Al-Mukhtar Journal of Basic Sciences 21 (2): 58-62, 2023
Doi:

Research Article **Open Access**

Approximate confidence intervals for the Population Proportion based on linear model

Nuri E. Mohamed

***Corresponding author:** n.almadni@wau.edu.ly
Department of Mathematics, Faculty of Science-, Wadi Shati University, Gurdha, Libya

Received:
10 September 2023

Accepted:
29 November 2023

Publish online:
31 December 2023

Abstract

When the linear model errors are non-normal, one might be interested in making inferences concerning proportion. The goal of this article is to construct approximate confidence intervals for the proportion founded on the supposed linear model which covers a true value of a proportion that is close to a specific nominal value of the level of significant.

Keywords: Linear model, central limit theorem, slutsky's lemma, asymptotic distribution, confidence intervals**.**

INTRODUCTION

In this paper study was derived the asymptotic confidence intervals by the *z*-quantile and by the *t*-quantile, to construct approximate confidence intervals of the proportion based on linear models.

Then it was introduced to the used model:
suppose that the sum random variable X_i is a Poisson distribution can be splits into two separate Poisson random variables Y_i, Z_i with means λ_1, λ_2 respectively. It means.,

$X_i \sim (\lambda_1, \lambda_2)$, and since E($Y_i$) = Var($Y_i$) = λ_1, and $E(Z_i)$ = Var(Z_i) = λ_2, consequently

E(X_i) = Var(X_i) = $\lambda_1 + \lambda_2$. Further,
P($Yi \mid Xi$) ~ Bin(X, p), where $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$, $X_i > 0$ (B(n, p) denotes the
Binomial distibution with parameters n, p).

The remainder of the paper was organized as follows: Description of the assumed model is given in Sec 2. Estimation of linear model parameter were given in Sec. 3. Section 4 to provide confidence intervals for the proportion.

The Linear model

Suppose, there are m observations of two relevant components of data, i.e., (Y_i, X_i); $i = 1, \ldots, m$ are $m \times 2$ dimensional observed data since every observational item refrenced by the subscript i, as well as, $Y_i, X_i > 0$; $\forall i = 1, \ldots, m$. The i^{th} observation represents unit, for example:

$Xi \equiv$number of units for a product i; $i = 1, \ldots, m$

Yi ≡number of damaged units for a product *i; i* = 1,….,*m*.

Consider the univariate linear model $Y_i = x_i p + \epsilon_i$, *i* = 1,….,*m*, with the following assumptions:

$E(Y_i) = x_i p$, $E(\epsilon_i) = 0$, also by variance relational to *xi* (*xi* is fixed variable), i.e,
$Var(Y_i) = Var(\epsilon_i) = \sigma^2 x_i$, compressing the model in vector form (for the i^{th} observation) yields:
$Y_i = Xp+\epsilon$, (*i* = 1,….,*m*), where

$Y = (Y_1, \dots Y_m)^T$, and the $m \times 1$ design vector.

$X = (x_1, \dots, x_m)^T$, then the heteroscedastic errors $(\epsilon_1, \dots \epsilon_m)^T = \epsilon$, through the expectations, $E(\epsilon) = \mathbf{0}_m$, and Var(∈) = σ²W, where **0**m =$(0, \dots, 0)^T$, and W = diag (x_i). The single model was weighted by the linear transformation

$$A_i Y_i = A_i\, x_i p + A_i\ \epsilon_i,$$

$$\tilde{Y}_i = \tilde{x}_i p + \tilde{\epsilon}_i,\ i = 1, \dots, m \qquad [2.1]$$

where, $A_i = \frac{1}{\sqrt{x_i}}$, given that $x_i > 0, \tilde{Y}_i = A_i\, Y_i = \frac{Y_i}{\sqrt{x_i}}, \tilde{x}_i = A_i\, x_i = \sqrt{x_i}$, $\tilde{\epsilon}_i = A_i\ \epsilon_i = \frac{\epsilon_i}{\sqrt{x_i}}$, it follows that $E(\tilde{\epsilon}_i) = 0$, and $Var(\tilde{\epsilon}_i) = \sigma^2$, $\forall i = 1, \dots, m$ (*homoscedastic errors*), as well as $Cov(\epsilon) = \sigma^2 I_m = Cov(\tilde{\boldsymbol{Y}})$.such that *Im* is an Identity Matrix of elements $m \times m$, also the weighted response vector $\tilde{\boldsymbol{Y}} = \left(\tilde{Y}_1, \dots, \tilde{Y}_m\right)^T$, and the weighted design vector $\tilde{\boldsymbol{X}} = (\tilde{x}_1, \dots, \tilde{x}_m)^T$, as well as the weighted error vector $\tilde{\epsilon} = \left(\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_m{}^T\right)$, where

$\sigma^2 = E\left(\tilde{Y}_i - \tilde{X}_i p\right)^2$

Estimation in linear models
Agreeing whether the demonstrate blunders are homoscedastic or heteroscedastic blunders we estimate the proportion *p*.

Heteroscedasticity
Since, error of the vector of the non-weighted ideal has covariance that is the variance proportional to the known invertible diagonal matrix *W;* so, it is the Generalized Least Squares Estimator, that is further the Best linear unbiased estimator. The covariance structure is given by

$$\operatorname{cov}(\epsilon) = \begin{bmatrix} \sigma^2 x_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 x_m \end{bmatrix} = \sigma^2 W$$

x_i are fixed, $i = 1, \dots, m$ and $W^{-1}W = I_m$, as well as $X^T W = 1^T{}_m$, where $1_m = (1, \dots 1)^T$. So, we have

$$\hat{p}_{GLS} = (X^T(\sigma^2 W)^{-1} X)^{-1} X^T (\sigma^2 W)^{-1} Y = (X^T W^{-1} X)^{-1} X^T W^{-1} Y$$

Homoscedasticity

It is familiar and defined very well; that the Weighted Least Squares Estimator is the Best linear unbiased estimator, in addition, due to the reason that, the weighted errors are homoscedastic therefore, the Weighted Least Squares Estimator functional to the Model 2.1 concludes in the OLS Estimator, hence is also the greatest LUE , according to Gauss-Markov's theorem *(1-3),* i.e., $\hat{p}_{WLS}$= $\hat{p}$. Since, $\mathrm{cov}(\tilde{\epsilon}) = \sigma^2 I_m$,

then

$$\hat{p}_{WLS} = \left(\tilde{X}^T(\sigma^2 I_m)^{-1}\tilde{X}\right)^{-1}\tilde{X}^T(\sigma^2 I_m)^{-1}\tilde{Y} = (X^T I_m X)^{-1}\tilde{X}^T I_m \tilde{Y}$$

$$= \left(\sum_{i=1}^{m}\left(\sqrt{x_i}\right)^2\right)^{-1}\sum_{i=1}^{m}\sqrt{x_i}\tilde{Y}_i = \frac{\sum_{i=1}^{m}\sqrt{x_i}\tilde{Y}_i}{\sum_{i=1}^{m}x_i} = \frac{\sum_{i=1}^{m}Y_i}{\sum_{i=1}^{m}x_i} = \hat{p},$$

$$\sqrt{x_i}\tilde{Y} = Y_i.$$

Asymptotic normal for the ratio $\hat{p}_m$

It supposed that the random errors $\tilde{\epsilon}_i$ are not Normally Distributed nonetheless are independently identical distributed random variables, $i = 1, \dots, m$, i.e.,$E(\tilde{\epsilon}_i) = 0$, and $\mathrm{Var}(\tilde{\epsilon}_i) = 0$. Furthermore, under a positive conditions on the project **X** we can demonstrate that in huge sample sizes, $\hat{p}$ obeys the asymptotic normal distribution.

And more additionally conditions on the couple of observations ***Xi, Yi*** are required, called (***Xi, Yi***) are independently identical distributed pairs of random variables, plus E(*Xi*) exist ⇒

($\tilde{X}_i$, $\tilde{Y}_i$**)** are then independently identical distributed random variables, furthermore $E\left(\tilde{X}^2{}_i\right)$ exists

$$i = 1, \dots, m\ , \tilde{X}_i = \sqrt{X_i}, \tilde{Y}_i = \frac{Y_i}{\sqrt{x_i}}, X_i > 0\ .$$

To arrive to the asymptotic distribution, one rephrases first the estimator $\hat{p}_m$as

$$\hat{p}_m = \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} = \left(\sum_{i=1}^{m}\tilde{X}_i\,\tilde{X}_i\right)^{-1}\sum_{i=1}^{m}\tilde{X}_i\,\tilde{Y}_i$$

$$= \left(\sum_{i=1}^{m}X_i\right)^{-1}\sum_{i=1}^{m}\tilde{X}_i\,(\tilde{X}_i\,p + \tilde{\epsilon}_i) = \left(\sum_{i=1}^{m}X_i\right)^{-1}\left(\sum_{i=1}^{m}\tilde{X}_i\,p + \sum_{i=1}^{m}\tilde{X}_i\,\tilde{\epsilon}_i\right)$$

$$= p + \left(\sum_{i=1}^{m}X_i\right)^{-1} + \sum_{i=1}^{m}\tilde{X}_i\,\tilde{\epsilon}_i \qquad (3.1)$$

Consequently

$$\sqrt{m}(\hat{p}_m - p) = \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^{-1} \frac{1}{\sqrt{m}}\left(\sum_{i=1}^{m} \sqrt{\tilde{X}_i}\,\tilde{\epsilon}_i\right) = \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^{-1} \frac{1}{\sqrt{m}}\left(\sum_{i=1}^{m} \sqrt{X_i}\tilde{\epsilon}_i\right) \quad (3.2)$$

The asymptotic of the equation 3.2, requests to confirm, the denominator in 3.2 is reliable, and the numerator submits the CLT. It direct to see (by the LLN)

$$= \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^{-1} \xrightarrow{p} \left(E(X_i)\right)^{-1},$$

Given that, E(*Xi*) > 0, and

$$= \left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)^{-1} \to \mu^{-1}, \qquad \mu \text{ is a constant.}$$

Along with the numerator

$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\sqrt{X_i}\tilde{\epsilon}_i \xrightarrow{D} N\left(0, \sigma^2 E(X_i)\right)$, where, the marginal or asymptotic variance

$$\operatorname{cov}\left(\sqrt{X_i}\tilde{\epsilon}_i, \sqrt{X_i}\tilde{\epsilon}_i\right) = \operatorname{Var}\left(\sqrt{X_i}\tilde{\epsilon}_i\right) = E\left(X_i \operatorname{Var}(\tilde{\epsilon}_i \setminus \tilde{X}_i)\right) + \operatorname{Var}\left(E\left(\sqrt{X_i}\tilde{\epsilon}_i \setminus \tilde{X}_i\right)\right)$$
$$= \sigma^2 E(X_i).$$

As a result, following the use of the Slutsky's lemma (see [Knight,. 2000], pp. 119-120), the equation 3.2 can be rewritten as

$$\sqrt{m}(\hat{p}_m - p) \xrightarrow{D} N\left(0, \sigma^2 E(X_i)\left(E(X_i)\right)^{-2}\right) \equiv N\left(0, \sigma^2\left(E(X_i)\right)^{-1}\right) \quad (3.3)$$

Approximate Confidence Intervals for the Population Proportion *p*

From 3.3, the asymptotic variance,

$$\operatorname{Var}\left(\sqrt{m}(\hat{p}_m)\right) = E\left(\operatorname{Var}\sqrt{m}(\hat{p}_m) \setminus \boldsymbol{X}^T\right) + \operatorname{Var}\left(E\sqrt{m}(\hat{p}_m) \setminus \boldsymbol{X}^T\right) = \frac{\sigma^2}{E(X_i)},$$

$$\text{as } \operatorname{Var}\left(E\sqrt{m}(\hat{p}_m) \setminus \boldsymbol{X}^T\right) = 0\ , \boldsymbol{X} = (X_1, X_2, \dots, X_m)^T.$$

conclude that , the estimated $(1-\alpha)\%$ asymptotic confidence interval for the amount *p* is given by

$\left[\hat{p}_m \pm z_{1-\frac{\alpha}{2}}\, s.e(\hat{p}_m)\right]$, where, the standard error of $\hat{p}_m$

$$,s.e(\hat{p}_m) = \frac{s_m}{\sum_{i=1}^{m} X_i}\ , \qquad s^2{}_m = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \hat{p}\,\tilde{X}_i\right)^2.$$

Further and since $\bar{X}_m$ and $s^2{}_m$ are consistent estimators for E(*Xi*) and σ^2 respectively, it follows that consistent estimator of the $\mathrm{Var}((\hat{p}_m))$ is $\frac{s^2{}_m}{\sum_{i=1}^{m} X_i}$ similarly, such as

$$\frac{\hat{p}_m - p}{\frac{s^2{}_m}{\sum_{i=1}^{m} X_i}} \sim t_{m-1} \xrightarrow{D} N(0,1)$$

Henceforth, the interval that is safety bounds provided by

$$\left[\hat{p}_m \pm t_{\left(m-1,1-\frac{\alpha}{2}\right)}\, s.e(\hat{p}_m)\right],$$

is the recommended more conservative confidence interval for *p*, where $s.e(\hat{p}_m) = \frac{s_m}{\sum_{i=1}^{m} x_i}$, as well as $t_{\left(m-1,1-\frac{\alpha}{2}\right)}$ is $\left(1-\frac{\alpha}{2}\right)$ percentile of the student t distribution with $(m-1)$ degrees of freedom.

CONCLUSION

Results obtained from this article two confidence intervals first is the asymptotic confidence interval and the second is the extra conservative asymptotic confidence interval for the population proportion which can give more reliable intervals to cover the true population proportion P. Hence we considered two confidence intervals, the asymptotic (following the normal quintile) in addition the proposed conservative (with the adjusted t-quintile) confidence intervals.

Duality of interest: The authors declare that they have no duality of interest associated with this manuscript..

Author contributions: The author did all the work related to the manuscript, including designing the research, collecting information, formulating theories and proofs, and preparing the entire paper.

Funding: There is no funding to support this manuscript

REFERENCES

A . Sen and M. Srivastava. Regression Analysis: Theory, Methods, and Application. *Springer-Verlage*, New York, Inc, (1990).

H. Stock, and W.C. Waston. Introduction to econometrics. Pearson Education, Inc, pp. 588-591, (2003) .

K.C. Knight,. Mathematical Statistics. Chapman & Hall/CRC. Texts in Statistical Science, pp. 120-149, (2000) .