations of the moment generating function [12].

## 4.2 Artificial Intelligence Sub-field of Machine Learning

A sub-field of artificial intelligence is machine learning. In supervised machine learning, algorithms learn from training data to predict responses when presented with new data. In the case of the median, one can start with real count data obtained from biological experiments in the laboratory from which one would compute the sample medians. This would serve as the training data.

Using the training data, the next step is to constrain the conditional expectations of the moment generating function for the median. Finally, use the training data and robust machine learning to directly predict the median count values.

# 5. Discussion and Conclusions

Biologists employ a wide range of biological procedures to obtain experimental results. Before conducting the decisive experiment that tests their hypothesis, they often devise treatments and modify their subjects. Nonetheless, executing the decisive experiment can be particularly complex, especially when measuring specific variables. In order to generate accurate data, biological researchers must be committed to utilizing appropriate statistical techniques. It is crucial for researchers to carefully consider the characteristics and properties of their data when analyzing results, as applying inappropriate statistical methods may lead to erroneous conclusions.

We presented the theoretical and methodological foundation for using artificial intelligence and machine learning to predict the median count. This research provides the functional formulation of conditional moment restrictions. In further analysis this work will be extended where medians obtained from simulated data will be compared to AI aided medians to conduct statistical root mean squared analysis to evaluate the quality of AI predictions.

## *References*

[1] Kim A., Mok B.R., Hahn S., Yoo J, Kim D.H., and Kim T.A. Alternative splicing variant of NRP/B promotes tumorigenesis of gastric cancer. BMB Rep. 2022 Jul;55(7):348-353. doi: 10.5483/BMBRep.2022.55.7.034. PMID: 35725010; PMCID: PMC9340087.

[2] He Y., Lu J, Ye Z., Hao S., Wang L., Kohli M., Tindall D.J, Li B., Zhu R., Wang L., Huang H. Androgen receptor splice variants bind to constitutively open chromatin and promote abiraterone-resistant growth of prostate cancer. Nucleic Acids Res. 2018 Feb 28;46(4):1895-1911. doi: 10.1093/nar/gkx1306. PMID: 29309643; PMCID: PMC5829742.

[3] Li Y., Gao Xx., Wei C., Guo R., Xu H., Bai Z., Zhou J., Zhu J., Wang W., Wu Y., Li J., Zhang Z., and Xie X. Modification of Mcl-1 alternative splicing induces apoptosis and suppresses tumor proliferation in gastric cancer. Aging (Albany NY). 2020 Oct 14;12(19):19293-19315. doi: 10.18632/aging.103766. Epub 2020 Oct 14. PMID: 33052877; PMCID: PMC7732305.

[4] Greenbaum, A., A. Rajput, and G. Wan, RON kinase isoforms demonstrate variable cell motility in normal cells. Heliyon, 2016. 2(9): p. e00153.

[5] St-Pierre A.P., Shikon V, and Schneider D.C. Count data in biology-Data transformation or model reformation? Ecol Evol. 2018 Feb 16;8(6):3077-3085. doi: 10.1002/ece3.3807. PMID: 29607007; PMCID: PMC5869353.

[6] Ramachandran, K.M. and Tsokos, C.P. Mathematical Statistics with Applications in R. San Diego, CA: Academic Press, 2020.

[7] Glenn Griesinger, N.L., Vrinceanu, D., Jackson, M., and Howell, W.C.. Elementary Statistics: A Guide to Data Analysis Using R. San Diego, CA, Cognella, 2023.

[8] Passini MA, Bu J, Richards AM, Kinnecom C, Sardi SP, Stanek LM, Hua Y, Rigo F, Matson J, Hung G, Kaye EM, Shihabuddin LS, Krainer AR, Bennett CF, and Cheng SH. Antisense oligonucleotides delivered to the mouse CNS ameliorate symptoms of severe spinal muscular atrophy. Sci Transl Med. 2011 Mar 2;3(72):72ra18. doi: 10.1126/scitranslmed.3001777. PMID: 21368223; PMCID: PMC3140425.

[9] Godavarthi, J.D., Polk, S., Nunez, L., Shivachar, Amruthesh, Glenn Griesinger, N.L., and Matin, A. (2020). Deficiency of splicing factor 1 (SF1) reduces intestinal polyp incidence in $Apc^{Min/+}$ mice. Biology, 9(1):1–13.

[10] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[11] Chu, J. T., and Hotelling, H. (1955). The Moments of the Sample Median. The Annals of Mathematical Statistics, 26(4), 593–606. http://www.jstor.org/stable/2236373.

[12] Kermer, H., Zhu, J., Muandet, K., and Scholkopf, B. (2022). Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions, Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA.

## Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

## Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

## Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

## Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)