BACI

Before-After-Control-Impact Power Analysis For Several Related Populations (Variance Known)

May 23, 2022

Richard A. Hinrichsen

Caveat: This study design tool is for an idealized power analysis built upon several simplifying assumptions (Table 1). For a specific design, a more accurate portrayal of power may require changing these assumptions and the underlying equations. (For example, assuming that variance will be estimated instead of known.) Therefore, this analysis should be treated as a rough guide to power.

INTRODUCTION

Currently there are many watershed projects underway in the Columbia Basin to determine the effects of various management actions on salmon survival. For example, there are a series of intensively monitored watersheds (IMW**s**) being established for the purpose of better understanding how salmon respond to approaches to restore habitat. When these projects are developed with a rigorous study design, it may be possible to identify the effectiveness of restoration and other management actions. This analysis was motivated by the need to design these studies so that they have a good chance of detecting significant survival rate changes when they occur. Using this tool can give an investigator a rough idea of the number of years to run an study and determine what statistical power can be expected based on what is known about the

variance-covariance structures for survival among the salmon populations studied. The framework for the analysis given here, although developed with salmon in mind, fits into the framework of the Before-After-Control-Impact (BACI) design. Such BACI-type designs find application beyond Columbia River salmon survival (Osenberg and Schmitt 1996).

An *a priori* power analysis is developed for a BACI-type design aimed at estimating a change in survival for several populations. The design includes a Before period where all populations receive no treatment followed by an After period where only the treatment populations receive treatment. This is a generalization of the BACI-type design where the control population and impact population are sampled one time before and one time after the treatment (Green 1979, Osenberg and Schmitt 1996). The assumptions for the analysis are given in Table 1. I assumed that in the absence of treatment, all populations have a common mean log survival. Because this assumption and others may not hold in practice, this analysis should be treated as a rough guide.

My goal is to demonstrate the probability of detecting an effect on survival when several related populations with a common mean survival are used in a BACI-type design. To accomplish this goal, I describe the study design in a statistically rigorous way, set up the likelihood function, and then use maximum likelihood theory to estimate power. Power is the probability of rejecting the null hypothesis of "no treatment effect" when it is false.

Table 1.—Assumptions used in power analysis[1].

A1	The observations of log(survival) follow a multivariate normal distribution.
A2	There is no serial dependence.
A3	All populations have a common mean log(survival) before treatment.
A4	After treatment, the control populations continue to have the same common mean as exhibited in the Before years, and the treatment populations also have a common mean, but shifted by an amount (the effect size) that is the same for all treatment populations.
A5	The measurement errors in log(survival) follow a multivariate normal distribution and the errors are independent of the error due to actual year-to-year environmental variability.

A6 The estimator of the treatment effect is a maximum likelihood estimate.

A7 The variance-covariance matrix representing the error in log(survival) is known.

[1] These are assumptions for an idealized design. For a specific application, a more accurate study design may require changing these assumptions and the underlying equations. Therefore, this analysis should be treated as a rough guide to power.

The R-code in Appendix A contains the R-code that implements this power analysis with the added assumptions that the variances in log(survival) are equal for all populations and the correlations in log(survival) are equal for each pair of populations. This is the intraclass covariance structure studied by R.A. Fisher (1925). The R-code for implementing this power analysis may be found in Appendix A (Venerables et al. 2010).

METHODS

To conduct the power analysis, I formulated a probability model and derived maximum likelihood estimators of the unknown parameters (Mood et al. 1974). These estimators were then used as the basis for testing the null hypothesis of "no treatment" effect. Power was then calculated as the probability that the null hypothesis is rejected.

The model.—I assumed that mean log(survival) before treatment was the same for each population and equal to μ_1. After treatment, the mean log(survival) of the treatment populations shifts by the amount δ for the treatment populations while the control populations continue to have a mean log(survival) of μ_1. I also assumed that year-to-year variability in log(survival) and measurement error followed a multivariate normal distribution with variance $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_y + \boldsymbol{\Sigma}_m$, where $\boldsymbol{\Sigma}_y$ is the variance-covariance matrix that describes year-to-year variability in the absence of measurement error, and $\boldsymbol{\Sigma}_m$ represents the variance-covariance matrix of the measurement error. In this implementation of the BACI design, the variance-covariance matrix, $\boldsymbol{\Sigma}$, was assumed to be known.

Maximum likelihood estimators.—To derive maximum likelihood estimators, an expression for the likelihood function is needed. For the model described above, the log-likelihood function is

$$l(\boldsymbol{\theta},\boldsymbol{\Sigma}) = C + (n/2)\ln|\boldsymbol{\Sigma}^{-1}| - (1/2)\sum_{t=1}^{n_1}(\boldsymbol{x}_t - [\boldsymbol{e} \quad \mathbf{0}]\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - [\boldsymbol{e} \quad \mathbf{0}]\boldsymbol{\theta}) \tag{1}$$

$$-(1/2)\sum_{t=n_1+1}^{n}(\boldsymbol{x}_t - [\boldsymbol{e} \quad \boldsymbol{e}_2]\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - [\boldsymbol{e} \quad \boldsymbol{e}_2]\boldsymbol{\theta});$$

where $\boldsymbol{\theta} = \begin{bmatrix}\mu & \delta\end{bmatrix}'$; $\boldsymbol{\Sigma}$ is the variance-covariance matrix; C is a constant that does not depend on the parameters; n_1 is the number of years prior to treatment; n is the total number of years of the study (including the n_1 Before years); $\mathbf{x}_t$ is a k-vector of observed survivals in year t; k is the number of populations (treatment + control) used in the study; $\mathbf{e}$ is a k-vector of 1s; $\mathbf{e}_2$ is a k-vector of k_1 0s followed by k_2 1s, where k_1 represents the number of control populations and k_2 represents the number of treatment populations. The vector $\mathbf{x}_t$ is arranged so that the k_1 control populations precede the k_2 treatment populations.

Using maximum likelihood theory, estimating equations for the Before mean, treatment effect are developed. In this problem, maximizing the likelihood function is equivalent to a solving the generalized least squares (GLS) problem of minimizing

$$SS = \left(\begin{bmatrix}\bar{\boldsymbol{x}}_1\\ \bar{\boldsymbol{x}}_2\end{bmatrix} - \begin{bmatrix}\boldsymbol{e} & 0\\ \boldsymbol{e} & \boldsymbol{e}_2\end{bmatrix}\begin{bmatrix}\mu\\ \delta\end{bmatrix}\right)' \begin{bmatrix}\boldsymbol{\Sigma}^{-1}n_1 & 0\\ 0 & \boldsymbol{\Sigma}^{-1}n_2\end{bmatrix}\left(\begin{bmatrix}\bar{\boldsymbol{x}}_1\\ \bar{\boldsymbol{x}}_2\end{bmatrix} - \begin{bmatrix}\boldsymbol{e} & 0\\ \boldsymbol{e} & \boldsymbol{e}_2\end{bmatrix}\begin{bmatrix}\mu\\ \delta\end{bmatrix}\right); \tag{2}$$

where $\bar{\mathbf{x}}_1$ represents the k-vector of sample means of log(survival) in the Before period, and $\bar{\mathbf{x}}_2$ represents the k-vector of sample means of log(survival) in the After period.

This generalized sum of squares may be written in the familiar form

$$SS = (\boldsymbol{y} - \boldsymbol{B\theta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{B\theta}); \quad (3)$$

where $\boldsymbol{y}' = [\bar{\boldsymbol{x}}_1' \quad \bar{\boldsymbol{x}}_2']$,$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{e} & 0 \\ \boldsymbol{e} & \boldsymbol{e}_2 \end{bmatrix}$, and $\boldsymbol{\Omega}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}n_1 & 0 \\ 0 & \boldsymbol{\Sigma}^{-1}n_2 \end{bmatrix}$. In this form, the GLS solution, the called the Aitken estimator (Press 2005), is known to be

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{B}^{\mathbf{T}}\boldsymbol{\Omega}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathbf{T}}\boldsymbol{\Omega}^{-1}\mathbf{y}\,. \quad (4)$$

Using matrix algebra, I derived the following estimator:

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\mu} \\ \hat{\delta} \end{bmatrix} = \frac{\begin{bmatrix} (\boldsymbol{e}_2{}'\Sigma^{-1}\boldsymbol{e}_2)(\boldsymbol{e}'\Sigma^{-1}\bar{\boldsymbol{x}}) - (\frac{n_2}{n})(\boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_2)(\boldsymbol{e}_2{}'\Sigma^{-1}\bar{\boldsymbol{x}}_2) \\ (\boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e})(\boldsymbol{e}_2{}'\Sigma^{-1}\bar{\boldsymbol{x}}_2) - (\boldsymbol{e}_2{}'\Sigma^{-1}\boldsymbol{e})(\boldsymbol{e}'\Sigma^{-1}\bar{\boldsymbol{x}}) \end{bmatrix}}{(\boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e})(\boldsymbol{e}_2{}'\Sigma^{-1}\boldsymbol{e}_2) - (\frac{n_2}{n})(\boldsymbol{e}_2{}'\Sigma^{-1}\boldsymbol{e})^2}, \quad (5)$$

where $\bar{\mathbf{x}}$ is a *k*-vector representing population-specific sample means over the entire duration of the study. Using GLS, the variance of the estimate of $\boldsymbol{\theta}$ (given that the variance-covariance matrix is known) is equal to

$$\text{var}\left(\widehat{\boldsymbol{\theta}}\right) = (\boldsymbol{B^T \Omega^{-1} B})^{-1} = \frac{\begin{bmatrix} n_2 \boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e}_2 & -n_2 \boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e} \\ -n_2 \boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e} & n \boldsymbol{e}' \boldsymbol{\Sigma^{-1}} \boldsymbol{e} \end{bmatrix}}{n n_2 \left(\boldsymbol{e}' \boldsymbol{\Sigma^{-1}} \boldsymbol{e}\right)\left(\boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e}_2\right) - \left(n_2 \boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e}\right)^2}. \quad (6)$$

In particular, the variance of the treatment effect estimator is

$$\text{var}\left(\hat{\delta}\right) = \frac{n \boldsymbol{e}' \boldsymbol{\Sigma^{-1}} \boldsymbol{e}}{n n_2 \left(\boldsymbol{e}' \boldsymbol{\Sigma^{-1}} \boldsymbol{e}\right)\left(\boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e}_2\right) - \left(n_2 \boldsymbol{e}_2{}' \Sigma^{-1} \boldsymbol{e}\right)^2}. \quad (7)$$

This formula shows how the variance of the treatment effect estimate depends on the values of k_1, k_2, n_1, n_2 and the entries of the variance-covariance matrix $\boldsymbol{\Sigma}$. Using this formula, it is then simple to calculate the standard error of the treatment effect estimate, which is equal to the square root of the variance:

$$\text{se}(\hat{\delta}) = \sqrt{\text{var}(\hat{\delta})}. \quad (8)$$

The coefficient of variation is

$$CV(\hat{\delta}) = se(\hat{\delta})/\delta. \quad (9)$$

Power.—Power is the probability of rejecting the null hypothesis of "no treatment" effect when the actual treatment effect is the alternative value of δ. Power depends on the true treatment effect, the probability of a type I error (usually called the alpha value), and the standard error of the estimator. By maximum likelihood theory the estimator of the treatment effect is asymptotically normally distributed. In this case it may be shown that the estimator is normally distributed regardless of the sample size. This occurs because the variance-covariance matrix is assumed known and the estimator is a linear combination of random variables x_t that are known to follow a multivariate normal distribution. Any linear combination of normally distributed random variables is also normally distributed. Thus, the power may be written as

$$\Pi(\delta) = \Phi(-z_{\alpha/2} - \delta/se(\hat{\delta})) + 1 - \Phi(z_{\alpha/2} - \delta/se(\hat{\delta})); \qquad (10)$$

where $\Phi(z)$ is the cumulative distribution function of a random variable that follows a standard normal distribution (a normal distribution with mean zero and standard deviation 1), and $z_{\alpha/2}$ is the critical value such that $\alpha/2$ probability lies to the right of the value $z_{\alpha/2}$ in a standard normal distribution. For example, when $\alpha = 0.05$, the critical value is equal to 1.96. Investigators often choose a design such that power of 0.8 is achieved (e.g, Peterman and Bradford 1987, Liermann and Roni 2008).

Example.—To illustrate the use of the BACI tool, I examined power over a range of inputs. I set the common variance to the alternative values of 0.1, 0.5, and 1.0, and set correlation to $\rho = 0.50$. The total number of years of the study ranged from 10 years to 30 years with equal numbers of Before and After years. The alternative total numbers of populations were 2 and 10, with half of the populations being used as control populations and the other half as treatment populations. I assumed that the standard deviation of log measurement error was log(1.10)=0.095. I also assumed that measurement errors were independent so that measurement error covariances between populations were all zero. I set the treatment effect equal to log(1.30), representing a 30% change in survival.

The results of this example are illustrated in Figure 1. Power increased as the number of years of experiment increased, number of populations increased, and the variance decreased. In the two-population design, a power of 0.80 was achieved when variance was 0.1 and the number of years was 24 or greater. In the 10-population design, power of 0.80 was achieved when variance was 0.1 and the number of years was 10 or greater, or when variance was 0.5 and the number of years was 24 or greater. Over the range of years examined, a power of 0.80 was never achieved when the variance was 1.0.

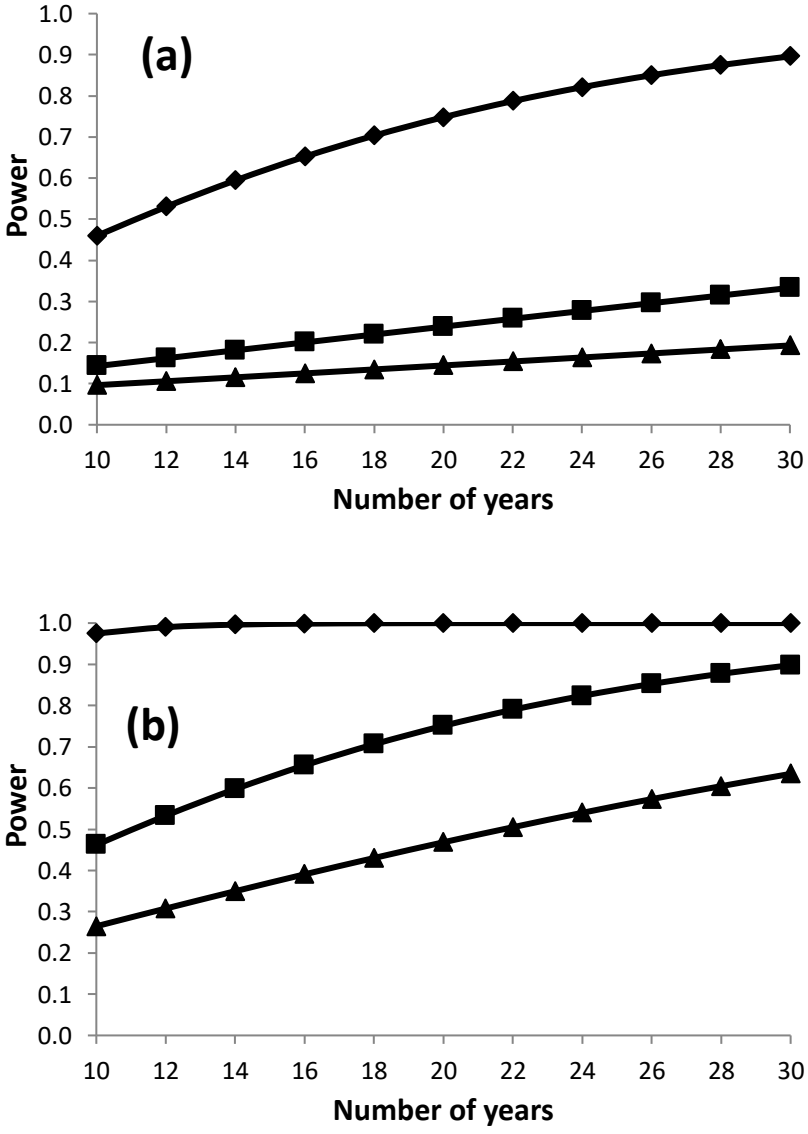

Figure 1.—Power varies with the number of years and number of populations used in the study. In (a), 2 populations were used and in (b), 10 populations were used. Curves were constructed using a year-to-year variance of 0.1 (diamonds), 0.5 (squares), and 1.0 (triangles). Equal numbers of treatment and control populations and equal numbers of Before and After years were assumed. Measurement error standard deviation was set at log(1.10) and the treatment effect was set to log(1.30).

ACKNOWLEDGEMENTS

This work was supported by the Bonneville Power Administration. Thanks to Charlie Paulsen, Rishi Sharma, Tracy Hillman, and Amber Parsons for their reviews.

REFERENCES

Fisher, R.A. 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, Scotland.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. Wiley and Sons, New York, New York.

Liermann, M., and P. Roni. 2008. More sites or more years? Optimal study design for monitoring fish response to watershed restoration. North American Journal of Fisheries Management. 28:935-943.

Mood, A.M, Graybill, F.A., and D.C. Boes. 1974. Introduction to the theory of statistics, Third Edition. McGraw-Hill, New York, New York.

Osenberg, C.W. and R.J. Schmidt. 1996. Detecting ecological impacts caused by human activities. In Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats, R.J Schmitt and C.W. Osenberg, Editors. Academic Press, New York, New York.

Peterman, W. M., and M. J. Bradford. 1987. Statistical power of trends in fish abundance. Canadian Journal of Fisheries and Aquatic Sciences 44: 1879:1889.

Press, S.J. 2005. Applied multivariate analysis: using Bayesian and frequentist methods of inference. Dover Publications, Inc. Mineola, New York

Venerables, W.N., Smith, D.M., and R Development Core Team. 2010. An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics Version 2.11.1 (2010-05-31). http://www.r-project.org

APPENDIX A. R-CODE

Table A.1. R-code used to calculate statistical power for the BACI-type design when variance is known.

```
#Program to estimate standard errors and power in a BACI-type design
#This R-code uses the added assumptions of a common variance and common correlation
#terms (the intraclass covariance structure studied by R. A. Fisher (1925). This assumption may
#be relaxed by specifying SIG2 as an input to the function baci in place of the input variables s2
#and rho
#inputs
#s2 is year-to-year variance (assumed equal for all populations)
#rho is the correlation of survivals between each pair of populations
#n1 number of Before years
#n2 number of After years
#k1 number of control populations
#k2 number of treatment populations
#me measurement error
#alpha -- prob. type I error (Probability of rejecting null hypothesis when true.)
#delta -- true treatment effect representing difference in natural log survival
#ln(Streatment/Scontrol)
#outputs
#se -- standard error
#cv -- coefficient of variation
#power -- probability of rejecting the null hypothesis of no effect when it is false
baci<-function(s2=1,rho=.9,n1=5,n2=5,k1=1,k2=1,me=log(1.10),alpha=0.05,delta=log(1.50)){
 check.inputs(s2,rho,n1,n2,k1,k2,me,alpha,delta)
 k<-k1+k2
 n<-n1+n2
 SIG2<-matrix(s2*rho,ncol=k,nrow=k)
 diag(SIG2)<-s2+me*me
 INVSIG2<-solve(SIG2)
 e<-rep(1,k)
 se<-n*t(e)%*%INVSIG2%*%e
```

```
 e2<-c(rep(0,k1),rep(1,k2))
 det<-n*t(e)%*%INVSIG2%*%e
 det<-det*n2*t(e2)%*%INVSIG2%*%e2-n2*n2*(t(e2)%*%INVSIG2%*%e)^2
 se<-sqrt(se/det)
#rule--reject when estimate exceeds q ses in absolute value (two-sided)
 q<-qnorm(1-alpha/2)
 power<-(1-pnorm(q*se,mean=delta,sd=se))+pnorm(-q*se,mean=delta,sd=se)
 return(list(s2=s2,rho=rho,n1=n1,n2=n2,k1=k1,k2=k2,me=me,
  alpha=alpha,delta=delta, se=se,cv=se/delta,power=power))
}

#make sure the inputs make sense
check.inputs<-function(s2,rho,n1,n2,k1,k2,me,alpha,delta){
 if(s2<0)stop("Variance, s2, must be nonnegative")
 if(me<0)stop("Measurement error standard deviation, me, must be nonnegative")
 if((s2+me*me)<=0)stop("Total variance, s2+me*me, must be positive")
 if(abs(rho)>1)stop("Correlation coefficient, rho, must be between -1 and 1")
 if(n1<=0)stop("Number of before years,n1, must be positive")
 if(round(n1)!=n1)stop("Number of before years,n1, must be an integer")
 if(n2<=0)stop("Number of after years, n2, must be positive")
 if(round(n2)!=n2)stop("Number of after years, n2, must be an integer")
 if(k1<=0)stop("Number of control populations, k1, must be positive")
 if(round(k1)!=k1)stop("Number of control populations, k1, must be an integer")
 if(k2<=0)stop("Number of treatment populations, k2, must be positive")
 if(round(k2)!=k2)stop("Number of treatment populations, k2, must be an integer")
 if((alpha<=0)|(alpha>=1))stop("Prob. of type I error, alpha, must be between 0 and 1")
 if(!is.double(delta))stop("The treatment effect, delta, must be a real number")
 return(NULL)
 }
```