Duris *et al. Journal of Statistical Distributions and Applications* (2018) 5:2
DOI 10.1186/s40488-018-0083-x

Journal of Statistical Distributions and Applications

RESEARCH Open Access

# Mean and variance of ratios of proportions from categories of a multinomial distribution

Frantisek Duris[1,2*], Juraj Gazdarica[3], Iveta Gazdaricova[3], Lucia Strieskova[3], Jaroslav Budis[4], Jan Turna[5] and Tomas Szemes[1,3,5]

*Correspondence:
fduris@dcs.fmph.uniba.sk
[1]Geneton s.r.o., Galvaniho 7, 82104 Bratislava, Slovakia
[2]Slovak Centre of Scientific and Technical Information, Lamacska cesta 7315/8A, 81104 Bratislava, Slovakia
Full list of author information is available at the end of the article

**Abstract**

Ratio distribution is a probability distribution representing the ratio of two random variables, each usually having a known distribution. Currently, there are results when the random variables in the ratio follow (not necessarily the same) Gaussian, Cauchy, binomial or uniform distributions. In this paper we consider a case, where the random variables in the ratio are joint binomial components of a multinomial distribution. We derived formulae for mean and variance of this ratio distribution using a simple Taylor-series approach and also a more complex approach which uses a slight modification of the original ratio. We showed that the more complex approach yields better results with simulated data. The presented results can be directly applied in the computation of confidence intervals for ratios of multinomial proportions.

**AMS Subject Classification:** 62E20

**Keywords:** Multinomial distribution, Ratio distribution, Mean, Variance

## 1 Introduction

Combinations of random variables (e.g., sums, products, ratios) regularly occur in many scientific areas. Particularly useful is the ratio of two random variables. For example, plant scientists use the ratio of leaf area to total plant weight (leaf area ratio) in the plant growth analysis (Poorter and Garnier 1996), and geneticists use the ratio of total genetic diversity distributed among populations to total genetic diversity in the pooled populations as a measure of population differentiation (Culley et al. 2002). The ratio of two fluorescent signals has several applications in fluorescence microscopy, e.g., estimating the DNA sequence copy number as a function of chromosomal location (Piper et al. 1995), and there are many (dimensionless) ratios employed in engineering (Mekic et al. 2012). In case of categorical data (i.e., from a binomial or multinomial distribution), there are numerous applications of ratios as well in consumer preference studies, election poll results, quality control, epidemiology, and so on.

Formally, a ratio distribution is a probability distribution constructed as the distribution of the ratio of two random variables, each having another (known) distribution. More particularly, given two random variables $Y_1$ and $Y_2$, the distribution of the random variable $Z$ that is formed as the ratio $Z = Y_1/Y_2$ is a ratio distribution. When using ratio distributions for theoretical and practical purposes, it is helpful to know its mean and variance, preferably in a computationally efficient form. In the case that $Y_1$ and $Y_2$ follow

normal distributions, and $\mu_{Y_2} = 0$, $Z$ is known as Cauchy distribution (Geary 1930; Fieller 1932; Hinkley 1969; Korhonen and Narula 1989; Marsaglia 2006). Other authors have addressed ratios of binomial proportions (also known as relative risk) (Koopman 1984; Bonett and Price 2006; Price and Bonett 2008), ratios of uniform distributions (Sakamoto 1943), Student's $t$ distributions (Press 1969), Weibull and gamma distributions (Basu and Lochner 1971; Provost 1989; Nadarajah and Kotz 2006), beta distributions (Pham-Gia 2000), Laplace and Bessel distributions (Nadarajah 2005; Nadarajah and Kotz 2005) and others. General notes on the product and ratio of two (not necessarily normal) random variables can also be found in (Frishman 1971; Van Kempen and Van Vliet 2000).

In our paper, we consider a ratio involving two or more random variables that jointly have a multinomial distribution. This situation is similar to relative risk or risk ratio which is the ratio of the probability of an event occurring (for example, developing a disease or being injured) in an exposed group to the probability of the event occurring in a comparison, non-exposed group. However, while the probabilities in the risk ratio are independent (in the sense that they describe two independent events in two independent groups), in our case, the probabilities are tied together through the covariance between multinomial categories. These ratios serve as a common framework for opinion polls, statistical quality control, and consumer preference studies. Confidence intervals for the odds ratio, which can be easily calculated, if the standard deviation is known, are especially important for applications. Nelson (1972) presented estimates, confidence intervals, and hypothesis tests for the odds ratio in trinomial distributions. Piegorsch and Richwine (2001) examined some types of confidence intervals in the context of analysis of genetic mutant spectra. Quesenberry and Hurst (1964) and Goodman (1965) explored methods for obtaining a set of simultaneous confidence intervals for the probabilities of a multinomial distribution. A comparison of performance of various confidence intervals also appeared in Alghamdi (2015); Aho and Bowyer (2015). To the best of our knowledge, however, there has been no analytical treatment of the ratio of multinomial proportions including derivations for formulae for the mean and variance of such a ratio.

A ratio between two or more random variables that jointly have a multinomial distribution also arises in the trending field of the non-invasive prenatal testing of common fetal aneuploidies such as trisomy of the $13^{th}$, $18^{th}$ or $21^{st}$ chromosome (Chiu et al. 2008; Sehnert et al. 2011; Lau et al. 2012; Minarik et al. 2015). We are currently working on implementation of this model into laboratory practice, and this paper represents a mathematical background of our work. In this paper, we discuss two solutions to the problem of mean and variance of the said ratio. More particularly, we derive asymptotic formulae for the mean and variance of the random variable $Z = Y_1/Y_2$, where $Y_1 = \sum_{k\in I} X_k$ and $Y_2 = \sum_{k\in J} X_k$, $I, J \subset \{1, ..., r\}$ and $I \cap J = \emptyset$, are sums of random variables $X_1, ..., X_r$ which together have a joint multinomial distribution.

## 2 Solution by Taylor series

There is a simple solution to the mean and variance of the ratio of multinomial proportions that can be derived by using the Taylor series. Formally, let a set of random variables $X_1, ..., X_r$ have a probability function

$$pr\,(X_1 = x_1, ..., X_r = x_r) = \frac{n!}{\prod_{i=1}^{r} x_i!} \prod_{i=1}^{r} p_i^{x_i},$$

where $x_i$ are non-negative integers such that $\sum x_i = n$ and $p_i$ are constants with $p_i > 0$ and $\sum p_i = 1$. The joint distribution of $X_1, ..., X_r$ is known as multinomial distribution. Let $u, v \in \{0,1\}^r$ be two binary vectors such that $\sum u_i > 0$, $\sum v_i > 0$ and $u_i v_i = 0$ for all $i$. We define

$$Z_0 = \frac{X \cdot u}{X \cdot v},$$

where $\cdot$ represents a scalar product and $X = (X_1, ..., X_r)$. Without loss of generality, we will restrict our explorations to $r = 3$ and $Z_0 = X_1/X_2$. This holds because the choice vectors $u, v$ have no common $X_i$; thus, the $X_i$s can be grouped to three disjoint sets: 1) $X_i$s selected by $u$, 2) $X_i$s selected by $v$, and 3) all others.

Also, the reader will note that the ratio $Z_0 = X_1/X_2$ can be viewed as a ratio of absolute quantities as well as a ratio of fractions or probabilities because $Z_0 = (X_1/n)/(X_2/n)$.

Before we proceed any further, observe that because of the possible zero in the denominator of $Z_0$, there is no analytical solution to the mean and variance of the ratio $Z_0$. A workaround for this problem is to rewrite this ratio using a function that does not have a singularity. Let $Z_0 = f(X_1, X_2) = X_1/X_2$ be a function of two random variables. Then, with $\mu = \left(\mu_{X_1}, \mu_{X_2}\right)$, we can use the Taylor series to approximate the function $f$ as

$$\begin{aligned} Z_0 = f(X_1, X_2) \approx & f(\mu) + \left(X_1 - \mu_{X_1}\right) \frac{\partial f}{\partial X_1}(\mu) + \left(X_2 - \mu_{X_2}\right) \frac{\partial f}{\partial X_2}(\mu) \\ & + \frac{1}{2}\left(X_1 - \mu_{X_1}\right)^2 \frac{\partial^2 f}{\partial X_1^2}(\mu) + \frac{1}{2}\left(X_2 - \mu_{X_2}\right)^2 \frac{\partial^2 f}{\partial X_2^2}(\mu) \\ & + \left(X_1 - \mu_{X_1}\right)\left(X_2 - \mu_{X_2}\right) \frac{\partial^2 f}{\partial X_1 \partial X_2}(\mu), \end{aligned}$$

from which we have

$$E(Z_0) \approx f(\mu) + \frac{1}{2}\frac{\partial^2 f}{\partial X_1^2}(\mu)\sigma_{X_1}^2 + \frac{1}{2}\frac{\partial^2 f}{\partial X_2^2}(\mu)\sigma_{X_2}^2 + \frac{\partial^2 f}{\partial X_1 \partial X_2}(\mu)\sigma_{X_1,X_2}. \tag{1}$$

Since $X_1$ and $X_2$ are terms of a random vector $X = (X_1, X_2, X_3)$ drawn from the multinomial distribution given by $(n, p_1, p_2, p_3)$, we have $\mu_{X_i} = np_i$ and $\sigma_{X_i}^2 = np_i(1 - p_i)$ for $i = 1, 2$, and $\sigma_{X_1,X_2} = -np_1p_2$. It follows easily that

$$E(Z_0) \approx \frac{p_1}{p_2} + \frac{1}{n}\left(\frac{p_1(1-p_2)}{p_2^2} + \frac{p_1}{p_2}\right) = \frac{p_1}{p_2}\left(1 + \frac{1}{np_2}\right). \tag{2}$$

For variance, we use a simpler approximation of $f$

$$f(X_1, X_2) \approx f(\mu) + \left(X_1 - \mu_{X_1}\right)\frac{\partial f}{\partial X_1}(\mu) + \left(X_2 - \mu_{X_2}\right)\frac{\partial f}{\partial X_2}(\mu),$$

from which we have

$$var(Z_0) \approx \frac{\partial f}{\partial X_1}(\mu)^2\sigma_{X_1}^2 + \frac{\partial f}{\partial X_2}(\mu)^2\sigma_{X_2}^2 + 2\frac{\partial f}{\partial X_1}(\mu)\frac{\partial f}{\partial X_2}(\mu)\sigma_{X_1,X_2}, \tag{3}$$

and finally

$$var(Z_0) \approx \frac{1}{n}\left(\frac{p_1(1-p_1)}{p_2^2} + \frac{p_1^2(1-p_2)}{p_2^3} + 2\frac{p_1^2}{p_2^2}\right) = \frac{1}{n}\left(\frac{p_1}{p_2}\right)^2\left(\frac{1}{p_1} + \frac{1}{p_2}\right). \tag{4}$$

## 3 Solution by a modified ratio

### 3.1 Definition

Let the symbols $X$, $u$, and $v$ have the same meaning as in Section 2. We define a new random variable $Z_1$ as

$$Z_1 = \frac{X \cdot u}{X \cdot v + 1}. \tag{5}$$

The $+1$ in the above definition serves to avoid zero in the denominator, and thus solves the problem with the singularity of $Z_0$. For the same reasons as in Section 2, we will restrict our explorations to $k = 3$ and $Z_1 = X_1/(X_2 + 1)$.

### 3.2 Sample space

The sample space $S_{Z_1} \subseteq \mathbb{Q}$ of the random variable $Z_1$ is limited by the sample space $S_X$ of the multinomially distributed random vector $X = (X_1, X_2, X_3)$. Therefore, if $X$ assumes values from the multinomial distribution given by $(n, p_1, p_2, p_3)$, then $Z_1$ cannot assume all rational values $a/(b+1)$ for some $a, b \in \mathbb{N}$, but only those that satisfy $a + b \leq n$ and $a, b \geq 0$. Furthermore, values $2/2$ and $4/4$ are considered identical; therefore, different outcomes of random vector $X$ may correspond with the same outcome of $Z_1$. In other words, each instance $(a, b, c)$ of $X$ corresponds with exactly one instance $a/(b+1)$ of $Z_1$, while an instance of $Z_1$ may correspond with multiple instances of $X$.

Naturally, the probability of a particular value of $Z_1$ can be determined by summing the probabilities of all (multinomial) vectors that are associated with this value. From this, it follows that if the initial multinomial probability distribution function of random vector $X$ is

$$pr\,(X_1 = a, X_2 = b, X_3 = c) = \binom{n}{a, b, c} p_1^a\, p_2^b\, p_3^c,$$

then the probability distribution function of random variable $Z_1$ is

$$pr\,(Z_1 = d) = \sum_{\substack{a,b,c \in \{0,\ldots,n\} \\ a+b+c=n \\ a/(b+1)=d}} \binom{n}{a, b, c} p_1^a\, p_2^b\, p_3^c,$$

which can be rewritten as

$$pr\,(Z_1 = d) = \sum_{b=0}^{n} \sum_{\substack{a=0 \\ a/(b+1)=d}}^{n-b} \binom{n}{b} \binom{n-b}{a} p_1^a\, p_2^b\, (1 - p_1 - p_2)^{n-a-b}.$$

### 3.3 Mean and variance

Now we can state the mean and variance of $Z_1$. The proofs of the statements can be found in the Appendix.

**Theorem 1** *Let $X = (X_1, X_2, X_3)$ be a random vector from the multinomial distribution given by $(n, p_1, p_2, p_3)$. The expected value of the random variable $Z_1$, given by* (5)*, is*

$$E(Z_1) = \frac{p_1}{p_2}\left(1 - (1 - p_2)^n\right).$$

**Theorem 2** *Let $X = (X_1, X_2, X_3)$ be a random vector from the multinomial distribution given by $(n, p_1, p_2, p_3)$, where*

$$n > \frac{1-p_2}{p_2}N + \frac{1-2p_2}{p_2}$$

*for some natural non-zero N. The variance of the random variable $Z_1$, given by* (5)*, is*

$$var(Z_1) = \left[\frac{p_1}{p_2(1-p_2)}\right]^2 \frac{\frac{1-p_2}{p_1} - 2}{n+2} + \frac{p_1}{p_2(1-p_2)} \frac{\frac{p_1}{1-p_2} - 1}{n+1} + \sum_{k=1}^{N} \frac{\left[\frac{p_1}{p_2(1-p_2)}\right]^2}{\binom{n+k+1}{k} p_2^k} \left[1 - \frac{k+2-\frac{1-p_2}{p_1}}{n+k+2}\right] + O\left(\frac{1}{n^{N+1}}\right).$$

**Corollary 1** *For $N = 1$ we have for the variance from Theorem* 2

$$var(Z_1) = \frac{1}{n}\left(\frac{p_1}{p_2}\right)^2 \left(\frac{1}{p_1} + \frac{1}{p_2}\right) + O\left(\frac{1}{n^2}\right).$$

*Observe that the formula for the variance is asymptotic in nature, and thus it may not work well for small n and certain configurations of $p_1$, $p_2$ and $p_3$. See Section* 5 *for more details.*

## 4 Approximate error of solution by a modified ratio

Let

$$Err = g(X_1, X_2) = \frac{X_1}{X_2} - \frac{X_1}{X_2+1} = \frac{X_1}{X_2(X_2+1)}$$

be a function of two random variables expressing the difference between $Z_0$ and $Z_1$. Analogous to the Eqs. (1)–(4) from Section 2 and with $f(X_1, X_2) = X_1/[X_2(X_2+1)]$, we have for the mean and variance of *Err*

$$\begin{aligned} E(Err) &\approx \frac{p_1}{p_2(1+np_2)} + \frac{p_1(1-p_2)\left(1+3np_2+3n^2p_2^2\right)}{np_2^2(1+np_2)^3} + \frac{p_1(1+2np_2)}{np_2(1+np_2)^2} \\ &= \frac{p_1\left[1+4np_2+(5-p_2)n^2p_2^2+n^3p_2^3\right]}{np_2^2(1+np_2)^3}, \end{aligned} \tag{6}$$

$$\begin{aligned} var(Err) &\approx \frac{(1-p_1)p_1}{np_2^2(1+np_2)^2} + \frac{(1-p_2)(p_1+2np_1p_2)^2}{np_2^3(1+np_2)^4} + \frac{2p_1^2(1+2np_2)}{np_2^2(1+np_2)^3} \\ &= \frac{p_1\left[p_2(1+np_2)^2 + p_1\left\{1+4np_2+(4-p_2)n^2p_2^2\right\}\right]}{np_2^3(1+np_2)^4}. \end{aligned} \tag{7}$$

It follows from the Eqs. (6) and (7) that $Z_1$ is an asymptotically ($n \to \infty$) unbiased estimator of the ratio of multinomial proportions $Z_0$. Moreover, the Eqs. (6) and (7) can be used to correct the mean and variance of the modified ratio $Z_1$ to better reflect the mean and variance of the original ratio $Z_0$. Let $Z_1^{cor} = Z_1 + Err$ be a new random variable. Since the expected value is linear, we have directly

$$\begin{aligned} E\left(Z_1^{cor}\right) &= E(Z_1) + E(Err) \approx \\ &\approx \frac{p_1}{p_2}\left(1-(1-p_2)^n\right) + \frac{p_1\left[1+4np_2+(5-p_2)n^2p_2^2+n^3p_2^3\right]}{np_2^2(1+np_2)^3}. \end{aligned}$$

For the variance, we have

$$var(Z_1^{cor}) = var(Z_1) + var(Err) + 2cov(Z_1, Err),$$

where

$$cov(Z_1, Err) = E(Z_1 \cdot Err) - E(Z_1) \cdot E(Err).$$

To approximate the value of $E(Z_1 \cdot Err)$, we use the Taylor series again, particularly Eq. (1). After some rearrangement, we get

$$\begin{aligned} E\left(\frac{X_1^2}{X_2(X_2+1)^2}\right) \approx & \frac{np_1^2}{p_2(1+np_2)^2} + \frac{(1-p_1)p_1}{p_2(1+np_2)^2} \\ & + \frac{p_1^2(1-p_2)\left(1+4np_2+6n^2p_2^2\right)}{p_2^2(1+np_2)^4} + \frac{2p_1^2(1+3np_2)}{p_2(1+np_2)^3} \\ = & \frac{p_1\left[p_2(1+np_2)^2 + p_1\left\{1+(5+2p_2)np_2+(8-p_2)n^2p_2^2+n^3p_2^3\right\}\right]}{p_2^2(1+np_2)^4} \end{aligned}$$

Thus, we can now easily calculate the value of $var\left(Z_1^{cor}\right)$ (equation omitted due to its length). In the next section, we shall discuss numerical simulations and performance of the presented formulae.

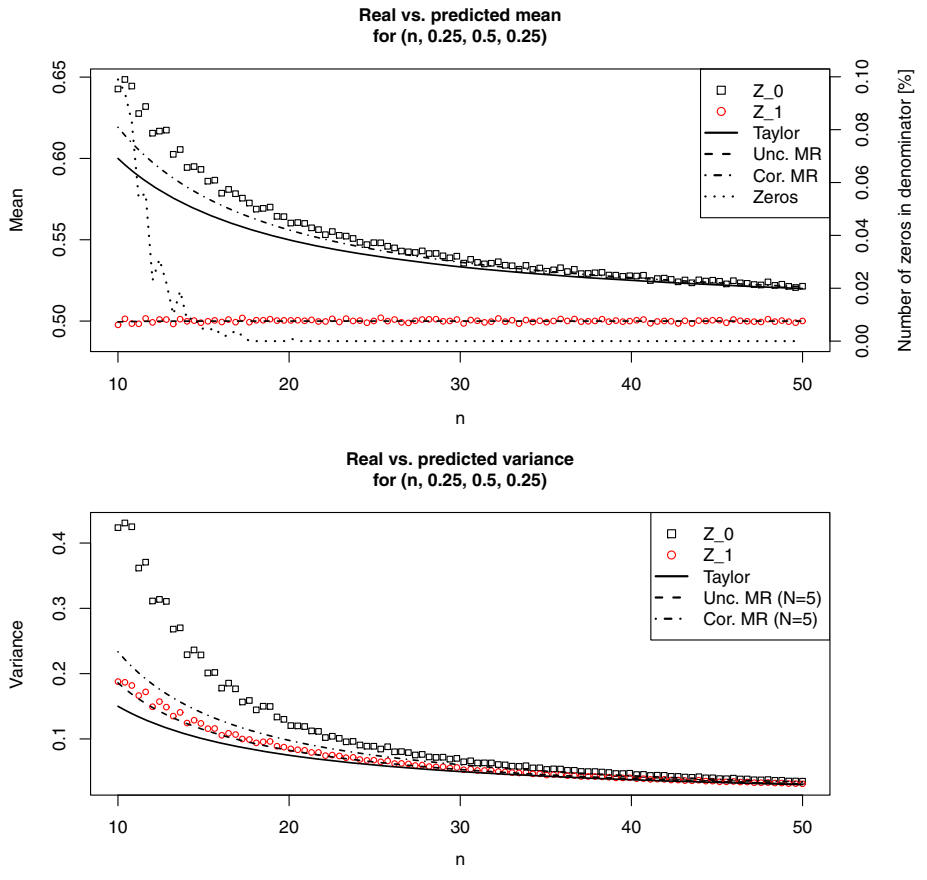


**Fig. 1** The simulation results based on the multinomial distribution given by $(n, 0.25, 0.5, 0.25)$, where $n$ ranges from 10 to 50. The mean and variance of the original ratios $Z_0$ (squares) as well as modified ratios $Z_1$ (red circles) are compared with models: the Taylor-series model (solid line), the modified ratio model (dashed line), and the corrected modified ratio model (dash-dot line). Additionally, in the upper plot, there is also information about the variable *zeros* on the secondary right axis (dots). In this case, the modified ratio model outperforms the Taylor series model for $Z_0$ data. Additionally, the uncorrected modified ratio model describes the $Z_1$ data very well

## 5 Numerical simulations

Numerical simulations were performed in the following way. We selected several multinomial distributions given by $(n, p_1, p_2, p_3)$ and for each such distribution, we sampled $10^5$ random vectors $(X_1, X_2, X_3)$. Vectors with $X_2 = 0$ were counted (variable *zeros*) and omitted from further calculations; that is, they were not replaced by new random vectors. For the vectors with $X_2 \neq 0$, we calculated the ratios $Z_0 = X_1/X_2$, while the ratios $Z_1 = X_1/(X_2 + 1)$ were calculated from all $10^5$ sampled vectors. Thus, we obtained $10^5 - zeros$ values of $Z_0$ and $10^5$ values of $Z_1$. From both sets we calculated the mean and variance of the sampled data. We compared these values with the predictions as follows below.

For the mean, we compared the means of the two data sets with the Taylor-series solution given by Eq. (2), and with the modified ratio (MR) solution given by Theorem 1 with and without the correction given by the Eq. (6).

For the variance, we compared the variances of the two data sets with the Taylor-series solution given by Eq. (4), and with the modified ratio solution given by Theorem 2 with and without the correction (the final formula for corrected variance of the modified ratio was omitted due to its length, but see Section 4 for calculation details). Note that for

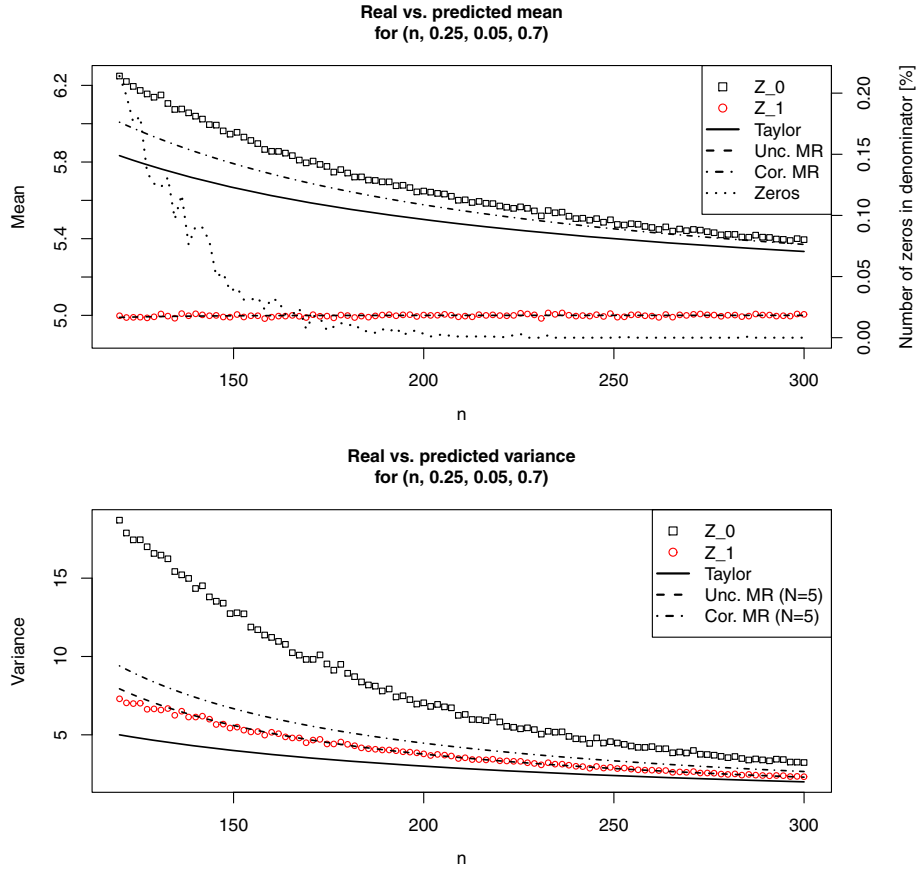


**Fig. 2** The simulation results based on the multinomial distribution given by $(n, 0.25, 0.05, 0.7)$, where $n$ ranges from 120 to 300. For the use of symbols see Fig. 1. Again, the modified ratio model outperforms the Taylor-series model for $Z_0$ data in this case, although the fit is not so close as in Fig. 1. The uncorrected modified ratio model still describes the $Z_1$ data very well

variance given by Theorem 2, we considered the case $N = 5$ so that its error $O\left(1/n^6\right)$ would not interfere with the correction.

Figure 1 shows the simulation results for the multinomial distribution given by ($n = 10, \ldots, 50, p_1 = 0.25, p_2 = 0.5, p_3 = 0.25$). The corrected modified ratio gives the best model of the mean and variance of $Z_0$. Observe also that the uncorrected modified ratio is a very precise model of $Z_1$.

In Fig. 2, when $p_2$ and $n$ are small, the discrepancy between the models and the data gets larger, although the corrected modified ratio still outperforms the Taylor-series approach. The uncorrected modified ratio is also a very good model of $Z_1$.

Figures 3 and 4 further explore the limits of the presented models. In Fig. 3, we compared the performance of the variance models in three multinomial distributions (with decreasing value of $p_2$) for various values of $N$ from Theorem 2. Note that with growing $N$, there also grows the minimal value of $n$ for which the Theorem 2 holds; therefore, the variance models start from a different $n$. It will be observed that all models have difficulty describing the initial part of the variance curve of the simulated data. However, one should keep in mind that the formula in Theorem 2 is only asymptotic.

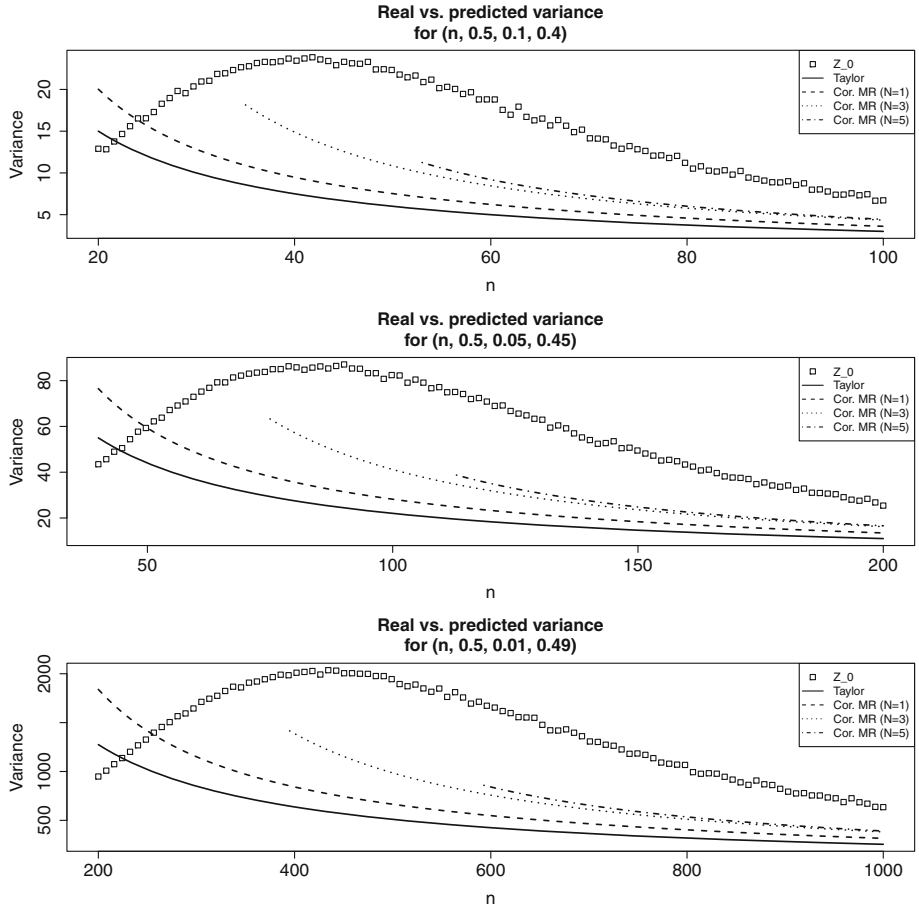


**Fig. 3** The simulation results based on three multinomial distributions and various values of $N$ from Theorem 2. Displayed are the results for variance. The simulation data for original ratios $Z_0$ (squares) are compared with models: the Taylor-series model (solid line) and the corrected modified ratio models with $N = 1$ (dashed line), $N = 3$ (dots), N=5 (dash-dot line). Observe that because of the condition on $n$ in Theorem 2, the modified ratio models do not start at the same value of $n$ for different $N$

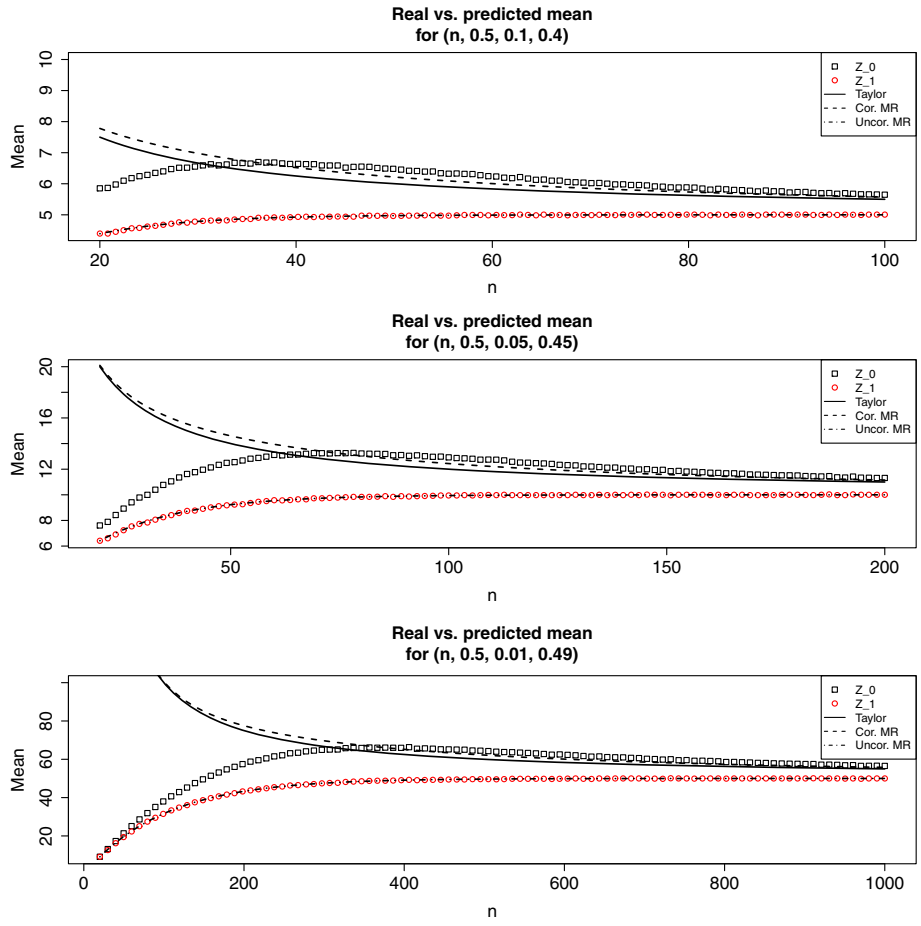


**Fig. 4** The results for mean for the data from Fig. 3. The simulation data for original ratios $Z_0$ (squares) are compared with models: the Taylor-series model (solid line) and corrected modified ratio model (dashed line). Observe the uncorrected modified ratio model (dash-dot line) which exactly models the modified ratios $Z_1$ (red circles) in all cases

In Fig. 4, we compared the models for mean on the same data as in Fig. 3. Again, for small values of $n$, the models fail to capture the real trend of the data. On a side note, the data for $Z_1$ are very well described by the uncorrected modified ratio model from Theorem 1.

The supplemental material contains a script (Additional file 1) to generate similar plots for the user-specified multinomial distribution $(n, p_1, p_2, p_3)$ and a range of $n$. Given the results from the simulation data, we encourage the reader to use this script and check whether the formulae presented in the paper will provide for a good approximation of $Z_0$ for his/hers particular multinomial distribution.

## Appendix

### Proof of Theorem 1

**Lemma 1** *Let $n \in \mathbb{N}$ and $R \in \mathbb{R}$. Then it holds*

$$\sum_{k=0}^{n} \binom{n}{k} R^k k = nR\,(1+R)^{n-1}\,.$$

*Proof* From $\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$ it directly follows that

$$\sum_{k=0}^{n} \binom{n}{k} R^k k = nR \sum_{k=0}^{n-1} \binom{n-1}{k} R^k = nR(1+R)^{n-1}.$$

□

*Proof of Theorem 1* From the definition of the expected value we have

$$E(Z_1) = \sum_{d \in S_{Z_1}} pr(Z_1 = d) \cdot d,$$

where $S_{Z_1}$ is a sample space of $Z_1$. By using

$$pr(Z_1 = d) = \sum_{b=0}^{n} \sum_{\substack{a=0 \\ a/(b+1)=d}}^{n-b} \binom{n}{b} \binom{n-b}{a} p_1^a \, p_2^b \, (1-p_1-p_2)^{n-a-b}$$

from Section 3.2, we can write

$$E(Z_1) = \sum_{d \in S_{Z_1}} \left( \sum_{b=0}^{n} \sum_{\substack{a=0 \\ a/(b+1)=d}}^{n-b} \binom{n}{b} \binom{n-b}{a} p_1^a p_2^b (1-p_1-p_2)^{n-a-b} \right) d.$$

Furthermore, because $\sum_{b=0}^{n} \sum_{a=0}^{n-b}$ enumerates all possible values of a random vector $(X_1, X_2, X_3) = (a, b, n-a-b)$ for the given $n$, it also enumerates all values of $Z_1$ including their multiplicities (see Section 3.2). Thus, we can simplify the expression of $E(Z_1)$ into

$$E(Z_1) = \sum_{b=0}^{n} \sum_{a=0}^{n-b} \binom{n}{b} \binom{n-b}{a} p_1^a p_2^b (1-p_1-p_2)^{n-a-b} \frac{a}{b+1}.$$

We rewrite this expression to separate the sums, thus obtaining

$$E(Z_1) = (1-p_1-p_2)^n \sum_{b=0}^{n} \binom{n}{b} \left( \frac{p_2}{1-p_1-p_2} \right)^b \frac{1}{b+1} \cdot \cdot \sum_{a=0}^{n-b} \binom{n-b}{a} \left( \frac{p_1}{1-p_1-p_2} \right)^a a. \tag{8}$$

Using Lemma 1, we have for (8)

$$\sum_{a=0}^{n-b} \binom{n-b}{a} \left( \frac{p_1}{1-p_1-p_2} \right)^a a = (n-b) \frac{p_1}{1-p_1-p_2} \left( \frac{1-p_2}{1-p_1-p_2} \right)^{n-b-1}.$$

By putting this back to $E(Z_1)$ and after some rearrangement of the terms, we get

$$E(Z_1) = (1-p_2)^n \left( \frac{p_1}{1-p_2} \right) \sum_{b=0}^{n} \binom{n}{b} \left( \frac{p_2}{1-p_2} \right)^b \frac{n-b}{b+1}. \tag{9}$$

We continue by splitting the following fraction into two terms

$$\frac{n-b}{b+1} = \frac{n+1}{b+1} - 1.$$

By this, the sum in (9) splits into two parts

$$E(Z_1) = A + B,$$

where

$$A = (1-p_2)^n \left(\frac{p_1}{1-p_2}\right) \sum_{b=0}^{n} \binom{n}{b} \left(\frac{p_2}{1-p_2}\right)^b \frac{n+1}{b+1},$$

$$B = (1-p_2)^n \left(\frac{p_1}{1-p_2}\right) \sum_{b=0}^{n} \binom{n}{b} \left(\frac{p_2}{1-p_2}\right)^b (-1).$$

With $\binom{n}{b}\frac{n+1}{b+1} = \binom{n+1}{b+1}$ and some rearrangement of the terms, we obtain

$$A = \frac{p_1}{p_2}\left(\frac{1}{1-p_2} - (1-p_2)^n\right),$$

and a straightforward calculation of $B$ yields

$$B = -\frac{p_1}{1-p_2}.$$

Finally, after putting $A$ and $B$ together, we get

$$E(Z_1) = A + B = \frac{p_1}{p_2} - \frac{p_1}{p_2}(1-p_2)^n = \frac{p_1}{p_2}\left(1-(1-p_2)^n\right).$$

□

**Proof of Theorem 2**

The proof of Theorem 2 relies on a series of lemmas and corollaries. For a better navigation through the proof, see Fig. 5 for the proof scheme.

**Lemma 2** *Let $n \in \mathbb{N}$ and $R \in \mathbb{R}$. Then it holds*

$$\sum_{k=0}^{n} \binom{n}{k} R^k k^2 = n(n-1)R^2(1+R)^{n-2} + nR(1+R)^{n-1}.$$

*Proof* From $\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$ and Lemma 1 it follows that

$$\begin{aligned}\sum_{k=0}^{n} \binom{n}{k} R^k k^2 &= nR \sum_{k=0}^{n-1} \binom{n-1}{k} R^k (k+1) \\ &= nR \sum_{k=0}^{n-1} \binom{n-1}{k} R^k k + nR \sum_{k=0}^{n-1} \binom{n-1}{k} R^k \\ &= n(n-1)R^2(1+R)^{n-2} + nR(1+R)^{n-1}.\end{aligned}$$

□

**Lemma 3** *Let $n \in \mathbb{N}$ and $R \in \mathbb{R}\setminus\{0\}$. Then, for any $N \in \mathbb{N}$ it holds*

$$\sum_{b=1}^{n} \binom{n}{b} \frac{R^b}{b} = \sum_{k=0}^{N} (A_{2k} - B_{2k}) + A_{2N+1},$$

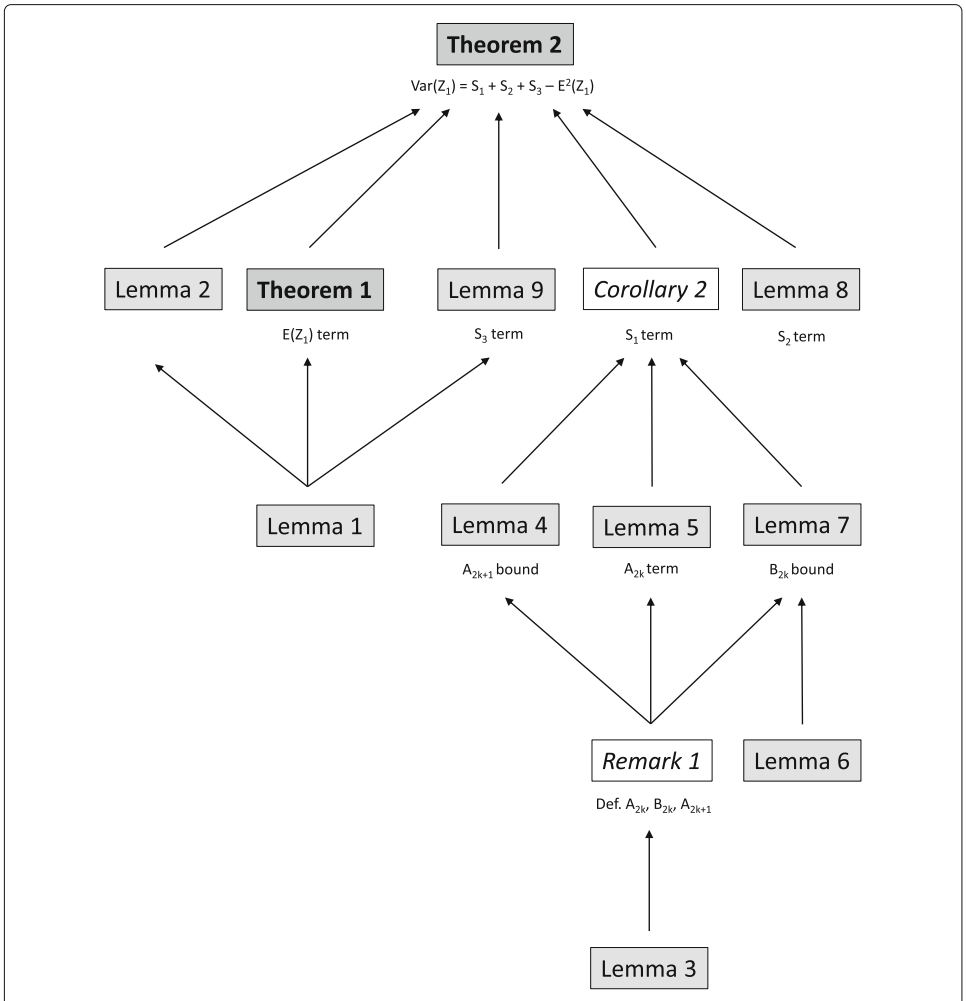


**Fig. 5** Scheme of the proof of Theorem 2

*where*

$$A_{2k} = \left(\prod_{i=1}^{k+1} \frac{1}{n+i}\right) \frac{k!}{R^{k+1}} (1+R)^{n+k+1},$$

$$B_{2k} = \left(\prod_{i=1}^{k+1} \frac{1}{n+i}\right) \frac{k!}{R^{k+1}} \sum_{b=0}^{k+1} \binom{n+k+1}{b} R^b,$$

$$A_{2k+1} = \left(\prod_{i=1}^{k+1} \frac{1}{n+i}\right) \frac{(k+1)!}{R^{k+1}} \sum_{b=k+2}^{n+k+1} \binom{n+k+1}{b} \frac{R^b}{b-(k+1)}.$$

*Proof* By induction on $N$. Let $N = 0$. Then, it follows

$$\sum_{b=1}^{n} \binom{n}{b} \frac{R^b}{b} = \sum_{b=1}^{n} \binom{n}{b} \frac{R^b}{b+1} \left(1 + \frac{1}{b}\right) = \sum_{b=1}^{n} \binom{n}{b} \frac{R^b}{b+1} + \sum_{b=1}^{n} \binom{n}{b} \frac{R^b}{b(b+1)}.$$

By using $\frac{n+1}{k+1}\binom{n}{k} = \binom{n+1}{k+1}$ and the binomial theorem, we can write

$$\begin{aligned}\sum_{b=1}^{n}\binom{n}{b}\frac{R^b}{b} &= \frac{1}{n+1}\frac{1}{R}\sum_{b=1}^{n}\binom{n+1}{b+1}R^{b+1} + \frac{1}{n+1}\frac{1}{R}\sum_{b=1}^{n}\binom{n+1}{b+1}\frac{R^{b+1}}{(b+1)-1}\\ &= \frac{1}{n+1}\frac{1}{R}\sum_{b=2}^{n+1}\binom{n+1}{b}R^{b} + \frac{1}{n+1}\frac{1}{R}\sum_{b=2}^{n+1}\binom{n+1}{b}\frac{R^b}{b-1}\\ &= A_0 - B_0 + A_1.\end{aligned}$$

The base of the induction holds. Assume that the lemma holds up to some natural $N$. We prove that it holds for $N+1$ as well. Consider the term $A_{2N+1}$. We have

$$\begin{aligned}A_{2N+1} &= \left(\prod_{i=1}^{N+1}\frac{1}{n+i}\right)\frac{(N+1)!}{R^{N+1}}\sum_{b=N+2}^{n+N+1}\binom{n+N+1}{b}\frac{R^b}{b+1}\left(1+\frac{N+2}{b-(N+1)}\right)\\ &= X_1 + X_2,\end{aligned}$$

where

$$\begin{aligned}X_1 &= \left(\prod_{i=1}^{N+1}\frac{1}{n+i}\right)\frac{(N+1)!}{R^{N+1}}\sum_{b=N+2}^{n+N+1}\binom{n+N+1}{b}\frac{R^b}{b+1},\\ X_2 &= \left(\prod_{i=1}^{N+1}\frac{1}{n+i}\right)\frac{(N+1)!}{R^{N+1}}\sum_{b=N+2}^{n+N+1}\binom{n+N+1}{b}\frac{R^b}{b+1}\frac{N+2}{b-(N+1)}.\end{aligned}$$

Furthermore, by the same trick with the binomial coefficient as above, we rewrite the terms $X_1$ and $X_2$ as

$$\begin{aligned}X_1 &= \left(\prod_{i=1}^{N+1}\frac{1}{n+i}\right)\frac{(N+1)!}{R^{N+1}}\frac{1}{n+N+2}\frac{1}{R}\sum_{b=N+2}^{n+N+1}\binom{n+N+2}{b+1}R^{b+1},\\ X_2 &= \left(\prod_{i=1}^{N+1}\frac{1}{n+i}\right)\frac{(N+1)!}{R^{N+1}}\frac{1}{n+N+2}\frac{1}{R}\sum_{b=N+2}^{n+N+1}\binom{n+N+2}{b+1}\frac{R^{b+1}(N+2)}{(b+1)-1-(N+1)}.\end{aligned}$$

After some rearrangement, we finally get (again using the binomial theorem)

$$\begin{aligned}X_1 &= \left(\prod_{i=1}^{N+2}\frac{1}{n+i}\right)\frac{(N+1)!}{R^{N+2}}\sum_{b=N+3}^{n+N+2}\binom{n+N+2}{b}R^{b} = A_{2(N+1)} - B_{2(N+1)},\\ X_2 &= \left(\prod_{i=1}^{N+2}\frac{1}{n+i}\right)\frac{(N+2)!}{R^{N+2}}\sum_{b=N+3}^{n+N+2}\binom{n+N+2}{b}\frac{R^b}{b-(N+2)} = A_{2(N+1)+1}.\end{aligned}$$

□

**Remark 1** *We will often use Lemma* 3 *with* $n+1$ *instead of* $n$*. Therefore, we restate the Lemma* 3 *with this change. Let* $n \in \mathbb{N}$ *and* $R \in \mathbb{R}\backslash\{0\}$*. Then, for any* $N \in \mathbb{N}$ *it holds*

$$\sum_{b=1}^{n+1}\binom{n+1}{b}\frac{R^b}{b} = \sum_{k=0}^{N}(A_{2k} - B_{2k}) + A_{2N+1},$$

*where*

$$A_{2k} = \left(\prod_{i=2}^{k+2} \frac{1}{n+i}\right) \frac{k!}{R^{k+1}} (1+R)^{n+k+2},$$

$$B_{2k} = \left(\prod_{i=2}^{k+2} \frac{1}{n+i}\right) \frac{k!}{R^{k+1}} \sum_{b=0}^{k+1} \binom{n+k+2}{b} R^b,$$

$$A_{2k+1} = \left(\prod_{i=2}^{k+2} \frac{1}{n+i}\right) \frac{(k+1)!}{R^{k+1}} \sum_{b=k+2}^{n+k+2} \binom{n+k+2}{b} \frac{R^b}{b-(k+1)}.$$

**Lemma 4** *Let $p_1, p_2 \in (0,1)$ be some real constants. Let $k, n$ be some non-zero natural numbers. Let $A_{2k+1}$ be the term from Remark 1. Furthermore, let $R = p_2/(1-p_2)$, and let*

$$A = (n+1)n\left(\frac{p_1}{1-p_2}\right)^2 + (n+1)\frac{p_1}{1-p_2},$$

$$D = \frac{(1-p_2)^n}{n+1}\frac{1-p_2}{p_2}.$$

*Then, for $\alpha \in [1, k+2]$, it holds*

$$ADA_{2k+1} \leq \alpha \frac{n}{(k+2)\binom{n+k+3}{k+2}} \frac{p_1}{p_2^{k+3}(1-p_2)} \left(\frac{p_1}{1-p_2} + \frac{1}{n}\right) = O\left(\frac{1}{n^{k+1}}\right).$$

*Proof* First of all, for $\alpha \in [1, k+2]$ we have

$$A_{2k+1} = \alpha \left(\prod_{i=2}^{k+3} \frac{1}{n+i}\right) \frac{(k+1)!}{R^{k+2}} \sum_{b=k+3}^{n+k+3} \binom{n+k+3}{b} R^b.$$

This follows easily by applying the inequality

$$\frac{k+2}{b+1} \geq \frac{1}{b-(k+1)} \geq \frac{1}{b+1}$$

to the term $A_{2k+1}$ from Remark 1, which holds for any natural $b, k$ except for pairs $b = k+1$ (in our case $b > k+1$). We can see this by solving the inequality

$$\frac{1+x}{b+1} \geq \frac{1}{b-(k+1)}$$

for $x$. By this, we get an upper and lower bound on the term $A_{2k+1}$, which differ by a multiplicative constant $k+2$. Finally, the lemma follows by extending the summation through index $b$ in the term $A_{2k+1}$ to a full range from 0 to $n+k+3$, by applying the binomial theorem and some simple rearrangement of the terms. The $O$ bound follows from the fact that $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$. □

**Lemma 5** *Let $p_1, p_2 \in (0,1)$ be some real constants. Let $k, n$ be some non-zero natural numbers. Let $A_{2k}$ be the term from Remark 1. Furthermore, let $R = p_2/(1-p_2)$, and let*

$$A = (n+1)n\left(\frac{p_1}{1-p_2}\right)^2 + (n+1)\frac{p_1}{1-p_2},$$

$$D = \frac{(1-p_2)^n}{n+1}\frac{1-p_2}{p_2}.$$

*Then, it holds*

$$ADA_{2k} = \frac{\left(\frac{p_1}{p_2(1-p_2)}\right)^2}{\binom{n+k+1}{k} p_2^k} \left(1 - \frac{k+2-\frac{1-p_2}{p_1}}{n+k+2}\right).$$

*Proof* The lemma follows easily by a straightforward multiplication of the terms $A$, $D$ and $A_{2k}$, and some rearrangement of the terms. □

The following lemma is an extension of one borrowed from Graham et al. (1994).

**Lemma 6** *Let $0 < \alpha < R/(1+R)$ for some real $R > 0$. Then, it holds*

$$\sum_{k \le \alpha n} \binom{n}{k} R^k = R^m 2^{nH(\alpha) - \frac{1}{2}\lg n + O(1)},$$

*where $m = \lfloor \alpha n \rfloor$ and*

$$H(\alpha) = \alpha \lg \frac{1}{\alpha} + (1-\alpha) \lg \frac{1}{1-\alpha}.$$

*Proof* First of all, we have

$$\frac{\binom{n}{k-1} R^{k-1}}{\binom{n}{k} R^k} = \frac{k}{n-k+1}\frac{1}{R} \le \frac{\alpha n}{n-\alpha n+1}\frac{1}{R} < \frac{\alpha}{1-\alpha}\frac{1}{R}.$$

Let $m = \lfloor \alpha n \rfloor = \alpha n - \epsilon$. It holds

$$\begin{aligned}\binom{n}{m} R^m < \sum_{k \le m} \binom{n}{k} R^k &< \binom{n}{m} R^m \left(1 + \frac{\alpha}{1+\alpha}\frac{1}{R} + \left(\frac{\alpha}{1-\alpha}\frac{1}{R}\right)^2 + \dots\right) \\ &= \binom{n}{m} R^m \frac{(1-\alpha)R}{(1-\alpha)R - \alpha}\end{aligned}$$

because

$$\frac{\alpha}{1-\alpha}\frac{1}{R} < 1,$$

which follows from $\alpha < R/(1+R)$. Thus,

$$\sum_{k \le m} \binom{n}{k} R^k = \binom{n}{m} R^m O(1).$$

By Stirling's approximation, we have

$$\begin{aligned}\log \binom{n}{m} &= -\frac{1}{2}\log n - (\alpha n - \epsilon)\log\left(\alpha - \frac{\epsilon}{n}\right) - ((1-\alpha)n + \epsilon)\log\left(1 - \alpha + \frac{\epsilon}{n}\right) + O(1) \\ &= -\frac{1}{2}\log n - n\alpha \log \alpha - n(1-\alpha)\log(1-\alpha) + O(1),\end{aligned}$$

and the lemma follows. □

**Lemma 7** *Let $p_1, p_2 \in (0,1)$ be some real constants. Let $k, n$ be some non-zero natural numbers such that*

$$n > \frac{1-p_2}{p_2}k + \frac{1-2p_2}{p_2}.$$

*Let $B_{2k}$ be the term from Remark 1. Furthermore, let $R = p_2/(1-p_2)$, and let*

$$A = (n+1)n\left(\frac{p_1}{1-p_2}\right)^2 + (n+1)\frac{p_1}{1-p_2},$$
$$D = \frac{(1-p_2)^n}{n+1}\frac{1-p_2}{p_2}.$$

*Then, it holds*

$$ADB_{2k} = n(1-p_2)^n \frac{p_1}{p_2(1-p_2)}\left(p_1 + \frac{1-p_2}{n}\right)\frac{2^{k+1}O(1)}{(k+1)(n+k+2)^{\frac{1}{2}}} = O\left(n^{\frac{1}{2}}(1-p_2)^n\right).$$

*Proof* Let $\alpha = (k+1)/(n+k+2)$. One can easily verify that $\alpha < R/(1+R) = p_2$ because of the choice of $n$. Thus, we can apply Lemma 6 to the sum from the term $B_{2k}$. From this, it follows that

$$\sum_{b=0}^{k+1}\binom{n+k+2}{b}\left(\frac{p_2}{1-p_2}\right)^b = \left(\frac{p_2}{1-p_2}\right)^{k+1} 2^{(n+k+2)H(\alpha)-\frac{1}{2}\lg(n+k+2)+O(1)}, \tag{10}$$

where

$$H(\alpha) = \alpha\lg\frac{1}{\alpha} + (1-\alpha)\lg\frac{1}{1-\alpha}.$$

Moreover, for $H(\alpha)$ we have

$$H(\alpha) = \frac{k+1}{n+k+2}\lg\left(\frac{2(n+k+2)}{k+1}\right) - O\left(\frac{1}{n^2}\right),$$

which follows from

$$\lg(1-\alpha) = -\sum_{i=1}^{\infty}\frac{\alpha^i}{i}.$$

Plunging this into (10), we get

$$\sum_{b=0}^{k+1}\binom{n+k+2}{b}\left(\frac{p_2}{1-p_2}\right)^b = \left(\frac{p_2}{1-p_2}\right)^{k+1}\frac{\left(\frac{2(n+k+2)}{k+1}\right)^{k+1}O(1)}{(n+k+2)^{\frac{1}{2}}}.$$

With this, we can write for the whole $B_{2k}$ term from Remark 1

$$B_{2k} = \frac{\left(\frac{2(n+k+2)}{k+1}\right)^{k+1}O(1)}{\binom{n+k+1}{k}(n+k+2)^{\frac{3}{2}}} \le \frac{2^{k+1}\binom{n+k+2}{k+1}O(1)}{\binom{n+k+1}{k}(n+k+2)^{\frac{3}{2}}} = \frac{2^{k+1}O(1)}{(k+1)(n+k+2)^{\frac{1}{2}}} \tag{11}$$

because $\left(\frac{n}{k}\right)^k \le \binom{n}{k}$. Similarly, with $\binom{n}{k} < \left(\frac{ne}{k}\right)^k$, we have for $B_{2k}$

$$B_{2k} \ge \frac{\left(\frac{2}{e}\right)^{k+1}O(1)}{(k+1)(n+k+2)^{\frac{1}{2}}},$$

if we use

$$\binom{n+k+1}{k}(n+k+2)^{\frac{3}{2}} = \binom{n+k+2}{k+1}(k+1)(n+k+1)^{\frac{1}{2}}.$$

Thus, we have

$$B_{2k} = \frac{2^{k+1}O(1)}{(k+1)(n+k+2)^{\frac{1}{2}}},$$

and the lemma easily follows by multiplying $B_{2k}$ with the term $AD$. □

**Corollary 2** *Let $p_1, p_2 \in (0,1)$ be some real constants. Let $n, N$ be some non-zero natural numbers such that*

$$n > \frac{1-p_2}{p_2}N + \frac{1-2p_2}{p_2}.$$

*Let $A_{2k}, B_{2k}$, $k = 0, ..., N$, and $A_{2N+1}$ be terms from Remark* 1. *Furthermore, let $R = p_2/(1-p_2)$, and let*

$$A = (n+1)n\left(\frac{p_1}{1-p_2}\right)^2 + (n+1)\frac{p_1}{1-p_2},$$
$$D = \frac{(1-p_2)^n}{n+1}\frac{1-p_2}{p_2}.$$

*Then, it holds*

$$AD\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b\frac{1}{b} = \left(\frac{p_1}{p_2(1-p_2)}\right)^2\sum_{k=0}^{N}\frac{1-\frac{k+2-\frac{1-p_2}{p_1}}{n+k+2}}{\binom{n+k+1}{k}p_2^k} + O\left(\frac{1}{n^{N+1}}\right).$$

*Proof* Follows from Lemmas 4, 5 and 7. □

**Lemma 8** *Let $p_1, p_2 \in (0,1)$ be some real constants and $n$ some non-zero natural number. Let*

$$B = (2n+1)\left(\frac{p_1}{1-p_2}\right)^2 + \frac{p_1}{1-p_2},$$
$$D = \frac{(1-p_2)^n}{n+1}\frac{1-p_2}{p_2}.$$

*Then, it holds*

$$BD\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b = 2\left(\frac{p_1}{1-p_2}\right)^2\frac{1}{p_2} + \frac{1}{n+1}\frac{p_1}{p_2(1-p_2)}\left(1-\frac{p_1}{1-p_2}\right) + O\left((1-p_2)^n\right).$$

*Proof* Straightforward by binomial theorem. □

**Lemma 9** *Let $p_1, p_2 \in (0,1)$ be some real constants and $n$ some non-zero natural number. Let*

$$C = \left(\frac{p_1}{1-p_2}\right)^2,$$
$$D = \frac{(1-p_2)^n}{n+1}\frac{1-p_2}{p_2}.$$

*Then, it holds*

$$CD\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b b = \left(\frac{p_1}{1-p_2}\right)^2.$$

*Proof* Straightforward by Lemma 1 and binomial theorem. □

*Proof of Theorem 2* The variance of the random variable $Z_1$ can be calculated as

$$var(Z_1) = E(Z_1^2) - E^2(Z_1).$$

By Theorem 1, we have

$$E(Z_1) = \frac{p_1}{p_2}\left(1 - (1-p_2)^n\right).$$

So, we only need to determine the value of $E\left(Z_1^2\right)$. From the definition of the expected value, we have

$$E\left(Z_1^2\right) = \sum_{b=0}^{n}\sum_{a=0}^{n-b}\binom{n}{b}\binom{n-b}{a}p_1^a p_2^b(1-p_1-p_2)^{n-a-b}\left(\frac{a}{b+1}\right)^2 = (1-p_1-p_2)^n V_1 V_2,$$

where

$$V_1 = \sum_{b=0}^{n}\binom{n}{b}\left(\frac{p_2}{1-p_1-p_2}\right)^b\left(\frac{1}{b+1}\right)^2,$$

$$V_2 = \sum_{a=0}^{n-b}\binom{n-b}{a}\left(\frac{p_1}{1-p_1-p_2}\right)^a a^2.$$

By application of Lemma 2 to $V_2$, we obtain

$$E(Z_1^2) = (1-p_2)^n\sum_{b=0}^{n}\binom{n}{b}\left(\frac{p_2}{1-p_2}\right)^b\left(\frac{1}{b+1}\right)^2 W,$$

$$W = (n-b)(n-b-1)\left(\frac{p_1}{1-p_2}\right)^2 + (n-b)\frac{p_1}{1-p_2}.$$

By using the equality

$$\binom{n}{b}\left(\frac{1}{b+1}\right)^2 = \binom{n+1}{b+1}\frac{1}{n+1}\frac{1}{b+1}$$

and adjustment of the summation borders, we get

$$E\left(Z_1^2\right) = \frac{(1-p_2)^n}{n+1}\cdot\frac{1-p_2}{p_2}\cdot\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b\frac{1}{b}W,$$

$$W = (n-b+1)(n-b)\left(\frac{p_1}{1-p_2}\right)^2 + (n-b+1)\frac{p_1}{1-p_2}.$$

Next, we split the term $W$ according to powers of $b$, thus obtaining

$$W = A - Bb + Cb^2,$$

where

$$A = (n+1)n\left(\frac{p_1}{1-p_2}\right)^2 + (n+1)\frac{p_1}{1-p_2},$$

$$B = (2n+1)\left(\frac{p_1}{1-p_2}\right)^2 + \frac{p_1}{1-p_2},$$

$$C = \left(\frac{p_1}{1-p_2}\right)^2.$$

If we set

$$D = \frac{(1-p_2)^n}{n+1}\cdot\frac{1-p_2}{p_2},$$

then we can write

$$E\left(Z_1^2\right) = D\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b\left(\frac{A}{b} - B + Cb\right) = S_1 + S_2 + S_3,$$

where

$$S_1 = AD\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b\frac{1}{b},$$
$$S_2 = -BD\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b,$$
$$S_3 = CD\sum_{b=1}^{n+1}\binom{n+1}{b}\left(\frac{p_2}{1-p_2}\right)^b b,$$

and by Corollary 2 ($S_1$) and Lemmas 8 ($S_2$) and 9 ($S_3$) we get

$$\begin{aligned}E(Z_1^2) = \sum_{k=0}^{N}\left(\frac{p_1}{p_2(1-p_2)}\right)^2\frac{1}{\binom{n+k+1}{k}p_2^k}\left(1 - \frac{k+2-\frac{1-p_2}{p_1}}{n+k+2}\right) - \\ - 2\left(\frac{p_1}{1-p_2}\right)^2\frac{1}{p_2} - \frac{1}{n+1}\frac{p_1}{p_2(1-p_2)}\left(1 - \frac{p_1}{1-p_2}\right) \\ + \left(\frac{p_1}{1-p_2}\right)^2 + O\left(\frac{1}{n^{N+1}}\right).\end{aligned}$$

The rest of the proof follows from adding the term $-E^2(Z_1)$ to the derived expression for $E\left(Z_1^2\right)$, separating the term for $k = 0$ from the rest of the sum, and simple rearrangement of the resulting terms. □

## Additional file

**Additional file 1:** A script written in language R to perform custom numerical simulations and produce graphical output. (R 10 kb)

**Acknowledgements**
This contribution is the result of implementation of the project *REVOGENE – Research centre for molecular genetics* (ITMS 26240220067) supported by the Research & Developmental Operational Programme funded by the European Regional Development Fund.

**Authors' contributions**
All authors contributed equally to the research. FD wrote the manuscript. JG prepared the figures. All authors read and approved the final manuscript.

**Competing interests**
The authors declare that they have no competing interests.

## Publisher's Note



**Author details**
[1]Geneton s.r.o., Galvaniho 7, 82104 Bratislava, Slovakia. [2]Slovak Centre of Scientific and Technical Information, Lamacska cesta 7315/8A, 81104 Bratislava, Slovakia. [3]Comenius University, Faculty of Natural Sciences, Ilkovicova 3278/6, 84104 Bratislava, Slovakia. [4]Comenius University Faculty of Mathematics, Physics and Informatics, Mlynska dolina, 84248 Bratislava, Slovakia. [5]Comenius University, Science Park, Ilkovicova 8, 84104 Bratislava, Slovakia.

Received: 9 August 2017 Accepted: 3 January 2018
Published online: 18 January 2018

**References**


Aho, K, Bowyer, RT: Confidence intervals for ratios of proportions: implications for selection ratios. Methods Ecol. Evol. **6**(2), 121–132 (2015)

Alghamdi, N: Confidence intervals for ratios of multinomial proportions (2015). Master's thesis, University if Nebraska at Omaha

Basu, A, Lochner, RH: On the distribution of the ratio of two random variables having generalized life distributions. Technometrics. **13**(2), 281–287 (1971)

Bonett, DG, Price, RM: Confidence intervals for a ratio of binomial proportions based on paired data. Stat. Med. **25**(17), 3039–3047 (2006)

Chiu, RW, Chan, KA, Gao, Y, Lau, VY, Zheng, W, Leung, TY, Foo, CH, Xie, B, Tsui, NB, Lun, FM, et al: Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of dna in maternal plasma. Proc. Natl. Acad. Sci. **105**(51), 20458–20463 (2008)

Culley, TM, Wallace, LE, Gengler-Nowak, KM, Crawford, DJ: A comparison of two methods of calculating gst, a genetic measure of population differentiation. Am. J. Bot. **89**(3), 460–465 (2002)

Fieller, E: The distribution of the index in a normal bivariate population. Biometrika. **24**, 428–440 (1932)

Frishman, F: On the arithmetic means and variances of products and ratios of random variables (1971). Technical report, DTIC Document

Geary, R: The frequency distribution of the quotient of two normal variates. J. R. Stat. Soc. **93**(3), 442–446 (1930)

Goodman, LA: On simultaneous confidence intervals for multinomial proportions. Technometrics. **7**(2), 247–254 (1965)

Graham, RL, Knuth, DE, Patashnik, O: Concrete Mathematics: A Foundation for Computer Science, 2nd edn, p. 492. Addison-Wesley Longman Publishing Co., Inc., Boston (1994). exercise 42

Hinkley, DV: On the ratio of two correlated normal random variables. Biometrika. **56**(3), 635–639 (1969)

Koopman, P: Confidence intervals for the ratio of two binomial proportions. Biometrics. **40**, 513–517 (1984)

Korhonen, PJ, Narula, SC: The probability distribution of the ratio of the absolute values of two normal variables. J. Stat. Comput. Simul. **33**(3), 173–182 (1989)

Lau, TK, Chen, F, Pan, X, Pooh, RK, Jiang, F, Li, Y, Jiang, H, Li, X, Chen, S, Zhang, X: Noninvasive prenatal diagnosis of common fetal chromosomal aneuploidies by maternal plasma dna sequencing. J. Matern. Fetal Neonatal Med. **25**(8), 1370–1374 (2012)

Marsaglia, G: Ratios of normal variables. J. Stat. Softw. **16**(4), 1–10 (2006)

Mekic, E, Sekulovic, N, Bandjur, M, Stefanovic, M, Spalevic, P: The distribution of ratio of random variable and product of two random variables and its application in performance analysis of multi-hop relaying communications over fading channels. Przegl. Elektrotechniczny. **88**(7A), 133–137 (2012)

Minarik, G, Repiska, G, Hyblova, M, Nagyova, E, Soltys, K, Budis, J, Duris, F, Sysak, R, Bujalkova, MG, Vlkova-Izrael, B, *et al*: Utilization of benchtop next generation sequencing platforms ion torrent pgm and miseq in noninvasive prenatal testing for chromosome 21 trisomy and testing of impact of in silico and physical size selection on its analytical performance. PloS ONE. **10**(12), 0144811 (2015)

Nadarajah, S: On the product and ratio of laplace and bessel random variables. J. Appl. Math. **2005**(4), 393–402 (2005)

Nadarajah, S, Kotz, S: On the ratio of pearson type vii and bessel random variables. Adv. Decis. Sci. **2005**(4), 191–199 (2005)

Nadarajah, S, Kotz, S: On the product and ratio of gamma and weibull random variables. Econ. Theory. **22**(2), 338–344 (2006)

Nelson, W: Statistical methods for the ratio of two multinomial proportions. Am. Stat. **26**(3), 22–27 (1972)

Pham-Gia, T: Distributions of the ratios of independent beta variables and applications. Commun. Stat. Theory Methods. **29**(12), 2693–2715 (2000)

Piegorsch, WW, Richwine, KA: Large-sample pairwise comparisons among multinomial proportions with an application to analysis of mutant spectra. J. Agric. Biol. Environ. Stat. **6**(3), 305–325 (2001)

Piper, J, Rutovitz, D, Sudar, D, Kallioniemi, A, Kallioniemi, O-P, Waldman, FM, Gray, JW, Pinkel, D: Computer image analysis of comparative genomic hybridization. Cytometry. **19**(1), 10–26 (1995)

Poorter, H, Garnier, E: Plant growth analysis: an evaluation of experimental design and computational methods. J. Exp. Bot. **47**(9), 1343–1351 (1996)

Press, SJ: The t-ratio distribution. J. Am. Stat. Assoc. **64**(325), 242–252 (1969)

Price, RM, Bonett, DG: Confidence intervals for a ratio of two independent binomial proportions. Stat. Med. **27**(26), 5497–5508 (2008)

Provost, S: On the distribution of the ratio of powers of sums of gamma random variables. Pak. J. Stat. **5**, 157–174 (1989)

Quesenberry, CP, Hurst, D: Large sample simultaneous confidence intervals for multinomial proportions. Technometrics. **6**(2), 191–195 (1964)

Sakamoto, H: On the distributions of the product and the quotient of the independent and uniformly distributed random variables. Tohoku Math. J. First Ser. **49**, 243–260 (1943)

Sehnert, AJ, Rhees, B, Comstock, D, de Feo, E, Heilek, G, Burke, J, Rava, RP: Optimal detection of fetal chromosomal abnormalities by massively parallel dna sequencing of cell-free fetal dna from maternal blood. Clin. Chem. **57**(7), 1042–1049 (2011)

Van Kempen, G, Van Vliet, L: Mean and variance of ratio estimators used in fluorescence ratio imaging. Cytometry. **39**(4), 300–305 (2000)