# AN ANALYSIS ON THE ACCURACY OF CONFIDENCE INTERVALS TO DETERMINE ALLELE FREQUENCY AS A FUNCTION OF SAMPLE SIZE

Gonçalo Maria

Faculdade de Ciências da Universidade de Lisboa

## Abstract

In population genetics it is often necessary to determine an allele's frequency, here we present an analysis of the margins of error of various confidence intervals as a function of sample size n for a significance level $\alpha$ in the two allele case. We aim to determine not only which methods provide the least margins of error but also measure the accuracy of these methods based on sample size.

## Introduction

### Hardy-Weinberg Equilibrium

The Hardy-Weinberg Equilibrium (HWE) (Hardy, 1908; Weinberg, 1908) states that the allele and genotype frequencies for a given locus will remain constant from generation to generation in the absence of outside factors such as: genetic drift, mate choice, natural selection etc. Therefore it is often useful to observe whether a population adheres to the expected HWE in regards to a specific locus to determine if outside pressures such as the ones listed above have an impact on the population's allele frequencies.

Here we will analyze tests for the simplest case where the number of alleles for the locus being observed is equal to two, for which the expected frequencies according to HWE for a frequency of allele $a$ is $p_a$ and a frequency of allele $A$ is $p_A = 1 - p_a$ are as follows:

$$p_{AA} = p_A^2$$
$$p_{aa} = p_a^2$$
$$p_{aA} = 2p_a p_A$$

´ Where $p_{ij}$ for $i, j \in \{a, A\}$ denotes the frequency of individuals with genotype $ij$ in a population.

### Testing for Hardy-Weinberg proportions

Many potential confidence intervals have been proposed for gauging allele frequencies such as the Wald confidence interval proposed by Gillespe in 2004; confidence intervals derived for the binomial distribution by Agresti and Coull in 1998 or the Clopper Pearson interval derived by Clopper and Pearson in 1934. However we will focus on the confidence intervals provided by Tak Fung and Kevin Keenan in 2014 for a given significance level $\alpha$ and compare them to proportion confidence intervals with the same significance level.

We analyse this method for the proportion confidence interval and for the confidence intervals defined by Tak Fung and Kevin Keenan in "Confidence Intervals for Population Allele Frequencies: The General Case of Sampling from a Finite Diploid Population of Any Size", for each method for constructing these intervals we generate samples and analyze the means of the margins of error for sample sizes 1 to 80 (except for one of the T. Fung, K. Keenan intervals cases), in the latter case in a population of 100 as the method described by T. Fung and K. Keenan requires the computation of combinations which can lead to very lengthy execution time.

Bibliography

Code Repository

## Methods

For comparing the adequacy of these two methods as a function of sample size we assume the most widely used significance level of $\alpha = 0.05$. First for each $n$ we simulate 10 samples with a $\hat{y}_{i,n}$: observed number of individuals with a copy of the $i$ allele in the sample, simulated as $p_{i,n} \times n$ where $p_{i,n}$ is a random number between zero and one, we then run another set of simulations (for n from 1 to 45 and population 100 for the T. Fung and K. Keenan confidence intervals) for $y_{i,n}$ values simulated as $p + d$ where $p$ is $\frac{1}{4}$ and $d$ is a randomly generated deviation, this last simulation is an attempt to understand how well the confidence intervals perform for HWE proportions expected of a Mendelian trait.

After running these simulations we plot the mean margins of error for each value $n$ as a function of $n$ for the two different methods of building confidence intervals and for the two different methods for simulating $\hat{y}_{i,n}$.

### Proportion Confidence Intervals

In order to compute the Proportion Confidence Intervals calculated as:

$$\hat{y}_{i,n} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{y}_{i,n}(\hat{y}_{i,n}-1)}{n}} \leq p_i \leq \hat{y}_{i,n} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{y}_{i,n}(\hat{y}_{i,n}-1)}{n}}$$

Python 3.8 was used, the code is available in the QR code.

### T. Fung, K. Keenan Confidence Intervals

To compute the Confidence Intervals described by T. Fung and K. Keenan we determined for each simulated sample:

$$[L_\alpha, U_\alpha] = [Min(P_{i,H_0,\alpha}), Max(P_{i,h_0,\alpha})]$$

where

$$P_{i,H_0,\alpha} = \{p_{i,0} : \hat{y}_{i,N} \in B_\alpha\}$$

with $B_\alpha$ the acceptance region determined as:

$$B_\alpha = \{y_{i,N,\frac{\alpha}{2}}, y_{i,N,1-\frac{\alpha}{2}}\}$$

and $y_{i,N,\frac{\alpha}{2}}$ and $y_{i,N,1-\frac{\alpha}{2}}$ represent the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the $y_{i,N}$ distribution of allele frequency for allele i in a population with size n as calculated per the methods highlighted by T. Fung and K. Keenan.

All these computations were run in Python 3.8 using a translation of the R code provided by T. Fung and K. Keenan in their article, the translated python code is available in the QR code.

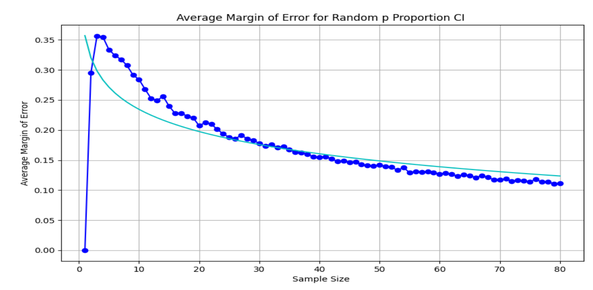
### Analyzing the data

After obtaining the scatter plots for each method of simulating the $\hat{y}_{i,N}$ and for each method for determining confidence intervals we fitted a logarithmic regression model to the data using python software.

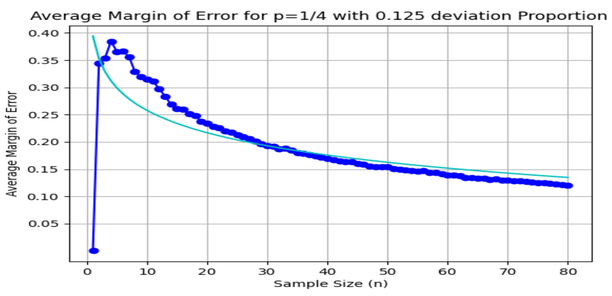
## Results

### Proportion Confidence Intervals

For the Proportion Confidence Intervals the following results were obtained:



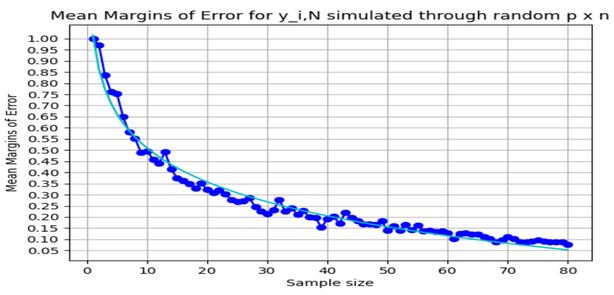

For randomly generated allele frequency $f(a) = p$ where $p$ is randomly generated between -1 and 1. In which the Logarithmic regression takes the expression: $y = 0.3632549 - 0.05483106log(x)$;



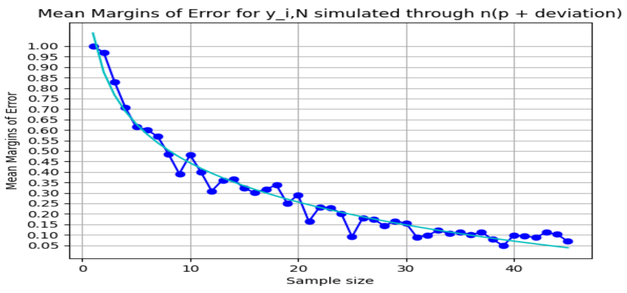

For allele frequency $f(a) = p + d$ where $p = \frac{1}{4}$ and $d$ is randomly generated between $-\frac{p}{2}$ and $\frac{p}{2}$ In which the Logarithmic regression takes the expression: $y = 0.40257321 - 0.06055756log(x)$.

### T. Fung and K.Keenan Confidence Intervals

For the T. Fung and K. Keenan Confidence Intervals the following results were obtained:




For randomly generated allele frequency $f(a) = p$ where $p$ is randomly generated between -1 and 1. In which the Logarithmic regression takes the expression: $y = 1.01541293 - 0.21969835log(x)$;




For number of observed individuals with one copy of the allele $n(p + d)$ where $p = \frac{1}{4}$ and $d$ is randomly generated between $-\frac{n}{10}$ and $\frac{n}{10}$ In which the Logarithmic regression takes the expression: $y = 1.06253648 - 0.26907592log(x)$.

## Discussion and Conclusion

The trend lines for the logarithmic regressions suggest that for both methods, after sample size approximately 35 the rate at which the margins of error diminish slows considerably. It is also important to note that given the high complexity of computation for the T. Fung and K. Keenan intervals and given that the Proportion Confidence Intervals have only a slightly bigger mean margin of error for the case where $\hat{y}_{i,N}$ does not follow the expected values for Mendelian inheritance: in cases involving the processing of many samples it might be preferable to use the Proportion Confidence Interval method.

However it should also be noted that if an allele is suspected to follow HWE with frequencies akin to those expected of Mendelian traits then the best approach would be to calculate the T. Fung and K. Keenan Intervals as these provide the fastest diminishing mean margins of error with a mean consistently below any of the other methods above sample size 30.

Contacto: GoncaloMaria2001@gmail.com