6-2021

A Geometric Approach to Conditioning and the Search for Minimum Variance Unbiased Estimators

David L. Farnsworth
Rochester Institute of Technology

James E. Marengo
Rochester Institute of Technology

A Geometric Approach to Conditioning and the Search for Minimum Variance Unbiased Estimators

James E. Marengo, David L. Farnsworth

School of Mathematical Sciences, Rochester Institute of Technology, Rochester, New York, USA
Email: jemsma@rit.edu, dlfsma@rit,edu

Abstract

Our purpose is twofold: to present a prototypical example of the conditioning technique to obtain the best estimator of a parameter and to show that this technique resides in the structure of an inner product space. The technique uses conditioning of an unbiased estimator on a sufficient statistic. This procedure is founded upon the conditional variance formula, which leads to an inner product space and a geometric interpretation. The example clearly illustrates the dependence on the sampling methodology. These advantages show the power and centrality of this process.

Keywords

Conditional Variance Formula, Conditioning, Geometric Representation, Minimum Variance Estimator, Rao-Blackwell Theorem, Sufficient Statistic, Unbiased Estimator

1. Introduction

We are given a coin that has probability θ of coming up heads when tossed once, where θ is an unknown real number in the interval (0, 1). We wish to estimate θ. Typically, we toss this coin a number of times, which is fixed in advance, and use the proportion of heads that appear as an estimator of θ. This approach not only makes sense intuitively, but also is optimal in that this estimator is unbiased and has the smallest possible variance based on the fixed number of tosses. It is the minimum variance unbiased estimator (MVUE) of θ [1]. The sample proportion is also the maximum likelihood estimator (MLE) of θ, because it maximizes the likelihood of obtaining the observed result of these coin tosses [1] [2].

Suppose that we proceed in a different way. We perform a "do until" experi-

ment. Instead of fixing the number of tosses in advance, we toss this coin until we obtain heads for the nth time, where n is fixed. The recorded data are $x_1, x_2, \cdots, x_n$, where x_1 is the number of tosses it takes to obtain heads for the first time, and for $2 \le k \le n$, x_k is the additional number of tosses it takes to obtain the kth occurrence of heads after we have obtained heads $k - 1$ times. These data are realizations of independent random variables $X_1, X_2, \cdots, X_n$, each of which has the geometric distribution with probability mass function (pmf)

$$p(x;\theta) = (1-\theta)^{x-1} \theta \quad \text{for} \quad x = 1, 2, 3, \cdots$$

[1] [2]. Our goal is to find the MVUE of θ based on this alternative data.

We use this example to illustrate how conditioning an unbiased estimator on a sufficient statistic can be used to find a MVUE. This shows how we can discover the best estimator by using the very important concept of conditioning. The example demonstrates that this is not simply of theoretical interest. Finding appropriate estimators is a consequential problem in inference, decision-making, and data reduction. In practical terms, an advantage of this method of using conditioning in this manner is that it is guaranteed to yield minimum variance and unbiased estimators. While they might be considered by some to be intuitive, other techniques, such as the method of moments, may not produce estimators with those desirable properties [1] [2].

2. The Minimum Variance and Unbiased Estimator (MVUE)

One way to find an unbiased estimator of θ for the geometric distribution and the data $x_1, x_2, \cdots, x_n$ is to start with X_1. Because $\Pr(X_1 = 1) = \theta$, the estimator

$$u(X_1) = \begin{cases} 1 & \text{if } X_1 = 1 \\ 0 & \text{if } X_1 > 1 \end{cases}$$

is an unbiased estimator of θ. However, this estimator ignores the values of $X_2, X_3, \cdots, X_n$, which makes it suspect. Indeed, it is not the MVUE of θ unless $n = 1$.

The random variable $Y = \sum_{k=1}^{n} X_k$ is the total number of tosses of the coin that it takes to obtain n heads. It has the negative binomial distribution with pmf

$$\Pr(Y = y) = \binom{y-1}{n-1} (1-\theta)^{y-n} \theta^n \quad \text{for} \quad y = n, n+1, \cdots$$

[1] [2]. The variable Y is a sufficient statistic for θ in the sense that the conditional, joint distribution of ($X_1, X_2, \cdots, X_n$), given the value of Y, does not depend upon θ, as we now demonstrate. For $x_1, x_2, \cdots, x_n \in \{0, 1\}$ and $y = \sum_{k=1}^{n} x_k$,

$$\Pr(X_k = x_k \text{ for } k = 1, 2, \cdots, n | Y = y)$$

$$= \frac{\prod_{k=1}^{n} \Pr(X_k = x_k)}{\Pr(Y = y)} = \frac{(1-\theta)^{\sum_{k=1}^{n} x_k - n} \theta^n}{\binom{y-1}{n-1}(1-\theta)^{y-n} \theta^n} = \frac{1}{\binom{y-1}{n-1}}. \tag{1}$$

when $y \ne \sum_{k=1}^{n} x_k$, the probability is zero.

The denominator of (1) is the number of ordered n-tuples of positive integers that sum to y. We conclude that, given $Y = y$, the conditional distribution of $(X_1, X_2, \cdots, X_n)$ is uniformly distributed on the set of all such n-tuples. The distribution does not depend upon θ. This means that, as long as the experimenters retain the value Y, they may discard the data $(x_1, x_2, \cdots, x_n)$ without losing any information about θ. The statistic Y is therefore sufficient or good enough, because it contains all the knowledge about the value of θ that is available in the data. In other words, Y drains the data of all the useful information that it has to convey about the value of θ.

The estimator

$$v(Y) = E\left(u(X_1)|Y\right)$$

is unbiased for θ, because

$$E\left(v(Y)\right) = E_Y\left(E_{X_1}\left(u(X_1)|Y\right)\right) = E\left(u(X_1)\right) = \theta.$$

Thus, $u(X_1)$ and $v(Y)$ are both unbiased estimators of θ which can be computed from the data $(x_1, x_2, \cdots, x_n)$.

According to the conditional variance formula

$$Var\left(u(X_1)\right) = E\left(Var\left(u(X_1)|Y\right)\right) + Var\left(E\left(u(X_1)|Y\right)\right) \qquad (2)$$

[1] [3] [4], so

$$Var\left(v(Y)\right) = Var\left(E\left(u(X_1)|Y\right)\right) \le Var\left(u(X_1)\right). \qquad (3)$$

This reasoning says that the MVUE of θ must be a function of Y, because we can use the conditional variance formula to condition any unbiased estimator of θ, which is not a function of Y, on Y itself to obtain another unbiased estimator of θ, which is a function of Y having a smaller variance. This is the content of the Rao-Blackwell theorem [3] [4].

There is only one function of Y that is an unbiased estimator of θ, which is demonstrated as follows. Suppose that there are two such estimators $v_1(Y)$ and $v_2(Y)$, then

$$E\left(v_1(Y)\right) = \theta = E\left(v_2(Y)\right)$$

and

$$0 = E\left(v_1(Y) - v_2(Y)\right) = \theta^n \sum_{y=n}^{\infty} \left(v_1(y) - v_2(y)\right)\binom{y-1}{n-1}(1-\theta)^{y-n}.$$

Because the series of powers of $1 - \theta$ is identically zero for $0 < \theta < 1$, all of its coefficients must be zero. Therefore, $v_1(y) = v_2(y)$ for $y = n, n+1, \cdots$, and the uniqueness is established. Thus, $v(Y) = E\left(u(X_1)|Y\right)$ must be the unique MVUE of θ.

To compute the estimator, proceed as follows:

$$E\left(u(X_1)|Y = y\right) = \sum_{x_1=1}^{\infty} u(x_1)\Pr\left(X_1 = x_1|Y = y\right)$$

$$= \Pr\left(X_1 = 1 \middle| Y = y\right)$$

$$= \frac{\Pr\left(X_1 = 1 \text{ and } Y = y\right)}{\Pr\left(Y = y\right)}$$

$$= \frac{\Pr\left(X_1 = 1\right)\Pr\left(\sum_{k=2}^{n} X_k = y - 1\right)}{\Pr\left(Y = y\right)}$$

where the independence of $X_1, X_2, \cdots, X_n$ yields the last equality. Further,

$$E\left(u\left(X_1\right) \middle| Y = y\right) = \frac{\theta \binom{y-2}{n-2}\left(1-\theta\right)^{(y-1)-(n-1)} \theta^{n-1}}{\binom{y-1}{n-1}\left(1-\theta\right)^{y-n} \theta^n} = \frac{n-1}{y-1}.$$

Hence, the MVUE of θ is

$$\hat{\theta} = \frac{n-1}{Y-1}.$$

3. Geometric Interpretation

There is a geometric interpretation of the conditional variance formula (2) that can provide a deeper understanding of these results. There is a long tradition of showing that variables and statistics reside in geometric spaces. Herr [5] gives a history of these representations. See [6] [7] [8] for other examples. The concept of projection, which is the essence of the present geometric approach, is widely used elsewhere, for example in machine learning. This approach places the process of computing MVUEs in the framework of the universally important concept of an inner product space and thus provides a deeper mathematical insight into this process.

The set of random variables on a fixed probability space which have finite second moment is a Hilbert space $\mathbb{L}^2$, where its inner product $\langle \bullet, \bullet \rangle$ and metric d are defined for points $U, V \in \mathbb{L}^2$ as

$$\langle U, V \rangle = E\left(UV\right)$$

and

$$d\left(U, V\right) = \sqrt{E\left(\left(U - V\right)^2\right)}$$

[9]. The conditional variance formula (2) is the "Pythagorean Theorem" in $\mathbb{L}^2$ as

$$d^2\left(u\left(X_1\right), E\left(u\left(X_1\right)\right)\right)$$
$$= d^2\left(u\left(X_1\right), E\left(u\left(X_1\right)\middle|Y\right)\right) + d^2\left(E\left(u\left(X_1\right)\middle|Y\right), E\left(u\left(X_1\right)\right)\right), \tag{4}$$

because

$$d^2\left(u\left(X_1\right), E\left(u\left(X_1\right)\middle|Y\right)\right) = E\left(\left(u\left(X_1\right) - E\left(u\left(X_1\right)\middle|Y\right)\right)^2\right)$$
$$= E_Y\left(E_{X_1}\left(\left(u\left(X_1\right) - E\left(u\left(X_1\right)\middle|Y\right)\right)^2 \middle| Y\right)\right) = E\left(Var\left(u\left(X_1\right)\middle|Y\right)\right)$$

and

$$d^2\left(E\left(u\left(X_1\right)\middle|Y\right),E\left(u\left(X_1\right)\right)\right)$$
$$=E\left(\left(E\left(u\left(X_1\right)\middle|Y\right)-E\left(u\left(X_1\right)\right)\right)^2\right)=Var\left(E\left(u\left(X\right)\middle|Y\right)\right),$$

by employing $E_Y\left(E_{X_1}\left(u\left(X_1\right)\middle|Y\right)\right)=E\left(u\left(X_1\right)\right)$. Equation (4) makes clear that $E\left(u\left(X_1\right)\middle|Y\right)$ is the orthogonal projection of $u\left(X_1\right)$ onto the subspace of random variables that are functions of Y, as the representational illustration in **Figure 1** shows.

If the unbiased estimator $u\left(X_1\right)$ is not a function of Y, then $d\left(u\left(X_1\right),E\left(u\left(X_1\right)\middle|Y\right)\right)>0$, and from (4)

$$d\left(E\left(u\left(X_1\right)\middle|Y,\right)E\left(u\left(X_1\right)\right)\right)<d\left(u\left(X_1\right),E\left(u\left(X_1\right)\right)\right).$$

That is,

$$Var\left(E\left(u\left(X_1\right)\middle|Y\right)\right)<Var\left(u\left(X_1\right)\right),$$

as in (3).

Because the hypotenuse of a right triangle is its longest side, **Figure 1** makes it clear that the variance of $v(Y)$ is strictly less than the variance of $u(X_1)$.

4. Conclusions

Using a prototypical example, we have shown how the powerful technique of conditioning can be used to find the minimum variance unbiased estimator (MVUE) of a parameter. The data collection technique and the observations' probability distribution determine the MVUE. Oftentimes, these MVUEs are not simply standard statistics, like means and variances, as the example illustrates. The geometry of the conditional variance formula shows how the minimum variance estimator is obtained as a projection.

Future endeavors might involve further development of the geometrical representation of this technique. Also, the example reveals how a feature of a
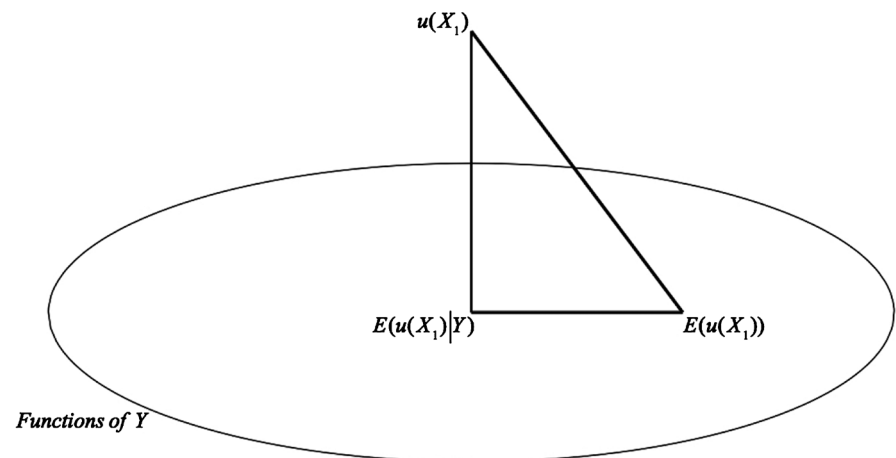

Figure 1. $E\left(u\left(X_1\right)\middle|Y\right)$ as the result of an orthogonal projection in the Hilbert space L^2.

physical population can be viewed as a parameter in different probability distributions, depending upon the sampling methods. In the example, θ is in a Bernoulli distribution and a geometric distribution. Topics for potential study are such choices, the ease of the sampling methods, the required sample sizes or efficiency, and the formulations of the MVUEs.

Acknowledgements

The authors want to thank an anonymous referee for many insightful comments.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Hogg, R.V., McKean, J.W. and Craig, A.T. (2018) Introduction to Mathematical Statistics. 8th Edition, Pearson, Boston.

[2] Devore, J.L. (2016) Probability and Statistics for Engineering and the Sciences. 9th Edition, Cengage, Boston.

[3] Peña, E.A. and Rohatgi, V. (1994) Some Comments about Sufficiency and Unbiased Estimation. *American Statistician*, **48**, 242-243.
https://doi.org/10.1080/00031305.1994.10476067

[4] Ross, S.M. (2019) Introduction to Probability Models. 12th Edition, Academic Press, London. https://doi.org/10.1016/B978-0-12-814346-9.00006-8

[5] Herr, D.G. (1980) On the History of the Use of Geometry in the General Linear Model. *American Statistician,* **34**, 43-47.
https://doi.org/10.1080/00031305.1980.10482710

[6] Farnsworth, D.L. (2000) The Geometry of Statistics. *College Mathematics Journal,* **31**, 200-204. https://doi.org/10.1080/07468342.2000.11974143

[7] Wood, G.R. and Saville, D.J. (2002) A New Angle on the t-Test. *Journal of the Royal Statistical Society, Series D* (*The Statistician*), **51**, 99-104.
https://doi.org/10.1111/1467-9884.00301

[8] Saville, D.J. and Wood, G.R. (2011) Statistical Methods: A Geometric Primer. CreateSpace, Scotts Valley.

[9] Rudin, W. (1986) Real and Complex Analysis. 3rd Edition, McGraw-Hill, New York.