Including Statistical Power for Determining

How Many Crashes Are Needed in Highway Safety Studies

Dominique Lord
Assistant Professor
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/458-3949, fax: 979/845-6481
E-mail: d-lord@tamu.edu

Byung-Jung Park
Graduate Assistant Researcher
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/845-2489, fax: 979/845-6481
E-mail: soldie@tamu.edu

Working Paper

May 13th, 2009

ABSTRACT

The objective of this short manuscript is to build on the work of Hauer (2008) and incorporate the concept associated with the power of a test (or Type II error) to calculate the necessary number of accidents in before-after studies. Statistical power plays an important role for minimizing false negatives in various types of studies. The consequences of ignoring the power in estimating the minimum sample size can have a significant impact in highway safety studies. In this case, a treatment that is deemed to be effective is rejected or not implemented because it was erroneously identified as ineffective. Using the same numerical example used by Hauer (2008), this manuscript shows that a much larger sample size is needed when the power is included in the study design.

Keywords: Highway safety; Statistical power; Type II error; Sample size; Study design

INTRODUCTION

In observational before-after studies, one of the most important questions that need to be answered is related to how many crashes are needed to estimate a change in safety with satisfactory precision. The short communication by Hauer (2008) has answered this question to some degree by means of the classical significance testing, which focuses primarily on controlling the probability of minimizing the type I error (i.e., α). The Type I error (aka false positive) is defined as the probability of rejecting the null hypothesis when it is actually true. The null hypothesis can usually be defined as "no effect in safety during before and after period." However, the major disadvantage of the classical significance test is that it dichotomizes the results into significant or not significant results with a p-value without taking into account the size of the observed effect (Singh, 2006). With this approach, whatever the magnitude of the observed effect may be, the null hypothesis can be rejected if the sample size is large enough. In highway safety studies, it may be more reasonable to focus on the precision for estimating the magnitude of the safety effect of a treatment rather than on the study's ability to reject the null hypothesis. The purpose of this article is to build on the work of Hauer (2008) and calculate the necessary number of crashes in before-after studies, by incorporating the statistical power of a study or test. The same example initially utilized by Hauer (2008) is used to illustrate this how to incorporate statistical power for determining the minimum sample size.

At the study design stage, if one is only interested in reducing the Type I error, it may be tempted to set the α value as low as possible (e.g., say 0.01). However, this results in the simultaneous increase in the likelihood of a Type II error (β), which in turn decreases the power ($1-\beta$).

Table 1 illustrates the four different situations used in hypothesis testing. The Type II error (aka false negative) is defined as the probability of failing to reject the null hypothesis when it is actually false. In other words, in highway safety studies, this would mean that although a treatment does reduce crashes, it is erroneously concluded that no change in safety has occurred. In short, the power can be interpreted as the probability that the false null hypothesis is correctly rejected.

Table 1. Matrix describing the hypothesis testing

	Do not reject H_0	Reject H_0
H_0 is True	Correct Decision 1-α: Confidence level	Type I error α: Significance level
H_0 is False	Type II error β	Correct Decision 1-β: Power of a test

POWER CALCULATION

The statistical power is defined as the probability of correctly detecting a true effect, if the effect actually exists. Figure 1 illustrates the basic principle underlying power calculations. The symbol d denotes the null value of the difference between the mean values for the groups being compared (typically, $d = 0$), and d_c denotes the value for the difference that is just significantly different from d at the significance level α. The symbol d^* is the magnitude of the actual difference in the mean values of the two groups, and it is the alternative hypothesis we wish to test. The portion of the distribution to the right of d_c represents the power. Based on the normal approximation of sampling distribution of the difference between the mean values for the groups and a few more assumptions, the power can be calculated according to the following equation (see Kelsey et al., 1986):

$$Z_\beta = \frac{d^* - d}{s.e.\left(d^*\right)} - Z_{\alpha/2} \tag{1}$$

Where, Z_β denotes the standard normal deviate corresponding to the position of d_c on the distribution around d^*, $Z_{\alpha/2}$ denotes the standard normal deviate corresponding to the position of d_c on the distribution around d, and $s.e.\left(d^*\right)$ denotes the standard error of d^*. Generally, a test with a power greater than 0.8 (or $\beta \le 0.20$) is considered statistically powerful.

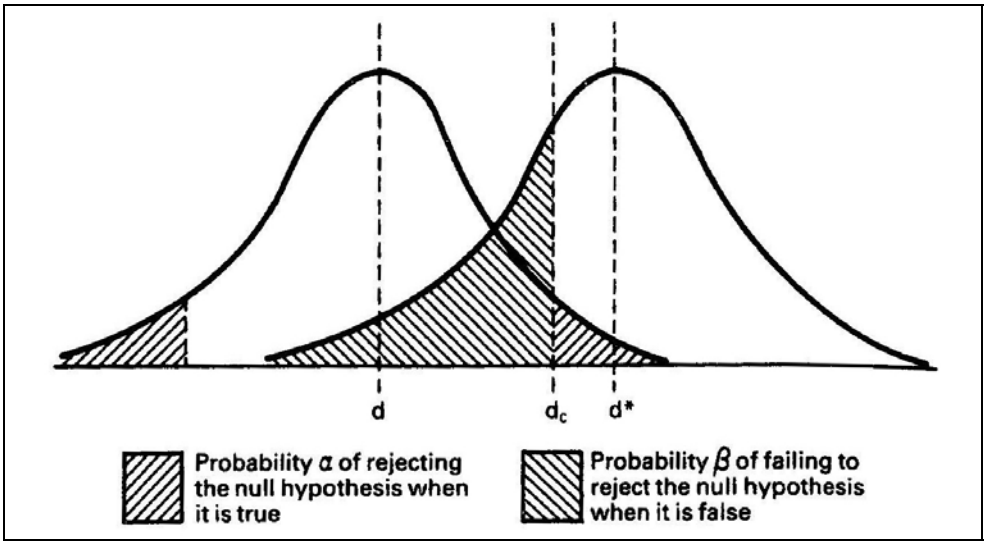

Figure 1. Concept of power (Kelsey et al., 1986)

COMPUTING THE SAMPLE SIZE

The required sample size can be determined to achieve the target statistical power. In this case, the significant level (α) and the size of the effect (the difference between the two mean values) should be pre-specified. In highway safety studies, the size of the effect can be interpreted as the change in safety between the before and after periods the researcher would like to detect; for example, a treatment is thought to reduce the expected number of crashes by 10% (often denoted as $\theta = 0.90$ in the safety literature). The question is, therefore, how many accidents are needed to be reasonably certain of distinguishing a true effect of this size from no effect at all.

Let x_1 and x_2 be the crash counts for c_1 and c_2 years or kilometers-years. Subscripts 1 and 2 represent the 'before' and 'after' periods, respectively. As mentioned in Hauer (2008) (and using the same notation), the difference $\mu_1 - \mu_2$ (d^* above) is estimated by $(x_1/c_1)-(x_2/c_2)$, and the variance of the estimated difference is given by $x_1/c_1^2 + x_2/c_2^2$. Since $d = 0$ in Equation (1), it can be rearranged in the following way:

$$\frac{d^*}{s.e(d^*)} = \frac{(x_1/c_1)-(x_2/c_2)}{\sqrt{(x_1/c_1^2)+(x_2/c_2^2)}} = Z_{\alpha/2} + Z_{\beta} \qquad (2)$$

Equation (2) provides an easy approach to compute the necessary sample size for a given significance level and power. For this exercise, we used *numerical example 2* from Hauer (2008). The example was slightly modified in order to incorporate the power analysis.

Numerical example 2: On a certain kind of road on which there are 1.5 reported crashes/kilometers-year an intervention is contemplated. How many kilometers of road are needed so that one can be 95% confident ($Z_{\alpha/2} = 1.96$) that in a before-after study a 10% reduction in expected accident frequency is detected, while ensuring the statistical power of 80% (i.e., $Z_{\beta} = 0.84$) (see Table 2 below)? Three years of 'before' and one year of 'after' data will be used, as described in the original example.

Let y_1 and y_2 be the 'before' and 'after' periods in years, and n be the number of kilometers of road. Then, $x_1 = 1.5 \times y_1 \times n = 4.5n$, and $x_2 = (1.5 \times 0.9) \times y_2 \times n = 1.35n$. The difference in mean values and the standard error are, $d^* = 1.5 - 1.5 \times 0.9 = 0.15$, and

$s.e.\left(d^*\right) = \sqrt{\frac{x_1}{\left(ny_1\right)^2} + \frac{x_2}{\left(ny_2\right)^2}} = \sqrt{\frac{4.5}{9n} + \frac{1.35}{n}}$, respectively. Therefore, solving Equation (2) for n with $Z_{\alpha/2} = 1.96$ and $Z_{\beta} = 0.84$ yields $n \approx 646$ km. This is nearly two times the length that was required by the approach without considering the power (Hauer (2008) estimated about 330 kilometers using $Z_{\alpha/2} = 2$). This implies that a 10% reduction in safety is so small to be correctly detected that it requires a large number of crashes. We realize that this significant increase in sample size may be difficult to collect in practice (Lord and Bonneson, 2005; Lord, 2006), but this is unfortunately necessary to properly estimate the effectiveness of treatments using a before-after study design.

Analogous to the *numerical example 3* in Hauer (2008), it is natural to ask about what would be a detectable change in the mean values with the same level of confidence (95%) and power (80%) when only 330 km and 100 km of highways are available. Again, using Equation (2), it can be shown that the reduction must be at least 13.8% and 24.0% for 330 km and 100 km, respectively, so that they are 'confidently and powerfully' detectable. The reduced sample size resulted in the increase of the size of the effect to produce the same level of confidence and power.

Table 2 below shows frequently used combinations of significance level and power. The values are computed for $Z_{\alpha/2=0.005} = 2.575$, $Z_{0.025} = 1.960$, and $Z_{0.05} = 1.645$, respectively. The value Z_{β} can be estimated by solving the equation shown on top of the last column. For $\alpha = 0.05$ and $1-\beta = 0.80$, we get $Z_{\beta} = 0.84$ $\left[\left(1.96 + Z_{\beta}\right)^2 = 7.849 \rightarrow Z_{\beta} = \sqrt{7.849} - 1.96 = 0.84\right]$.

Table 2. Frequent combinations of significance level and power (Kelsey et al., 1986)

Significance level (α)	Power ($1-\beta$)	$(Z_{\alpha/2}+Z_{\beta})^2$
0.01	0.80	11.679
	0.90	14.879
	0.95	17.814
	0.99	24.031
0.05	0.80	7.849
	0.90	10.507
	0.95	12.995
	0.99	18.372
0.10	0.80	6.183
	0.90	8.564
	0.95	10.822
	0.99	15.770

REFERENCES

Hauer, E., 2008. How many accidents are needed to show a difference? Accident Analysis Prevention 40(4), 1634-1635.

Kelsey, J. L., Thompson, W. G., and Evans, A. S., 1986. Methods in Observational Epidemiology, Oxford University Press, Inc., New York, USA.

Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma Models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention, 38(4), 751-766

Lord, D., and J.A. Bonneson, 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transportation Research Record 1908, 88-95.

Singh, G., 2006. A shift from significance test to hypothesis test through power analysis in medical research. Journal of Postgraduate Medicine 52(2), 148-150.