*Brazilian Journal of Biometrics*, **43**, 1–18 (2025)
doi:10.28951/bjb.v43i3.754


ARTICLE

# Sample size estimation of skewed distributions in medical research.

Sadanandam Vemula and Kalpanapriya Dhakshanamoorthy*

Department of Mathematics, Vellore institute of Technology, Vellore, India.
*Corresponding author. Email: dkalpanapriya@vit.ac.in

(Received: July 30, 2024; Revised: December 4, 2024; Accepted: December 19, 2024; Published: June 9, 2025)

**Abstract**

The primary goal of calculating sample size is to ascertain the minimum number of samples required to identify meaningful changes in treatment outcomes, clinical parameters, or associations following data collection. Determining the sample size is the initial and crucial step in organizing a clinical trial. An improper assessment of this number could result in the approval of an ineffective medication or the rejection of an effective one. Sample size estimations should align with the intended analysis methodology. We will use generalized linear models (GLMs) to analyze the data, frequently employing normal approximations for non-normal distributions. The Binomial, Negative Binomial, Poisson, and Gamma families are specific cases where we utilize GLM theory to derive sample size formulas when comparing two means. We evaluated the performance of normal approximations by simulating various distributions using the log-link and identity-link functions. First, we examined the extent of errors in normal approximations for discrete probability distributions. Next, we applied GLM theory to derive sample size equations, which were evaluated through case studies and simulations. The Negative Binomial and Gamma distributions under study are well-suited for calculations on the link function (log) scale, often providing greater accuracy than normal approximations. However, the Binomial and Poisson distributions offer minimal advantage. The proposed method effectively calculates sample sizes when comparing the means of highly skewed outcome variables.

**Keywords**: Sample size; Power; Skewed distribution; Generalised linear model; Discrete distribution.

## 1. Introduction

For clinical and community research, as well as basic biological studies, including animal research, accurately determining the appropriate sample size is essential. A study's limitations often stem from the fact that, in fundamental biomedical science, researchers frequently employ the sample size used in comparable studies (Yan *et al.,* 2017). Understanding the scope of the research in terms of the desired techniques of analysis, the appropriate study designs, and the characteristics

of the proposed study population is crucial (Bolarinwa, 2020; Boneau, 1960). Charan proposed a method that employs power analysis to ascertain the sample size for animal studies, a process that closely resembles the determination of sample sizes in clinical and population studies. The sample size is critical for both testing and estimating the accuracy of diagnostic tests in medicine. In a medical setting, a small sample size results in an inaccurate assessment of accuracy with a broad confidence interval, which is not helpful to decision-makers (Hajian-Tilaki, 2014). Conversely, an excessively large sample size is a waste of resources, particularly when the new diagnostic test is expensive.

Lachin explains that we frequently use the normal approximation when calculating sample sizes, especially for non-Gaussian data analyzed with generalized linear models (GLMs). Even when using Poisson regression, several medical statistics textbooks still estimate sample sizes for rates using a normal approximation. A statistical procedure that is not in line with the sample size calculation could lead to a discrepancy between the nominal and actual power values. The analysis of biomedical data has frequently employed generalized linear models. For statistical inference, GLMs frequently use the Wald test and the likelihood ratio (LR) test. However, in small and intermediate samples, the Wald and LR tests may be too liberal. In recent clinical studies, the adoption of the Negative Binomial (NB) model for count data has increased. Clinical studies typically prefer it over the Poisson model in instances of overdispersed count data. The determination of sample size is one of the challenges in using the Negative Binomial model in clinical trial design. In practice, simulation techniques are extensively used to estimate sample sizes. Data with a Binomial distribution, similar to that of logistic regression, can also be modeled using generalized linear models (Lee & Conway, 2022). Disease mapping research frequently employs the Poisson distribution to investigate spatial variance in disease counts. Unfortunately, equidispersion, or the equivalency of the variance and mean, is a well-known drawback of the Poisson distribution. This assumption is frequently violated, leading to overdispersion in the data, where the variance exceeds the mean. Underdispersion, where the variance is smaller than the mean, occurs less frequently. To mitigate overdispersion, we frequently allow the Poisson mean to fluctuate between covariate groups or geographic units by using fixed and random effects.

As a generalization of exponential distributions, the gamma distribution belongs to the family of continuous probability distributions (Tripathi *et al.,* 1993). Leonhard Euler, a Swiss mathematician, is credited with creating the gamma distribution function, according to Nagar, Correa, and Gupta (Nagar *et al.,* 2013). Many researchers have developed and investigated this function due to its perceived significance. A study on assessing the homogeneity of the gamma distribution's characteristics (shape and scale) was carried out, among others, by Bhattacharya (Bhattacharya, 2002) and Bhaumik (Bhaumik *et al.,* 2009). A study of the shape, scale, and position of the gamma distribution's probability density function (pdf) was conducted by Chen and Kotz (Chen & Kotz, 2013). Numerous scholars, including Schickedanz and Krause, have also studied and developed bivariate gamma distributions. They investigated the application of the generalized likelihood ratio (GLR) for testing scale parameters from two gamma-distributed data sets.

The current study examines a dichotomous predictor variable, which compares two means. To determine whether the normal approximation works with certain cumulative distribution functions and situations, we used the Berry-Esseen theorem and calculated the relevant distributions. We specifically concentrate on the Negative Binomial and Gamma distributions, in part because they might describe skewed data, for which normal approximations are less likely to be adequate. For the former, we partially reproduce the work of Zhu and Lakkis. We derive a general formula that includes, for example, the Poisson and Binomial distributions. These methods are applied to real-world research cases. The sample size equations used in this article to compare the means of the Negative Binomial, Binomial, Poisson, and Gamma distributions were developed using the GLM method. We conducted a simulation study using the logit, log, and identity link functions to assess the performance of the normal approximation for various distributions with varying effect sizes and

a constant power of 90%.

## 2. Methods

For discrete probability distributions, we investigate the magnitude of errors in normal approximations. After that, we create sample size formulas using GLM theory, and we evaluate them using simulations and worked instances. To determine whether the normal approximation works with certain cumulative distribution functions and situations, we used the Berry-Esseen theorem and calculated the relevant distributions. We used R software for computation throughout.

**Using the Berry-Esseen theorem:**

Let $R_1, R_2, \ldots, R_n$ be independent and identically distributed zero-mean random variables with positive variance $\sigma^2$. The standardized mean of a random variable is given by

$$S_n = \sum_{k=1}^{n} \frac{R_k}{\sigma\sqrt{n}}. \tag{1}$$

The Berry-Esseen theorem, where $\Phi$ is the CDF of the standard Gaussian distribution and $F_n(\gamma)$ is the CDF of $S_n$, applies as follows (Feller, 1971). The theorem states that $\rho < \infty$ is the absolute third central moment, and $C$ is a distribution-independent positive constant, then

$$\left|F_n(\gamma) - \Phi(\gamma)\right| \leq \frac{C\rho}{\sigma^3\sqrt{n}}. \tag{2}$$

You can use the Berry-Esseen method even when a direct calculation from the distribution is not practical. We can express the bound using a finite sum and the third non-absolute central moment. These bounds can be used to evaluate the sufficiency of the Gaussian distribution assumptions underlying popular sample size calculations (Stonehouse & Forrester, 1998). In the section that follows, we describe a sample size technique that may be more reliable.

### 2.1 Sample Sizes Based on Generalized Linear Model Theory

An exponential family distribution yields vectors of independent responses, $Y_i\,(i = 1, \ldots, N)$, which are characterized by GLMs. The covariates in the model, $x_{ij}$, are made up of a linear combination of unknown regression coefficients. These can be expressed as exponential family distributions that produce vectors of independent responses, $Y_i\,(i = 1, \ldots, N)$, described by GLMs. The model's covariates, $x_{ij}$, as a linear combination of unknown regression coefficients, can be represented as

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Here, $\eta_i$ is linked to $\mu_i$, the mean of $Y_i$, through the link function $\eta_i = g(\mu_i)$. The sample size for a hypothesis associated with the mean of such a distribution on the scale of the link function can be determined using the variance of its maximum likelihood estimate (MLE). For GLMs, the covariance matrix of the parameter estimates is approximately

$$(X^T W X)^{-1} \tag{3}$$

where $W$ is the weighted diagonal matrix and $X$ is the design matrix (Zelterman, 2005). To understand how the sample size affects the variance of the parameter estimates, consider comparing the means of two groups of sizes $N_0$ and $N_1$. Here, $X$ comprises two columns and $N$ rows, where

$N_0 + N_1 = N$. The $N_0$ 0's and $N_1$ 1's in the second column follow all of the 1's in the first column, which represents the intercept. The letter $W$ is defined by

$$W = \frac{\left(\frac{d\mu}{d\eta}\right)^2}{V(\mu)} \tag{4}$$

where the variance function that connects $Y$'s variance to its mean is denoted by $V(\mu)$ (McCullagh, 2019). We can clearly see that the diagonal of $W$ consists of $N_0$ copies of $w_0$ and $N_1$ copies of $w_1$. We are interested in comparing the two means of the second diagonal element of the $2 \times 2$ matrix, which is given by Equation (3). Basic matrix algebra shows that this element is given by

$$\left[(N_0 w_0)^{-1} + (N_1 w_1)^{-1}\right].$$

The guidelines from Lachin regarding sample size are utilized in this comparison. In Lachin's notation, the subscripts 0 and 1 denote the null and alternative hypotheses, respectively. In this article, we use $O$ and $A$ to represent the null and alternative hypotheses, while 0 and 1 refer to the reference (control) and intervention groups, respectively. Additionally, we will use $\lambda$ as the generic parameter instead of $\mu$, with $\mu$ serving as the mean denoter. Furthermore, we will adopt a different subscript notation for standard normal deviations, such that $z_p$ represents the standard normal deviation for the lower tail area $p$. Our statistic ($X$ in Lachin's notation) is the estimate of the transformed mean difference generated by the GLM (Zhang *et al.,* 2007). Typically, we use a log transformation, or logit for a binomial. This statistic's mean is $\lambda_O$ under the null hypothesis and $\lambda_A$ under the alternative hypothesis. Its standard deviations are $\sum_O$ and $\sum_A$, respectively (Lachin, 1981). These values lead to the formulation of Lachin's equation:

$$\left|\lambda_A - \lambda_O\right| = Z_{1-\alpha/2} \sum_O - Z_{1-\beta} \sum_A . \tag{5}$$

Following Lachin's lead, the proportions in the groups are denoted by

$$Q_0 = \frac{N_0}{N}, \quad Q_1 = \frac{N_1}{N}.$$

Our method is to use a normal approximation on the scale of the link function. Often, we use the logarithm to approximate the scale of the link function, but the identity link yields more well-known equations. We consider two methods for approximating the variance under the $H_0$(Null hypothesis). Both groups use the reference value in the first technique, which we refer to as Method-1 in honor of Zhu and Lakkis (Zhu & Lakkis, 2014). Using the above matrix algebra, we have

$$\sum_0 = \sqrt{\frac{1}{Q_1 N}\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2} + \frac{1}{Q_0 N}\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}}$$

$$= \sqrt{\left(\frac{1}{Q_1} + \frac{1}{Q_0}\right)\frac{1}{N}\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}}$$

and

$$\sum_A = \sqrt{\frac{1}{Q_1 N}\frac{V(\mu_1)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_1}\right)^2}} + \sqrt{\frac{1}{Q_0 N}\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}}.$$

Hence for Method 1, we obtain

$$|\lambda_A - \lambda_O| = \frac{Z_{1-\alpha/2}\sqrt{\left(\frac{1}{Q_0}+\frac{1}{Q_1}\right)\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}} + Z_{1-\beta}\sqrt{\frac{V(\mu_1)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_1}\right)^2}\left(\frac{1}{Q_1}\right)+\left(\frac{1}{Q_0}\right)\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}}}{\sqrt{N}}$$

$$\sqrt{N} = \frac{Z_{1-\alpha/2}\sqrt{\left(\frac{1}{Q_0}+\frac{1}{Q_1}\right)\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}} + Z_{1-\beta}\sqrt{\frac{V(\mu_1)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_1}\right)^2}\left(\frac{1}{Q_1}\right)+\left(\frac{1}{Q_0}\right)\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}}}{g(\mu_0)-g(\mu_1)}. \quad (6)$$

According to Zhu and Lakkis (Zhu & Lakkis, 2014), the test characteristics are often better if the intervention arm is utilized under the null hypothesis (method 2) rather than $\mu_1$, so $\sum_0 = \sum_A$, then

$$\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2} = \frac{V(\mu_1)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_1}\right)^2}$$

and we derive

$$\sqrt{N} = \frac{(Z_{1-\alpha/2}+Z_{1-\beta})\sqrt{\frac{V(\mu_1)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_1}\right)^2}\left(\frac{1}{Q_1}\right)+\left(\frac{1}{Q_0}\right)\frac{V(\mu_0)}{\left(\frac{d\mu}{d\eta}\Big|_{\mu=\mu_0}\right)^2}}}{g(\mu_0)-g(\mu_1)}. \quad (7)$$

We can readily apply the general equations (6) and (7) to determine the distributional situations of interest. With the exception of references to earlier work using Method 1, we will apply equation (7).

## 2.2 Binomial distribution

Assume that $X$ is a Binomial variable with $n$ distinct independent events, each with a probability of $p$. The probability mass function gives the likelihood of obtaining precisely $X$ successes in $n$ separate Bernoulli trials. By assuming that $n = 1$, it would be possible to understand why the literature occasionally fails to include $n$ in the variance function (Cundill & Alexander, 2015).

We use the Exponential family function:

$$f(x) = e^{x\theta + n\log(1-p)+\log\binom{n}{x}}$$

$$\theta = \log\left(\frac{p}{1-p}\right)$$

$$b(\theta) = n\log\left(1+e^{\theta}\right), \quad a(\phi) = 1$$

$$\mu = n\frac{e^{\theta}}{1+e^{\theta}}$$

$$\mu = \frac{ne^{\theta}}{1+e^{\theta}}$$

$$p = \frac{\mu}{n}$$

$$\theta = \log\left(\frac{\mu}{n-\mu}\right)$$

$$\frac{d\mu}{d\theta} = \frac{\mu(n-\mu)}{n}$$

For $n = 1$ and $\theta = \eta$, we get

$$\frac{d\mu}{d\eta} = \mu(1-\mu)$$

and

$$V(\mu) = \mu(1-\mu)$$

canonical logit link is

$$\frac{d\mu}{d\eta} = \mu(1-\mu).$$

Therefore, using equation (7), we have

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\frac{1}{\mu_1(1-\mu_1)} + \frac{1}{Q_0}\frac{1}{\mu_0(1-\mu_0)}}}{\sqrt{d}\left[\text{logit}(\mu_0) - \text{logit}(\mu_1)\right]}. \tag{8}$$

The related equation on the identity link scale of difference in proportions is

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\mu_1(1-\mu_1) + \frac{1}{Q_0}\mu_0(1-\mu_0)}}{\sqrt{d}\left[(\mu_0) - (\mu_1)\right]}. \tag{9}$$

## 2.3 Negative Binomial Distribution

The number of failures in a sequence of independently distributed Bernoulli trials that must occur before a given number of successes is known as the Negative Binomial distribution (NBD), and it is a discrete probability distribution. A little $k$ indicates a large variance as $k \to \infty$; then the distribution tends to be Poisson. Poisson regression is modified by Negative binomial regression, which reduces the need for equidispersion (Holodinsky *et al.,* 2021). We have extensively evaluated both overdispersed count data and recurrent event data using the NB regression. A Poisson-gamma mixture can be used to represent the NB distribution (Tang *et al.,* 2021). Let $X$ be a random variable with a variance function of $V(\mu)$, and let it follow the Negative Binomial distribution with a population mean of $\mu$ and a dispersion parameter of $k$, then we have $v(\mu) = \mu + \frac{\mu^2}{k}$. The logarithmic link function is

$$\frac{d\mu}{d\eta} = \mu$$

$$\eta = \log(\mu)$$

using equation (6), we have

$$\sqrt{N} = \frac{Z_{1-\alpha/2}\sqrt{\left(\frac{1}{Q_0} + \frac{1}{Q_1}\right)\left(\frac{\mu_0 + \frac{\mu_0^2}{k_0}}{\mu_0^2}\right)} + Z_{1-\beta}\sqrt{\left(\frac{1}{Q_1}\frac{\mu_1 + \frac{\mu_1^2}{k_1}}{\mu_1^2} + \frac{1}{Q_0}\frac{\mu_0 + \frac{\mu_0^2}{k_0}}{\mu_0^2}\right)}}{\log(\mu_0) - \log(\mu_1)}$$

$$\sqrt{N} = \frac{Z_{1-\alpha/2}\sqrt{\left(\frac{1}{Q_0} + \frac{1}{Q_1}\right)\left(\frac{1}{\mu_0} + \frac{1}{k_0}\right)} + Z_{1-\beta}\sqrt{\frac{1}{Q_1}\left(\frac{1}{\mu_1} + \frac{1}{k_1}\right) + \frac{1}{Q_0}\left(\frac{1}{\mu_0} + \frac{1}{k_0}\right)}}{\log(\mu_0) - \log(\mu_1)}. \tag{10}$$

In the unique scenario where sample sizes are equal, $Q_0 = Q_1 = 0.5$, $k$-parameter $k_0 = k_1$, and after changing in equation (7), we get

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\left(\frac{1}{\mu_1} + \frac{1}{k_1}\right) + \frac{1}{Q_0}\left(\frac{1}{\mu_0} + \frac{1}{k_0}\right)}}{\log(\mu_0) - \log(\mu_1)}. \tag{11}$$

Applying equation (7) to the identity scale with variances equal to $\mu_i + \frac{\mu_i^2}{k_i}$ (i=0,1) yields a normal approximation:

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\left(\mu_1 + \frac{\mu_1^2}{k_1}\right) + \frac{1}{Q_0}\left(\mu_0 + \frac{\mu_0^2}{k_0}\right)}}{\mu_0 - \mu_1}. \tag{12}$$

We were able to use simulations to figure out the real power sample sizes from equations (11) and (12) by making repeated datasets of the calculated sizes and analyzing them using the Wald and GLM tests.

## 2.4 Poisson distribution

The Poisson distribution with parameter $\lambda$ is stated to exist for a discrete random variable $X$ with mean $\mu$. We can allow the value of $k$ in equation (11) to reach infinity, or we can also utilize the log link, $\nu(\mu) = \mu$ in equation (7).

**Natural exponential family:**

$$\begin{aligned}
f(x) &= \left[e^{\frac{\theta x - b(\theta)}{a(\psi)} + c(x,\psi)}\right] \\
\log(f(x)) &= x\log(\lambda) - \lambda - \log(x!) \\
f(x) &= e^{x\log(\lambda) - \lambda - \log(x!)} \\
\theta &= \log(\lambda) \\
\lambda &= e^{\theta} \\
b(\theta) &= e^{\theta} \\
\mu &= \frac{d}{d\theta}(e^{\theta}) = e^{\theta} = \lambda \\
V(x) &= a(\psi)\frac{d^2}{d\theta^2}(e^{\theta}) = e^{\theta} = \lambda
\end{aligned}$$

canonical link is

$$\log(\lambda) = \theta = \eta$$

we obtain

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\left(\frac{1}{\mu_1} + \frac{1}{k_1}\right) + \frac{1}{Q_0}\left(\frac{1}{\mu_0} + \frac{1}{k_0}\right)}}{\log(\mu_0) - \log(\mu_1)}. \tag{13}$$

This is compared using simulation for the situation of $Q_0 = Q_1 = 0.5$ on the scale of the identity link with the normal approximation that follows, which is derived from equation (12) by letting $k$ tend to infinity

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{2(\mu_1 + \mu_0)}}{\mu_0 - \mu_1}. \tag{14}$$

### 2.5 Gamma distribution

The Gamma distribution can be parameterized using the shape parameter $k$ and the scale parameter $\theta$, denoted as $Y \sim \gamma(k, \theta)$, where $Y$ represents a gamma-distributed random variable. The sum of identical independent exponentials and the exponential distribution are examples of special situations. Applications commonly use models of right-skewed data. The Gamma distribution is commonly used to simulate waiting periods in econometrics and other applied sciences, making the parameterization with $k$ and $\theta$ more popular in these fields. In the event where $Y$ is a random variable with scale parameter $\theta$ and shape parameter $k$, then $E(Y) = k\theta$.

$$\begin{aligned} V(\mu) &= \frac{k}{\theta^2} \\ &= k\left(\frac{\mu}{k^2}\right) \\ V(\mu) &= \frac{\mu^2}{k}. \end{aligned}$$

Log-link function

$$\begin{aligned} \eta &= \log(\mu) \\ \frac{d\mu}{d\eta} &= \mu. \end{aligned}$$

As noted by (Forbes *et al.,* 2011), the variance function holds for the Gamma distribution. Using equation (7), we obtain

$$\sqrt{N} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\frac{1}{K_1} + \frac{1}{Q_0}\frac{1}{K_0}}}{\log(\mu_0) - \log(\mu_1)}. \tag{15}$$

## 3. Results

Supplementary file (Table 4) displays the Berry-Esseen limits and related values derived from the computation of the non-Gaussian CDFs for the case with a fixed sample size of 100. The normal approximation performs better for larger means, as expected, according to both approaches. Berry-Esseen bounds are frequently substantially broader than explicit computation results. Therefore, we focus on the latter strategy. The results for Binomial distributions with different sample sizes and proportions ($\mu$) are displayed in supplementary file (Figure 1). As expected, differences in the CDF of the normal approximation tend to be larger for lower sample numbers and values of $\mu$ that are further from 0.5. For parameter values discovered in some research investigations, the differences are non-negligible, especially for small values of $\mu$, say between 1 and 5%, which would be anticipated to approximate Poisson. This tends to maintain the concern that power estimates based on common approximations might not be precise.

We must relate a distribution of interest ($Y$) to use the Berry-Esseen theorem. We assume a discrete distribution with a zero mean. Let $Y$ be a non-negative discrete random variable with mean $\mu_Y$ and variance $\sigma^2$, and define $R = Y - \mu_Y$. We can then estimate the third central moment of $R$, $m_3$, using $f$, the probability density function of $Y$. The first sum's terms are either negative or zero, while the second sum's terms are positive. We can express the third absolute central moment

of $R$ as a finite sum in the following way

$$\begin{aligned} m_3 &= E[|(R-\mu_R)^3|] \\ &= \sum_{\gamma>0} |(Y-\mu_Y)^3| f(\gamma) \\ &= \sum_{\gamma=0}^{\mu_\gamma} |(\gamma-\mu_Y)^3| f(\gamma) + \sum_{\mu_\gamma+1}^{\infty} |(\gamma-\mu_Y)^3| f(\gamma) \\ &= -\sum_{\gamma=0}^{\mu_\gamma} (\gamma-\mu_Y)^3 f(\gamma) + \sum_{\mu_\gamma+1}^{\infty} |(\gamma-\mu_Y)^3| f(\gamma) \\ &= E[(Y-\mu_Y)^3] - 2\sum_{\gamma=0}^{\mu_\gamma} (\gamma-\mu_Y)^3 f(\gamma). \end{aligned}$$

## 3.1 GLM Method for Binomial Distribution

As the magnitude of the effect size increases, fewer samples are needed to achieve a power of 0.9. For example, detecting small effect sizes (such as 0.2) requires approximately 3,400 samples, while large effect sizes (such as 0.8) can be identified with fewer than 100 samples. Equations on both the log scale (Equation 8) and the untransformed scale (Equation 9) were analyzed with $n = 1$ for various values of $\mu_0$ and efficacy. We conducted 10,000 simulations for each configuration of parameters. Both models demonstrated identical power when $\mu = 0.5$. The observed trends are consistent with lower Poisson means, specifically for $\mu_0 = 0.1$ and 0.05, and both methods tend to show conservative behavior at higher efficacies. While the two approaches are generally similar, sample size estimates from the logit-link method tend to be marginally higher than those from the identity-link method. Detailed results are presented in Table 1, while the comparison of log-scale and identity-scale results with $\mu_0 = 0.1$ and varying efficacies is available in supplementary file (Table 1). Figure 1 illustrates the sample size requirements for detecting various effect sizes in a Binomial distribution with a mean of 0.5, comparing log-link and identity-link functions at 90% power.

For smaller effect sizes like 0.2, the required sample sizes are 3,398 for the logit-link and 3,383 for the identity-link approach. For moderate effect sizes, such as 0.5, the sample size requirements decrease to 372 for the logit-link and 357 for the identity-link. For larger effect sizes like 0.8, the sample sizes required are 91 for the logit-link and 74 for the identity-link. These results highlight that detecting smaller effects demands considerably more resources, while larger effects are easier to identify with fewer samples, making them more practical for studies with limited resources. The slight variation between the logit-link and identity-link functions indicates that both are effective, though the choice of method should be based on theoretical considerations relevant to the study design.

All calculations are based on simulations repeated 10,000 times with a mean of 0.5, ensuring the robustness of the estimates. Effect size plays a critical role in determining sample size requirements, and both the logit and identity-link functions yield comparable results, with the identity-link being marginally more efficient in certain cases. Researchers should carefully evaluate the assumptions and methodology to ensure accurate and reliable results.

**Table 1.** Comparing the Binomial means with various efficacies using logit and Identity link functions and each data values are simulated 10,000 times with mean 0.5

| Effect size | Power | Sample size determination using logit link | Sample size determination using Identity link |
|---|---|---|---|
| 0.2 | 0.9 | 3398 | 3383 |
| 0.23 | 0.9 | 2483 | 2468 |
| 0.25 | 0.9 | 2053 | 2039 |
| 0.28 | 0.9 | 1580 | 1565 |
| 0.30 | 0.9 | 1343 | 1329 |
| 0.35 | 0.9 | 928 | 913 |
| 0.38 | 0.9 | 757 | 743 |
| 0.4 | 0.9 | 666 | 652 |
| 0.43 | 0.9 | 554 | 540 |
| 0.45 | 0.9 | 492 | 478 |
| 0.48 | 0.9 | 415 | 401 |
| 0.5 | 0.9 | 372 | 357 |
| 0.53 | 0.9 | 317 | 302 |
| 0.55 | 0.9 | 286 | 271 |
| 0.58 | 0.9 | 246 | 231 |
| 0.6 | 0.9 | 223 | 208 |
| 0.63 | 0.9 | 193 | 177 |
| 0.65 | 0.9 | 176 | 160 |
| 0.68 | 0.9 | 153 | 137 |
| 0.7 | 0.9 | 140 | 124 |
| 0.73 | 0.9 | 122 | 106 |
| 0.75 | 0.9 | 113 | 96 |
| 0.78 | 0.9 | 99 | 82 |
| 0.8 | 0.9 | 91 | 74 |

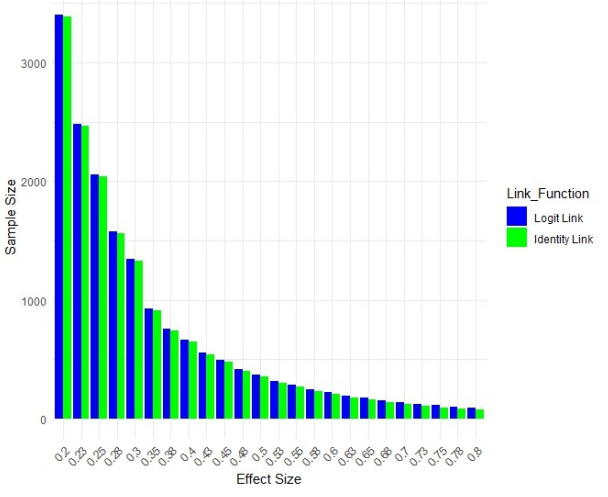


**Figure 1.** Visualization of sample size estimation for detecting various effect sizes in a Binomial distribution with mean $\mu_0 = 0.5$, comparing log-link and identity-link functions at 90% power.

## 3.2 GLM Method for Negative Binomial Distribution

As the effect size increases, the required sample size to achieve a power of 0.9 decreases. For smaller effect sizes, such as 0.2, a larger sample size is needed (around 2,110 for the log-link and 2,122 for the identity-link) compared to larger effect sizes. For moderate effect sizes like 0.5, the sample sizes required are 263 for the log-link and 274 for the identity-link. Similarly, for large effect sizes like 0.8, the required sample sizes are 81 for the log-link and 87 for the identity-link. In general, both the log-link and identity-link methods provide similar results, with a slight difference in sample sizes. The log-link approach tends to require slightly smaller sample sizes compared to the identity-link for most effect sizes. These results are based on 10,000 simulations with a mean of $\mu = 0.75$, ensuring the robustness of the sample size estimates.

These findings emphasize that larger effect sizes can be detected with fewer samples, making studies of larger effects more feasible in resource-constrained settings. The slight difference between the log-link and identity-link methods suggests that the choice of link function is not crucial but may depend on theoretical considerations specific to the study design. Using the null hypothesis that both means are equal to 0.75, a power of 90%, and a significance level of 5% (two-tailed), the sample size was determined for the given method. The key parameters were $\mu = 0.75$, $k_0 = k_1 = 1$, and $Q_0 = Q_1 = 0.5$, derived from a crash data analysis using maximum likelihood bootstrapped likelihood estimation. We compared equation (11) on a logarithmic scale and equation (12) on an identity scale. The results indicated that, in the case of a negative binomial distribution, the sample size values for the identity link function were slightly higher than those for the log-link function. As the effect size increased, the sample sizes for both methods decreased accordingly. The detailed results are presented in Table 2.

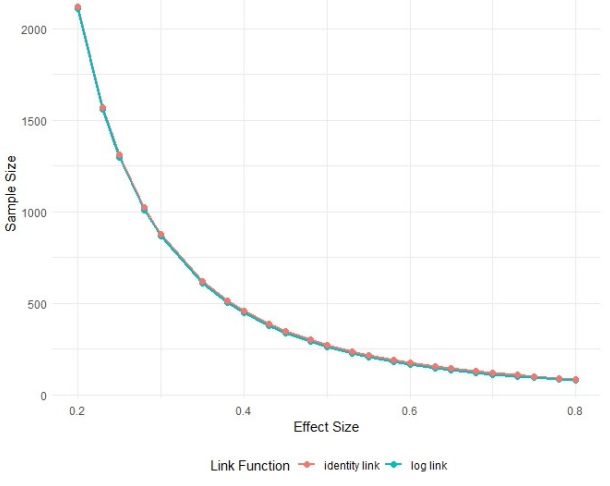


**Figure 2.** Visualization of sample size estimation for detecting various effect sizes in a Negative Binomial distribution with mean $\mu_0 = 0.75$, comparing log-link and identity-link functions at 90% power.

The simulations were conducted in two different configurations: (a) the dispersion parameter $k$ was allowed to vary from 0.1 to 10, with the Poisson distribution acting as the limiting case when $k = \infty$; (b) the efficacy, defined as $1 - \left(\frac{\mu_1}{\mu_0}\right)$, was varied from 0.2 to 0.8. When using the log-link function, the efficacy was maintained at 20%, which closely matched the nominal power. In contrast, the identity-link method showed a slightly conservative trend. As the mean difference and effectiveness increased, the log-link method consistently approximated the nominal power, while the identity link resulted in an overestimation of the sample size by more than 50% for the higher

**Table 2.** Comparing the means of Negative binomial distribution using nominal power 0.90, various efficacies using log-link and Identity link with mean $\mu = 0.75$, each data value is simulated 10,000 times.

| Effect size | Power | Sample size determination using log link | Sample size determination using Identity link |
|---|---|---|---|
| 0.2 | 0.9 | 2110 | 2122 |
| 0.23 | 0.9 | 1558 | 1571 |
| 0.25 | 0.9 | 1298 | 1310 |
| 0.28 | 0.9 | 1010 | 1022 |
| 0.30 | 0.9 | 866 | 878 |
| 0.35 | 0.9 | 610 | 621 |
| 0.38 | 0.9 | 504 | 516 |
| 0.4 | 0.9 | 448 | 460 |
| 0.43 | 0.9 | 378 | 389 |
| 0.45 | 0.9 | 339 | 350 |
| 0.48 | 0.9 | 290 | 301 |
| 0.5 | 0.9 | 263 | 274 |
| 0.53 | 0.9 | 228 | 238 |
| 0.55 | 0.9 | 208 | 218 |
| 0.58 | 0.9 | 182 | 192 |
| 0.6 | 0.9 | 167 | 177 |
| 0.63 | 0.9 | 148 | 157 |
| 0.65 | 0.9 | 136 | 146 |
| 0.68 | 0.9 | 121 | 130 |
| 0.7 | 0.9 | 113 | 121 |
| 0.73 | 0.9 | 101 | 109 |
| 0.75 | 0.9 | 95 | 102 |
| 0.78 | 0.9 | 86 | 92 |
| 0.8 | 0.9 | 81 | 87 |

efficacy values. Figure 2 presents the sample size estimation for detecting various effect sizes in a Negative Binomial distribution with a mean of 0.75, comparing log-link and identity-link functions at 90% power.

## 3.3 GLM method for Poisson distribution

We compared equations (13) and (14), which use the logarithmic and identity scales, respectively, with the aim of determining the required sample sizes for detecting various effect sizes in a Poisson distribution. These comparisons were made using a mean of $\mu_0 = 2.514$ and a nominal power of 90 %. Based on 10,000 simulations, both methods showed comparable results, with the log-link function providing slightly higher sample size estimates than the identity-link function, particularly for smaller effect sizes. As the effect size increased, the required sample size decreased. For small effect sizes (e.g., 0.2), large sample sizes (378 and 376 for log and identity links, respectively) were required, while larger effect sizes (e.g., 0.8) required significantly fewer samples (19 and 16 for log and identity links, respectively). Overall, both methods were effective, but the identity-link function was marginally more efficient in certain cases. The findings align with the results from a non-homogeneous Poisson process applied to COVID-19 data analysis in Kuwait, as discussed in (Al-Dousari *et al.,* 2021).

**Table 3.** Comparing the means of Poisson distribution using nominal power 0.90,various efficacies using log-link and Identity-link functions with mean $\mu_0 = 2.514$, each data value simulated 10,000 times

| Effect size | Power | Sample size determination using log link | Sample size determination using Identity link |
|---|---|---|---|
| 0.2 | 0.9 | 378 | 376 |
| 0.23 | 0.9 | 281 | 280 |
| 0.25 | 0.9 | 236 | 234 |
| 0.28 | 0.9 | 184 | 183 |
| 0.30 | 0.9 | 160 | 158 |
| 0.35 | 0.9 | 138 | 137 |
| 0.38 | 0.9 | 96 | 94 |
| 0.4 | 0.9 | 86 | 84 |
| 0.43 | 0.9 | 73 | 71 |
| 0.45 | 0.9 | 66 | 64 |
| 0.48 | 0.9 | 57 | 55 |
| 0.5 | 0.9 | 52 | 50 |
| 0.53 | 0.9 | 46 | 44 |
| 0.55 | 0.9 | 42 | 40 |
| 0.58 | 0.9 | 38 | 36 |
| 0.6 | 0.9 | 35 | 32 |
| 0.63 | 0.9 | 31 | 29 |
| 0.65 | 0.9 | 30 | 27 |
| 0.68 | 0.9 | 27 | 24 |
| 0.7 | 0.9 | 25 | 22 |
| 0.73 | 0.9 | 23 | 20 |
| 0.75 | 0.9 | 22 | 19 |
| 0.78 | 0.9 | 20 | 17 |
| 0.8 | 0.9 | 19 | 16 |

For $\mu_0 = 2.514$ power is 0.90, and for different effect sizes, 10000 simulations were run for each data set using log-link and identity-link functions, the results are shown in Table 3. The Poisson

distribution simulation results with $\mu_0$ = 2.412, 3.223 are given in the Supplementary file Table 2 and Table 3. Figure 3 illustrates the sample size requirements for detecting various effect sizes in a Poisson distribution (mean = 2.514) using log-link and identity-link functions, assuming 90% power.

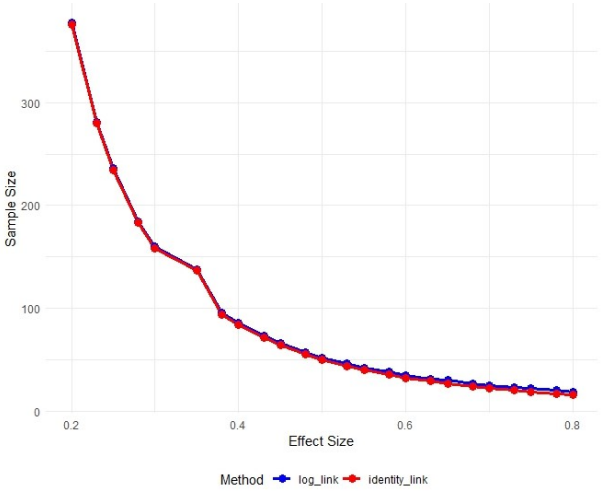


**Figure 3.** Visualization of sample size comparison for detecting various effect sizes in a Poisson distribution with mean $\mu_0$ = 2.514 using log-link and identity-link functions at 90% power.

## 3.4 GLM Method for Gamma Distribution

The Gamma distribution is a continuous probability distribution with two parameters (Kurniasari *et al.,* 2018). As in Yue et al. (2001), we used the problem of estimating the association parameter using the product-moment correlation coefficient with a mean of 9.68, a shape parameter of 2.50, and a scale parameter of 0.258 (Yue *et al.,* 2001). As previously, we compare the power of the sample sizes obtained from equation (15) with the sample sizes obtained via the proper normal estimation on the original scale. The sample size values for the two tests are the same, however, the identity-link method's sample size values are slightly higher when compared to the log-link approach. The results shown in Table 4 show that the sample size based on the link function's scale maintains close to nominal power, whereas the normal approximation overestimates the needed sample size by at least 50% for the larger mean differences. In this case, the likelihood ratio test produced higher estimated powers for both tests (not shown). Once more, the difference scale showed much higher power than the logarithmic scale due to the identical sample size inputs for both test protocols. Figure 4 presents the estimated sample sizes required to detect various effect sizes in a Gamma distribution (mean = 9.68), comparing log-link and identity-link functions at 90% power, based on 10,000 simulation replications.

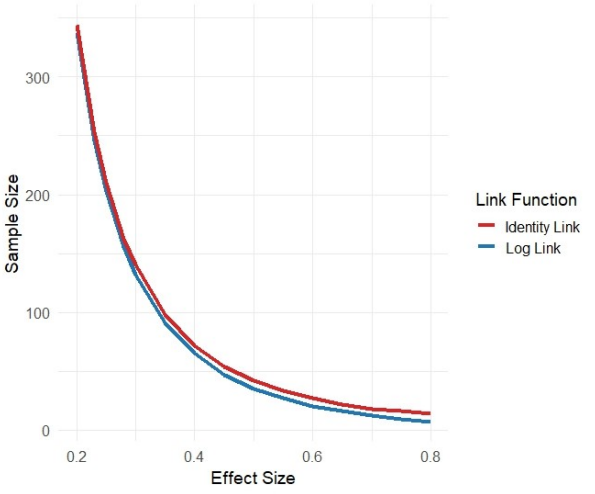


**Figure 4.** Sample size estimation for detecting various effect sizes in a Gamma distribution with mean $\mu = 9.68$, comparing log-link and identity-link functions at 90 % power, based on simulations repeated 10,000 times.

**Table 4.** Comparing the means of gamma distribution using log-link, Identity link with various efficacies, nominal power 0.90 and each data value simulated 10,000 times($\mu = 9.68$)

| Effect size | Power | Sample size determination using log link | Sample size determination using Identity link |
|---|---|---|---|
| 0.2 | 0.9 | 338 | 344 |
| 0.23 | 0.9 | 246 | 253 |
| 0.25 | 0.9 | 203 | 210 |
| 0.28 | 0.9 | 156 | 162 |
| 0.30 | 0.9 | 132 | 140 |
| 0.35 | 0.9 | 91 | 98 |
| 0.4 | 0.9 | 65 | 72 |
| 0.45 | 0.9 | 47 | 54 |
| 0.5 | 0.9 | 35 | 42 |
| 0.55 | 0.9 | 27 | 33 |
| 0.6 | 0.9 | 20 | 27 |
| 0.65 | 0.9 | 16 | 22 |
| 0.7 | 0.9 | 12 | 18 |
| 0.75 | 0.9 | 9 | 16 |
| 0.8 | 0.9 | 7 | 14 |

## 4. Discussions

When you look at the Poisson and Binomial results in terms of the difference in rates or proportions, they show that figuring out sample sizes on the scale of the link function, such as log rates or log odds, doesn't help much. On the other hand, sample size calculations based on mean differences can be quite conservative for the Negative Binomial and Gamma distributions, which include additional parameters that can indicate skewness. This results in bigger numbers that far exceed the required power. However, for the example studies, sample size estimations on the log scale stay rather close to the nominal power. Even in cases where we will employ generalized linear models to analyze the data, we frequently estimate sample sizes for discrete data using normal approximations to distributions. Based on the differences in CDF between the exact distributions and the normal approximation, as determined by distribution functions or the Berry-Esséen theorem, the inaccuracy could be significant and illogical. In theory, the rate of convergence of the normal approximation to that determined by the central limit theorem can be estimated using Berry-Esséen and related theorems (Feller, 1957; Korolev & Shevtsova, 2012). But frequently, their boundaries turned out to be noticeably broader than those found by calculating the relevant distribution's CDF. When analysis-stage robustness is taken into account, the t-test conducts well under some significant departures from normality (Heeren & D'Agostino, 1987).

## 5. Conclusions

The technique is particularly useful for distributions such as the negative binomial and gamma distributions, which, due to their parameters, can exhibit significant asymmetry, making normal estimation of the sample mean less accurate. Our approach is straightforward to apply and aligns closely with the generalized linear models (GLMs) commonly used to compare the means of non-normal distributions. This method is expected to be especially beneficial in scenarios where substantial asymmetry is anticipated in the response variable and where normal estimates are likely to be less reliable. We have demonstrated the advantages of these strategies by presenting examples across various distributions.

### Acknowledgments

We sincerely thank all the authors for their significant contributions to this study. Finally, we express our gratitude to our institute for their financial and moral support, which made this research possible.

### Conflicts of Interest

We have no conflicts of interest while preparing the manuscript.

### Author Contributions

**Conceptualization**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Data curation**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Formal analysis**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Funding acquisition**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Investigation**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Methodology**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Project administration**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Software**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Resources**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Supervision**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Validation**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Visualization**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Writing – original draft**: VEMULA, S.; DHAKSHANAMOORTHY, K. **Writing – review and editing:** VEMULA, S.; DHAKSHANAMOORTHY, K.

Supplementary-File-1: Comparison of means of Binomial and Poisson Distributions using Log-link and identity-link functions.

## References


1. Al-Dousari, A., Ellahi, A. & Hussain, I. Use of non-homogeneous Poisson process for the analysis of new cases, deaths, and recoveries of COVID-19 patients: A case study of Kuwait. *Journal of King Saud University-Science* **33,** 101614 (2021).
2. Bhattacharya, B. Tests of parameters of several gamma distributions with inequality restrictions. *Annals of the Institute of Statistical Mathematics* **54,** 565–576 (2002).
3. Bhaumik, D. K., Kapur, K. & Gibbons, R. D. Testing parameters of a gamma distribution for small samples. *Technometrics* **51,** 326–334 (2009).
4. Bolarinwa, O. A. Sample size estimation for health and social science researchers: The principles and considerations for different study designs. *Nigerian Postgraduate Medical Journal* **27,** 67–75 (2020).
5. Boneau, C. A. The effects of violations of assumptions underlying the t test. *Psychological bulletin* **57,** 49 (1960).
6. Chen, W. W. & Kotz, S. The Riemannian structure of the three-parameter Gamma distribution. *Scientific Research* (2013).
7. Cundill, B. & Alexander, N. D. Sample size calculations for skewed distributions. *BMC medical research methodology* **15,** 1–9 (2015).
8. Feller. *An introduction to Probability Theory and its Applications* (Wiley and Sons, 1971).
9. Feller, W. *An introduction to probability theory and its applications. 2* (Wiley, 1957).
10. Forbes, C., Evans, M., Hastings, N. & Peacock, B. *Statistical distributions* (John Wiley & Sons, 2011).
11. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics* **48,** 193–204 (2014).
12. Heeren, T. & D'Agostino, R. Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in medicine* **6,** 79–90 (1987).
13. Holodinsky, J. K., Yu, A. Y., Kapral, M. K. & Austin, P. C. Comparing regression modeling strategies for predicting hometime. *BMC Medical Research Methodology* **21,** 1–18 (2021).
14. Korolev, V. & Shevtsova, I. An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums. *Scandinavian Actuarial Journal* **2012,** 81–105 (2012).
15. Kurniasari, D., Warsono, W., Widiarti, W. & Usman, M. Estimation of Generalized Gamma Distribution Parameter with Probability Weighted Moment Method. *Science International Lahore* **30,** 1–6 (2018).
16. Lachin, J. M. Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials* **2,** 93–113 (1981).
17. Lee, C. S. & Conway, C. *The role of generalized linear models in handling cost and count data* 2022.
18. McCullagh, P. *Generalized linear models* (Routledge, 2019).
19. Nagar, D. K., Roldán-Correa, A. & Gupta, A. K. Extended matrix variate gamma and beta functions. *Journal of Multivariate Analysis* **122,** 53–69 (2013).
20. Stonehouse, J. M. & Forrester, G. J. Robustness of the t and U tests under combined assumption violations. *Journal of Applied Statistics* **25,** 63–74 (1998).

21. Tang, Y., Zhu, L. & Gu, J. An improved sample size calculation method for score tests in generalized linear models. *Statistics in Biopharmaceutical Research* **13,** 415–424 (2021).

22. Tripathi, R. C., Gupta, R. C. & Pair, R. K. Statistical tests involving several independent gamma distributions. *Annals of the Institute of Statistical Mathematics* **45,** 773–786 (1993).

23. Yan, F., Robert, M. & Li, Y. Statistical methods and common problems in medical or biomedical science research. *International journal of physiology, pathophysiology and pharmacology* **9,** 157 (2017).

24. Yue, S., Ouarda, T. B. & Bobée, B. A review of bivariate gamma distributions for hydrological application. *Journal of Hydrology* **246,** 1–18 (2001).

25. Zelterman, D. *Discrete distributions: applications in the health sciences* (John Wiley & Sons, 2005).

26. Zhang, Y., Ye, Z. & Lord, D. Estimating dispersion parameter of negative binomial distribution for analysis of crash data: bootstrapped maximum likelihood method. *Transportation Research Record* **2019,** 15–21 (2007).

27. Zhu, H. & Lakkis, H. Sample size calculation for comparing two negative binomial rates. *Statistics in medicine* **33,** 376–387 (2014).