Published in final edited form as:

Stat Commun Infect Dis. 2011 October 1; 3(1): . doi:10.2202/1948-4690.1036.

Sample Size for a Binomial Proportion with Autocorrelation

Amalia S. Magaret, PhD* and

Department of Laboratory Medicine, University of Washington, Box 359928, Seattle, WA 98195

Jeffrey Stanaway, MPH

Department of Epidemiology, University of Washington

Abstract

A flexible sample size computation is desired for a binomial outcome consisting of repeated binary measures with autocorrelation over time. This type of outcome is common in viral shedding studies, in which each individual's outcome is a proportion: the number of samples on which virus is detected out of number of samples assessed. Autocorrelation between proximal samples occurs in some conditions such as herpes infection, in which reactivation is episodic. We determine a sample size computation that accounts for: (1) participant-level differences in outcome frequency, (2) autocorrelation in time between samples, and (3) varying number of samples per participant. In addition, we develop a computation appropriate for crossover designs that accounts for the dependence of the investigational treatment effect on the pretreatment detection frequency. The computations are validated through comparison with real and simulated data, and sensitivity to misspecification of parameter values is examined graphically.

1 Introduction

In studies of persons with genital or oral herpes simplex virus (HSV) infection, a common measure evaluated is the frequency of viral detection on the skin and mucosa, also known as viral shedding (Krone et al., 1998; Schacker et al., 1998; Wald et al., 2000; Corey et al., 2004; Gupta et al., 2004; Posavad et al., 2004). Genital and oral swabs are self-collected daily (or multiple times per day), and the presence of virus detected by polymerase chain reaction (PCR) is recorded for each sample. High frequencies of viral shedding have been associated with disease severity (Wald et al., 1995, 2003) and risk of transmission (Brown et al., 2003; Andrews et al., 2006; Corey et al., 2004). Since shedding is common and more objective and frequent than the presence of lesions, changes in shedding frequency can be used to demonstrate the efficacy of interventions aimed at reducing disease burden or transmissibility. Since shedding is intermittent, multiple samples increase precision: we have previously shown that 77% of HSV2-seropositive persons shed at least once over 30 consecutive daily samples (Magaret et al., 2009).

Sample size computations are readily available for binary outcomes when observed singly (Newcombe, 1998) or in a paired design (Nam and Blackwelder, 2002; Agresti and Min, 2005; Bonett and Price, 2006). Diggle et al. (2002) provide sample size for repeated measures binary outcomes with exchangeable correlation. In contrast, our studies rely on comparison of per-person frequencies, computed as the number of samples on which virus is detected out of number of samples assessed. These proportions vary for several reasons. First, disease severity differs between persons, and therefore some persons experience more frequent shedding than others. Second, herpes reactivations are episodic, and therefore

*corresponding author. Vaccine and Infectious Disease Institute, Fred Hutchinson Cancer Research Center, amag@u.washington.edu
ph: (206) 520-4316 fax: (206) 720-4371.

detection of virus tends to occur in “bursts” inducing greater association in detection between chronologically proximal versus distal samples. Third, the number and timing of per-person samples collected can vary depending on the trial’s secondary objectives, such as a desire to describe episode duration precisely. Therefore we required a sample size computation that included these elements of variability but could be adapted to any number of per-person samples.

In addition, we wanted to compute sample size for both cohort designs (in which each participant receives a single treatment or exposure) and crossover designs (in which participants experience both treatments sequentially). Crossover studies are commonly performed to evaluate potential therapies for herpesvirus infections as they increase efficiency relative to the cohort design. Crossover is feasible in contexts in which the investigational treatment (such as an antiviral) has a short-lived effect and participants can quickly begin the second arm of the trial.

Hereafter we have generalized “viral detection” as “event” and “shedding frequency” as “outcome or event frequency”, since these methods can be applied to any binomial outcome. These methods could be applied to any longitudinal assessment of an intermittent event such as a transient laboratory abnormality or recurring illness. Outcome frequencies resulting from repeated sampling have been reported in fields such as climatology (Wilks, 1997), pulmonary medicine (Ho et al., 2007; Hertz-Picciotto et al., 2007; Linares et al., 2010), and in the agricultural literature when detecting herd illness rates (Clough et al., 2003) or the presence of pest infestation (Schwertman and Smith, 1998).

We considered interventions that reduce the outcome frequency on a relative rather than absolute scale when compared to a standard treatment or placebo. This relative reduction is assumed to apply equally to all participants. For example, an investigational drug that is estimated to reduce the event frequency by 80% will cause a person who has a 60% event frequency on standard treatment to experience the event on only 12% of samples, and will reduce another person’s frequency from 20% to 4%. In contrast, hypothesizing an absolute reduction due to the investigational treatment might cause us to anticipate some negative outcome frequencies among those receiving that treatment. A proportionate effect is biologically plausible in that those who have low initial outcome frequencies will not substantially benefit from interventions.

We introduced notation (Section 2) and formulas (Section 3) for computing the variability of the observed outcome based on these assumptions: the outcome frequency varies between participants according to a prior Beta distribution, and the correlation between events over time is best described using a lag 1 autoregressive model (AR1) (Abraham and Ledolter, 1983) which also varies according to a Beta prior. In Section 4, the variability estimates from these formulas were compared to the observed variability in a large dataset of HSV-2 seropositive persons. Simulations were then performed (Section 5) comparing observed power to power based on variance formulas. Misspecification of parameter values used in the sample size computation was described graphically, (Section 6) to demonstrate the precision needed to achieve the desired power and to identify any regions of critical importance. Lastly, we briefly considered departures from assumptions (Section 7) and described the accompanying statistical code designed for use in R (Section 8).

2 Notation

Let X_{ijt} indicate the binary outcome (1=event, 0=no event) for participant i at time point j during treatment T . The number of samples for each participant on *each arm* (if participating in a crossover study) will be m ; the outcome frequency for participant i will be

$\frac{Y_{iT}}{m}$ where $Y_{iT} = \sum_{j=1}^m X_{ijT}$. We do not allow m to vary by individual since it is selected by design; it is understood, however, that failure to provide all samples will diminish power. Let $T \in \{A, B\}$ where treatment A is the standard treatment and B the investigational one.

We also introduce the following parameter notation. Let p_A be a random Beta-distributed variable (α, β) whose realizations define the true outcome frequencies p_{iA} for each individual i on standard treatment A . Let $E(p_A) = \mu$ and $\text{Var}(p_A) = \sigma^2$. Let K indicate the proportional reduction in the outcome frequency on the investigational treatment B . For cohort studies, p_B is also a random Beta-distributed variable, with mean $K\mu$ and variance $K^2\sigma^2$. For crossover studies, we assume p_B is *not* a random variable but determined by K , such that $p_{iB} = Kp_{iA}$, $\forall i$. That is, we assume that all participants of crossover studies experience the same reduction. This is done in part to avoid having to make more complex assumptions about the relationship between K and p_A , and in part because preliminary data on the variability of K is unlikely to be available.

Correlation between time points from the same participant will be summarized using a lag 1 autoregressive (AR1) correlation coefficient ϕ such that ϕ^k is the correlation between measurements k time points apart. We will allow the individual realizations ϕ_i to vary according to a prior Beta distribution with shape and scale parameters ε and ζ , such that $E(\phi) = \Phi$. We assume that these measures are estimable from previous data. The assumptions inherent in this AR1 model are biologically appropriate as HSV shedding follows an episodic pattern, and substantial variability in shedding characteristics and clinical severity occur between persons (Wald et al., 1995, 2000; Mark et al., 2008).

3 Methods

3.1 Cohort designs

We derived the variance estimate for a binomial proportion with autocorrelation using the variance decomposition formula (Kelly, 1994) and conditioning on the conjugate prior (Lehmann and Casella, 1998) such that $\text{Var}(Y_{iA}) = E[\text{Var}(Y_{iA}|p_{iA})] + \text{Var}[E(Y_{iA}|p_{iA})]$ where $p_{iA} \sim \text{Beta}(\alpha, \beta)$. The variance of a mean of a series of samples with autoregressive correlation was described by Leith (1973) and more explicitly by Wilks (1997) and includes computation of a variance inflation factor. The variance of the proportion for a single

individual is below, where $\gamma = \left(1 + \frac{2}{m} \sum_{j=1}^{m-1} \frac{\Gamma(\varepsilon+\zeta)\Gamma(\varepsilon+m-j)}{\Gamma(\varepsilon)\Gamma(\varepsilon+\zeta+m-j)}\right)$. For derivation, see appendix A (Section 10.1).

$$\text{Var}\left(\frac{Y_{iA}}{m}\right) = \frac{\alpha\beta((\alpha+\beta)\gamma+m)}{m(\alpha+\beta)^2(\alpha+\beta+1)} \quad (1)$$

Power or sample size computations for a standard two-sample test (Friedman et al., 1998) can be performed by assuming that the mean and variance of p_B in the investigational treatment group B are $E(p_B) = K\mu$ and $\text{Var}(p_B) = K^2\sigma^2$, respectively, where

$\mu = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Rearranging the expectation and variance of a Beta distributed variable, the hyperparameters γ and δ of p_B can be computed as follows.

$$\gamma = \frac{\mu^2}{\sigma^2}(1 - K\mu) - K\mu$$

$$\delta = \frac{\mu}{K\sigma^2}(1 - K\mu)^2 - (1 - K\mu) \quad (2)$$

Then assuming the same autocorrelation is present while on treatment B, the variability of $\frac{Y_{iB}}{m}$ can be computed by plugging in γ and δ to equation (1) in place of α and β . If prior data collected while on treatment B are available, estimates for γ and δ may be obtained hence.

3.2 Crossover studies

We assume that, conditioned on the value of the individual outcome frequency (p_{iA}), any time dependence between time points on different treatments is zero. The variance of the difference in outcome between treatments within a single individual is shown here, where

$$\gamma = \left(1 + \frac{2}{m} \sum_{j=1}^{m-1} \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + m - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + m - j)}\right). \text{ For derivation, see Appendix A (Section 10.2).}$$

$$\text{Var}\left(\frac{Y_{iA} - Y_{iB}}{m}\right) = \frac{\gamma}{m} \left[(K+1) \frac{\alpha}{\alpha + \beta} - (K^2 + 1) \left(\frac{\alpha\beta + \alpha^2(\alpha + \beta + 1)}{(\alpha + \beta)^2(\alpha + \beta + 1)} \right) \right] + (K-1)^2 \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}, \quad (3)$$

Standard power and sample size computations for a one-sample test (Friedman et al., 1998) can be performed using this variance estimate.

4 Application

We compared measures of accrued data from study participants seen at the University of Washington Virology Research Clinic (UW-VRC) against our formulas. We first described extraction of the parameter values used in variance computation from a one half 'estimation' subset of these data. This was done separately for those participants enrolled in observational studies versus in investigational drug studies. Subsequently, variability estimates from equations (1) and (3) based on parameters drawn from the estimation half of the dataset were compared against raw variance estimates from the remaining (testing) half of the data.

4.1 Data description

HSV Shedding frequency has been evaluated in a series of studies at the UW-VRC over the last three decades. We included shedding characteristics of study participants who performed daily home collection of genital swabs for a period of at least 30 days. For these participants, HSV detection was performed by PCR which is the most sensitive method (Ryncarz et al., 1999). Clinical findings from these studies have been previously reported (Schacker et al., 1998; Krone et al., 1998; Wald et al., 2000; Gupta et al., 2004; Posavad et al., 2004; Corey et al., 2004).

Two cohorts of study participants are described: (1) the natural history cohort who received no intervention, and (2) the crossover cohort who participated in interventional studies of suppressive antiviral therapy. Between 1989 and 2008, 531 HSV-2 seropositive, HIV-1 seronegative participants performed daily home collection of genital swabs for at least 30 days (median, 59 days; range, 30–564) and were included in the natural history cohort. Of 38,750 swabs collected, 7,231 (19%) were HSV DNA positive by PCR.

Over that same time interval, 156 participants completed placebo-controlled crossover trials of interventions to decrease HSV shedding through suppressive antiviral therapy (median

duration 59 days, range 30–564). Several trials were combined in which participants took either acyclovir 400mg twice daily or famciclovir 250mg twice daily. These drugs have previously been shown to have a similar reduction on the frequency of HSV shedding (between 75% and 80%; Gupta et al., 2004; Sacks, 2004). Shedding characteristics from the crossover cohort participants are summarized separately. Of 15,651 swabs collected, 8,142 were taken while off drug (of which 2054 or 25% were positive) and 7,509 were taken while on suppressive antiviral therapy (of which 368 or 5% were positive).

4.2 Parameter values used in variance estimation

We used a randomly selected one-half estimation subset of the observed data to estimate parameters needed for variance computation. To estimate the Beta distribution parameters, we derived the marginal distribution of Y_{iA} over a Beta prior for p_A as shown in equation (4). Maximum likelihood was used to obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ from individual outcomes Y_{iA} over m samples. The individual shedding frequencies used were not taken from consecutive samples but samples separated chronologically, in order to increase precision of the individual proportions and minimize the influence of autocorrelation. Therefore, we selected shedding indicators (from two to ten per participant) spaced a week apart in time. One week was chosen as the estimates of ϕ_j tend to fall to below 0.1 at this time (data not shown). For example, the number of samples with detectable HSV DNA out of 3 samples was computed as $Y_{jT} = X_{j(1)T} + X_{j(8)T} + X_{j(15)T}$. Resulting estimates $\hat{\alpha}$ and $\hat{\beta}$ were similar regardless of the number of samples used.

$$\begin{aligned} f(Y_{iA}) &= \int_0^1 f(Y_{iA}|p_{iA})f(p_{iA})dp_{iA} \\ &= \int_0^1 \binom{m}{Y_{iA}} p_{iA}^{Y_{iA}} (1-p_{iA})^{(m-Y_{iA})} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_{iA}^{(\alpha-1)} (1-p_{iA})^{(\beta-1)} dp_{iA} \quad (4) \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(Y_{iA}+\alpha)\Gamma(m-Y_{iA}+\beta)\Gamma(m+1)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+m)\Gamma(Y_{iA}+1)\Gamma(m-Y_{iA}+1)} \end{aligned}$$

The distribution of the autoregressive correlation ϕ was found by estimating ϕ_j separately from each participant's event data using time-series analysis. The hyperparameters ϵ and ζ for the Beta distribution of ϕ were estimated using maximum likelihood. It is important to note that estimation of ϕ_j for each individual requires consecutive samples and may be affected by substantial missingness.

Additional parameter estimation was needed to compute the variability of a difference in proportion in crossover studies (for use in equation (3)). To estimate these parameters, we evaluated the 156 participants who participated in placebo-controlled treatment trials of antivirals in a crossover design, as they were recruited to have more severe clinical disease and were anticipated to shed more frequently while on placebo.

Parameter estimates obtained from the natural history cohort include $\hat{\alpha} = 1.0$ and $\hat{\beta} = 6.2$ (mean shedding frequency $\hat{\mu} = 14\%$) as well as $\hat{\epsilon} = 2.1$ and $\hat{\zeta} = 1.8$ (mean AR1 correlation 0.54). The autocorrelation varied widely between participants, nearly spanning the range (0, 1), confirming the need to incorporate ϕ as a random variable and not a fixed parameter as originally proposed. From the crossover cohort we obtained, from data accrued while on placebo, $\hat{\alpha} = 0.7$, $\hat{\beta} = 2.1$ ($\hat{\mu} = 25\%$) as well as $\hat{\epsilon} = 0.9$ and $\hat{\zeta} = 1.0$ (mean AR1 correlation 0.47). We also estimated $\hat{K} = 0.27$, a 73% relative reduction in shedding on suppressive antiviral therapy, similar to that previously reported.

4.3 Comparison of expressions with real data

Parameter values extracted from the estimation half of the observed data were then re-applied to the formulas (1) and (3). We compared the formula results to the sample variances from the remaining testing half, using periods of observation from between 3 to 60 samples per participant. As shown in Figure 1, the formulas provide accurate variance estimates for the remaining data. Points generally fall near the line ($y = x$) showing agreement between formulaic and observed variance, except for the crossover variance estimates with 60 samples per participant, for which the formulaic variance is more than two times as high as the observed. While the other variance estimates were computed over at least 80 participants in each group, the questionable estimate was computed with only 12 participants in the estimation half and 11 participants in the testing half, as few participants contributed sufficient samples in each arm.

5 Power accuracy: simulation

We compared power derived from the variance formulas to that computed from simulated data, matching assumptions. True proportions p_A were simulated randomly according to a Beta distribution (α, β). Using these frequencies, the autoregressive correlation was induced by computing the probability of an event at each time conditioned on the previous time's simulated results. The individual AR1 correlation measures were also simulated using a Beta distribution ($\phi \sim \text{Beta}(\epsilon, \zeta)$). Parameters p_A and ϕ are assumed to be uncorrelated, as substantiated in our natural history cohort of over 500 participants showing a Pearson correlation coefficient of 0.1. The conditional probability for the event indicator at each time $j + 1$ is simulated based on the previous event at time j .

$$\begin{aligned} p(X_{i(j+1)T}=1 | p_{iT}, X_{iT}=1) &= \phi_i(1 - p_{iT}) + p_{iT} \\ p(X_{i(j+1)T}=1 | p_{iT}, X_{iT}=0) &= (1 - \phi_i)p_{iT} \end{aligned} \quad (5)$$

The expected value of $X_{iT}|p_{iT}$ is easily shown to be p_{iT} , for all j (see Appendix 11.1).

We first confirmed the accuracy of the variance formulas over a wide range of characteristics: varying mean outcome frequency ($\mu \in \{5\%, 10\%, 20\%, 40\%, 60\%, 80\%\}$), varying shape ($\alpha \in \{1\mu, 2\mu, 5\mu, 10\mu, 20\mu, 50\mu\}$), and treatment efficacy ($K \in \{0.2, 0.5, 0.8\}$). The Beta distributions assumed for the outcome are shown in Figures 2a–f and provide a comprehensive range of means and variances, from low to high and from narrow to broad, respectively. Variability in ϕ was assumed to be large (standard deviation 0.23, matching the variability in our data). We varied the average AR1 correlation so that ($\Phi \in \{0.2, 0.5, 0.8\}$) using (ϵ, ζ) $\in \{(0.4, 1.7), (2.0, 2.0), (1.7, 0.4)\}$ as shown in Figure 3. We did not consider as many conditions for ϕ as for p_A , however, we considered hyperparameters with the greatest variability so as to include the "worst case".

Simulations were performed over 1000 repetitions of each set of parameters, creating between 5 and 90 samples or event indicators for each of 1000 participants. The average bias between the sample variance and that anticipated by the formulas was less than 1% in all cases (results not shown). These results strongly suggest that the variance formulas accurately represent the data for most reasonable scenarios.

We then assessed through simulation the adequacy of the formulas to predict sample size needed for both cohort and crossover designs. The Beta distributions were simulated either with $\alpha = 1.0$, $\beta = 6.2$, $\epsilon = 2.1$ and $\zeta = 1.8$, (as from the natural history cohort) or with $\alpha = 0.7$, $\beta = 2.1$, $\epsilon = 0.9$ and $\zeta = 1.0$ (as from our crossover cohort). We considered treatment effects of $K \in \{0.2, 0.5, 0.8\}$, corresponding to 80%, 50% and 20% relative reductions in

outcome frequency, investigational versus standard. The number of samples per participant was 10, 30 or 60. The number of participants was computed to provide 80% power to detect a difference in event frequencies between treatments, using variance formulas (1) and (3). We computed power on simulated data sets using standard two-sample t-tests (for unequal variances) and paired t-tests (or one-sample tests for a difference). 6000 repetitions were performed for each set of parameters to provide a 95% confidence interval for power of $\pm 1\%$.

Table 1 shows that the anticipated power of 80% based on the formula (1) is achieved within the 1% margin in most cases. For large treatment effects ($K = 0.2$), the proposed computations may be a bit conservative. This is likely due to the fact that many participants will have no events observed on the investigational treatment, making variability less than anticipated.

Similarly, Table 2 shows that the observed power based on simulated data and the paired t-test is close to that estimated under equation (3), though the estimate can be a bit conservative when the treatment effect is large.

6 Misspecification of parameters

In Sections 4 and 5, we have evaluated the accuracy of the variance formulas when the parameter values are known or can be estimated from the data, however, these values may not be known precisely. In order to increase the applicability of our findings, we sought to determine the acceptable degree of misspecification of α , β , μ , ϵ , ζ and ϕ in determining

outcome variability, $\text{Var}(\frac{Y_{IA}}{m})$. We used the same broad range of underlying Beta distributions as described in Figures 2 and 3.

We considered four types of misspecification: (i) where α and β are both misspecified but relatively scaled such that the mean μ is still accurate, and ϵ and ζ remain accurate; (ii) where both α and β are accurate and where ϵ and ζ are both misspecified but relatively scaled such that the mean Φ is still accurate; (iii) where ϕ is misspecified by varying ζ , but α , β and ϵ are accurate; and (iv) where μ is misspecified by varying β but α , ϵ and ζ remain accurate. We considered a variability estimate within 25% of the true variability to be an acceptable degree of error and show the range of parameter estimates satisfying this criterion for each true parameter value. Figures 4a–d show the limits of acceptability for estimates on the y-axes versus true parameter values on the x-axes. Since sample size is directly proportional to the variability, these values ensure 25% error in sample size.

Figure 4a shows the range of acceptable estimates for α in case (i) above where the mean μ is accurately specified. The axes are given as multiples of μ , since the findings are the same for all values of the mean and depend only on the relationship between α and μ . For example, when $\alpha = 10\mu$ and Φ is 0.5, estimates for α between 6μ and 23μ will be adequate according to our requirement of less than 25% bias in variance. This translates to: for a specific outcome frequency of $\mu = 0.2$, α of $10 * 0.2 = 2.0$ being adequately represented by estimates between $\hat{\alpha} = 6 * 0.2 = 1.2$ and $\hat{\alpha} = 23 * 0.2 = 4.6$. As α increases relative to μ , the range of acceptable values for α widens. Misspecification of β when μ and Φ are accurate can be expressed using the same figure with the axis labels changed so that μ is replaced by $(1 - \mu)$.

Figure 4b provides acceptable ranges for $\hat{\epsilon}$ when Φ , α and β are correctly specified (case (ii) above). The misspecification ranges shown are for a high value of $\Phi = 0.80$. Lower values of Φ showed even wider limits, indicating that estimated values of ϵ and ζ may have little impact on variance when Φ is accurate. Misspecification of ζ when μ and Φ are accurate

can be expressed using the same figure with the axis labels changed so that Φ is replaced by $(1 - \Phi)$.

Figure 4c provides acceptable ranges for $\hat{\Phi}$ when other parameters are correctly specified (case (iii)). For example, when $\alpha = 50\mu$ and Φ is 0.80, estimates for Φ between 0.73 and 0.85 will be adequate according to our requirement of less than 25% bias in variance. The acceptable range for Φ widens as its value decreases and as the variability in outcome frequency increases (as α decreases relative to μ).

Figure 4d shows acceptable ranges for $\hat{\mu}$ when $\alpha = 5\mu$ (case (iv)). When $\mu = 0.05$, almost no deviation is tolerated in achieving an accurate sample size. At $\mu = 0.2$ and $\Phi=0.8$, $\hat{\mu}$ between 0.16 and 0.25 are acceptable. Therefore accuracy in the true outcome frequency is most critical for extreme values, high or low. Graphs of the misspecification of μ looked nearly identical at $\alpha = 50\mu$, indicating that the mean outcome frequency has the greatest influence on variability of outcome frequency.

7 Violations of Assumptions

The assumption of a constant proportionate effect on outcome frequency due to the intervention may not hold in the crossover design, indicating differences in treatment efficacy between persons. Simulations were performed (not shown) in which K was made to vary by individual, and the resulting data were compared to the variance estimate in equation (3) in which a fixed value for K (the mean) was applied. In some cases where the treatment effect was made to vary widely in the simulated data, simulated outcome variability was up to two times higher than that indicated by the formula. Therefore, a constant relative treatment effect must hold.

The assumptions of an underlying Beta distribution for outcome frequency or for AR1 correlation may not be accurate. We did not specifically simulate alternative distributions to assess this assumption. Nevertheless, we did simulate Beta distributions over a wide range of shapes and feel optimistic that most distributions with range (0, 1) would be well approximated by one of our selections.

8 Software

Programs have been written for use in R (version 2.11.1) to perform sample size or power computation for autocorrelated binomial outcomes. Two functions are available for download at this journal's website; see the supplemental content to this article. The first function (*estabphi*) uses preliminary data to compute Beta hyperparameters α and β for the outcome and ϵ and ζ for the autoregressive correlation. This function requires repeated measures binary data of the same frequency of collection as is planned for the study to be designed. The code is commented to describe the form of the required input data and interpretation of the output. The second function (*formbbinom*) computes sample size or power based on the formulas for either cohort or crossover design studies, and can use the input parameters estimated by the function *estabphi*.

9 Conclusions

Viral shedding data frequently involve multiple samples, and shedding rates commonly constitute the outcome of interest in studies of interventions or group comparisons. The variability of individual shedding rates is not captured by any methods available to date, as available methods either fail to incorporate individual variability and autocorrelation, or they are based on a previous study and so cannot be adapted to a different number of per-participant samples.

We have developed sample size computations for both cohort and crossover study designs, and provided software with which to implement these computations. Most input parameter values can be obtained from preliminary data of the same frequency of collection as planned for the study to be designed; an estimate of the treatment effect must also be supplied. Comparison to our existing cohorts demonstrate accuracy of the formulas in the context of HSV viral shedding. Simulations confirm that the variability of proportions can be adequately represented by implementing a Beta prior distribution for both outcome frequency and sample-to-sample autocorrelation. Power computations indicate that standard sample size estimates based on these variance formulas are highly accurate, though can be somewhat conservative for large effect sizes.

Tolerance for misspecification of the mean $\mu = \frac{\alpha}{\alpha + \beta}$ is especially low in the extremes (below $\mu = 20\%$ and above $\mu = 80\%$). When μ is correctly specified, the consequences are slight of misspecifying the hyperparameters α and β , except when the distribution is very wide (when α is small relative to μ). The average autocorrelation measure Φ generally has at least 5–10% flexibility in either direction without inducing a greater than 25% bias in the estimated sample size, and is very insensitive to accuracy of its hyperparameters ϵ and ζ .

While these computations were derived for estimation with viral shedding data, they could be applied to any repeated measures binary outcome for which autocorrelation in time is anticipated. As described in Section 1, we have found some examples in the literature, but many other applications may also be appropriate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Appendix A

10.1 Variability in proportions with autocorrelation

$$\begin{aligned}
 \text{Var}\left(\frac{Y_{iT}}{m}\right) &= \text{Var}\left(\frac{1}{m} \sum_{j=1}^m X_{ijt}\right) \\
 &= \text{E} \left[\text{Var} \left(\frac{1}{m} \sum_{j=1}^m X_{ijt} | p_{iT} \right) \right] + \text{Var} \left[\text{E} \left(\frac{1}{m} \sum_{j=1}^m X_{ijt} | p_{iT} \right) \right] \\
 &= \text{E} \left[\text{E} \left[\text{Var} \left(\frac{1}{m} \sum_{j=1}^m X_{ijt} | p_{iT}, \phi_i \right) | p_{iT} \right] + \text{Var} \left[\text{E} \left(\frac{1}{m} \sum_{j=1}^m X_{ijt} | p_{iT}, \phi_i \right) | p_{iT} \right] \right] + \text{Var} \\
 &\quad \left[\text{E} \left(\frac{1}{m} \sum_{j=1}^m X_{ijt} | p_{iT} \right) \right]
 \end{aligned}$$

$$\begin{aligned}
&= E \left[E \left[\frac{1}{m} p_{it}(1 - p_{it}) \left(1 + \frac{2}{m} \sum_{j=1}^{m-1} j \phi_i^{(m-j)} \right) | p_{it} \right] + \text{Var} (p_{it} | p_{it}) \right] + \text{Var} (p_{it}) \\
&= E \left[\frac{1}{m} p_{it}(1 - p_{it}) \left(1 + \frac{2}{m} \sum_{j=1}^{m-1} \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + m - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + m - j)} \right) + 0 \right] \\
&\quad + \text{Var} (p_{it}) = \frac{\gamma}{m} E [p_{it}(1 - p_{it})] + \text{Var} (p_{it}) \\
&= \frac{\gamma}{m} [E(p_{it}) - \text{Var}(p_{it}) - E^2(p_{it})] + \text{Var} (p_{it}) \\
&= \frac{\gamma}{m} \left[\frac{\alpha}{\alpha + \beta} - \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} - \frac{\alpha^2}{(\alpha + \beta)^2} \right] + \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\
&= \frac{\alpha\beta((\alpha + \beta)\gamma + m)}{m(\alpha + \beta)^2(\alpha + \beta + 1)}, \quad (6)
\end{aligned}$$

where $\left(1 + \frac{2}{m} \sum_{j=1}^{m-1} j \phi_i^{(m-j)}\right)$ is the variance inflation factor described by Wilks, and $\gamma = \left(1 + \frac{2}{m} \sum_{j=1}^{m-1} \frac{\Gamma(\varepsilon + \zeta) \Gamma(\varepsilon + m - j)}{\Gamma(\varepsilon) \Gamma(\varepsilon + \zeta + m - j)}\right)$ is its expectation when ϕ_i follows the Beta distribution with parameters (ε, ζ) .

10.2 Variability of difference, with autocorrelation

$$\begin{aligned}
\text{Var}\left(\frac{Y_{iA} - Y_{iB}}{m}\right) &= \text{Var}\left(\frac{1}{m} \sum_{j=1}^m X_{ijA} - \frac{1}{m} \sum_{j=1}^m X_{ijB}\right) \\
&= E \left[\text{Var} \left(\left(\frac{1}{m} \sum_{j=1}^m X_{ijA} - \frac{1}{m} \sum_{j=1}^m X_{ijB} \right) | p_{iA} \right) \right] + \text{Var} \\
&\quad \left[E \left(\left(\frac{1}{m} \sum_{j=1}^m X_{ijA} - \frac{1}{m} \sum_{j=1}^m X_{ijB} \right) | p_{iA} \right) \right]
\end{aligned}$$

$$=E \left[\text{Var} \left(\frac{1}{m} \sum_{j=1}^m X_{ijA} | p_{iA} \right) + \text{Var} \left(\frac{1}{m} \sum_{j=1}^m X_{ijB} | p_{iA} \right) - 2 \text{Cov} \left(\frac{1}{m} \sum_{j=1}^m X_{ijA}, \frac{1}{m} \sum_{j=1}^m X_{ijB} | p_{iA} \right) \right] + \text{Var} (p_{iA} - K p_{iA})$$

$$=E \left[\text{Var} \left(\frac{1}{m} \sum_{j=1}^m X_{ijA} | p_{iA} \right) + \text{Var} \left(\frac{1}{m} \sum_{j=1}^m X_{ijB} | p_{iA} \right) \right] + \text{Var} (p_{iA} - K p_{iA})$$

$$= \frac{\gamma}{m} E [p_{iA} (1 - p_{iA}) + K p_{iA} (1 - K p_{iA})] + \text{Var} (p_{iA} - K p_{iA})$$

$$= \frac{\gamma}{m} [(K+1)E(p_{iA}) - (K^2+1)E(p_{iA}^2)] + \text{Var} (p_{iA} - K p_{iA})$$

$$= \frac{\gamma}{m} \left[(K+1) \frac{\alpha}{\alpha+\beta} - (K^2+1) \left(\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2} \right) \right] + (K-1)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (7)$$

$$\text{where } \gamma = \left(1 + \frac{2}{m} \sum_{j=1}^{m-1} \frac{\Gamma(\varepsilon+\zeta)\Gamma(\varepsilon+m-j)}{\Gamma(\varepsilon)\Gamma(\varepsilon+\zeta+m-j)} \right).$$

Appendix B

11.1 Expectation of shedding indicators X_{ijT}

$$E(X_{i(j+1)T})$$

$$= p(X_{i(j+1)T}=1 | X_{iT}=1) * p(X_{iT}=1) \\ + p(X_{i(j+1)T}=1 | X_{iT}=0) * p(X_{iT}=0)$$

$$= (\phi(1 - p_{iT}) + p_{iT}) * p_{iT} + ((1 - \phi)p_{iT}) * (1 - p_{iT})$$

$$= \phi[(1 - p_{iT}) * p_{iT} - p_{iT} * (1 - p_{iT})] + p_{iT} [p_{iT} + 1 - p_{iT}]$$

$$= \phi * 0 + p_{iT} * 1$$

$$=P_{IT} \quad (8)$$

References

- Abraham, B.; Ledolter, J. Statistical Methods for Forecasting. New York, USA: Wiley; 1983.
- Agresti A, Min Y. Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*. 2005; 24:729–740. [PubMed: 15696504]
- Andrews W, Kimberlin D, Whitley R, Cliver S, Ramsey P, Deeter R. Valacyclovir therapy to reduce recurrent genital herpes in pregnant women. *American Journal of Obstetrics and Gynecology*. 2006; 194:774–781. [PubMed: 16522412]
- Bonett D, Price R. Confidence intervals for a ratio of binomial proportions based on paired data. *Statistics in Medicine*. 2006; 25:3039–3047. [PubMed: 16345058]
- Brown Z, Wald A, Morrow R, Selke S, Zeh J, Corey L. Effect of serologic status and cesarean delivery on transmission rates of herpes simplex virus from mother to infant. *Journal of the American Medical Association*. 2003; 289:203–209.
- Clough H, Clancy D, O'Neill P, French N. Bayesian methods for estimating pathogen prevalence within groups of animals from faecal-pat sampling. *Preventive Veterinary Medicine*. 2003; 58:145–169. [PubMed: 12706055]
- Corey L, Wald A, Patel R, Sacks S, Tyring S, Warren T, JM D, Paavonen J, Morrow R, Beutner K, Stratchounsky L, Mertz G, Keene O, Watson H, Tait D, Vargas-Cortes M. Once-daily valacyclovir to reduce the risk of transmission of genital herpes. *New England Journal of Medicine*. 2004; 350:11–20. [PubMed: 14702423]
- Diggle, P.; Heagerty, P.; Liang, K.; Zeger, S. *Analysis of Longitudinal Data*. 2nd edition. Oxford, England: Oxford; 2002.
- Friedman, L.; Furberg, C.; DeMets, D. *Fundamentals of clinical trials*. 3rd edition. New York, USA: Springer; 1998.
- Gupta R, Wald A, Krantz E, Selke S, Warren T, Vargas-Cortes M, Miller G, Corey L. Valacyclovir and acyclovir for suppression of shedding of herpes simplex virus in the genital tract. *Journal of Infectious Diseases*. 2004; 190:1374–1381. [PubMed: 15378428]
- Hertz-Picciotto I, Baker R, Yap P, Dostal M, Joad J, Lipsett M, Greenfield T, Herr C, Benes I, Shumway R, Pinkerton K, Sram R. Early childhood lower respiratory illness and air pollution. *Environmental Health Perspectives*. 2007; 115:1510–1518. [PubMed: 17938744]
- Ho W, HARTLEY W, Myers L, Lin M, Lin Y, Lien C, Lin R. Air pollution, weather, and associated risk factors related to asthma prevalence and attack rate. *Environmental Research*. 2007; 104:402–409. [PubMed: 17316602]
- Kelly, D. *Introduction to Probability*. New York, USA: Macmillan; 1994.
- Krone M, Tabet S, Paradise M, Wald A, Corey L, Celum C. Herpes simplex virus shedding among human immunodeficiency virus-negative men who have sex with men: site and frequency of shedding. *Journal of Infectious Diseases*. 1998; 178:978–982. [PubMed: 9806024]
- Lehmann, E.; Casella, G. *Theory of Point Estimation*. 2nd edition. New York, USA: Springer; 1998.
- Leith C. The standard error of time-average estimates of climatic means. *Journal of Applied Meteorology*. 1973; 12:1066–1069.
- Linares B, Guizar J, Armador N, Garcia A, Miranda V, Perez J, Chapela R. Impact of air pollution on pulmonary function and respiratory symptoms in children. longitudinal repeated-measures study. *BMC Pulmonary Medicine*. 2010; 10:1–9. [PubMed: 20051135]
- Magaret A, Johnston C, Wald A. Use of the designation “shedder” in mucosal detection of herpes simplex virus dna involving repeated sampling. *Sexually Transmitted Infections*. 2009; 85:270–275. [PubMed: 19211593]
- Mark K, Wald A, Magaret A, Selke S, Olin L, Huang M, Corey L. Rapidly cleared episodes of herpes simplex virus reactivation in immunocompetent adults. *Journal of Infectious Diseases*. 2008; 198:1141–1149. [PubMed: 18783315]

- Nam J, Blackwelder W. Analysis of the ratio of marginal probabilities in a matched-pair setting. *Statistics in Medicine*. 2002; 21:689–699. [PubMed: 11870810]
- Newcombe R. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*. 1998; 17:857–872. [PubMed: 9595616]
- Posavad C, Wald A, Kuntz S, Huang M, Selke S, Krantz E, Corey L. Frequent reactivation of herpes simplex virus among HIV-1-infected patients treated with highly active antiretroviral therapy. *Journal of Infectious Diseases*. 2004; 190:693–696. [PubMed: 15272395]
- Ryncarz A, Goddard J, Wald A, Huang M, Roizman B, Corey L. Development of a high-throughput quantitative assay for detecting herpes simplex virus dna in clinical samples. *Journal of Clinical Microbiology*. 1999; 37:1941–1947. [PubMed: 10325351]
- Sacks S. Famciclovir suppression of asymptomatic and symptomatic recurrent anogenital herpes simplex virus shedding in women: a randomized, double-blind, double-dummy, placebo-controlled, parallel-group, single-center trial. *Journal of Infectious Diseases*. 2004; 189:1341–1347. [PubMed: 15073669]
- Schacker T, Zeh J, Hu H, Hill E, Corey L. Frequency of symptomatic and asymptomatic herpes simplex virus type 2 reactivations among human immunodeficiency virus-infected men. *Journal of Infectious Diseases*. 1998; 178:1616–1622. [PubMed: 9815213]
- Schwertman N, Smith D. Optimal sample size for detection of an infestation. *Journal of Agricultural, Biological, and Environmental Statistics*. 1998; 3:359–369.
- Wald A, Huang M, Carrell D, Selke S, Corey L. Polymerase chain reaction for detection of herpes simplex virus (hsv) dna on mucosal surfaces: comparison with hsv isolation in cell culture. *Journal of Infectious Diseases*. 2003; 188:1345–1351. [PubMed: 14593592]
- Wald A, Zeh J, Selke S, Ashley R, Corey L. Virologic characteristics of subclinical and symptomatic genital herpes infections. *New England Journal of Medicine*. 1995; 333:770–775. [PubMed: 7643884]
- Wald A, Zeh J, Selke S, Warren T, Ryncarz A, Ashley R, Krieger J, Corey L. Reactivation of genital herpes simplex virus type 2 infection in asymptomatic seropositive persons. *New England Journal of Medicine*. 2000; 342:844–850. [PubMed: 10727588]
- Wilks D. Resampling hypothesis tests for autocorrelated fields. *American Meteorological Society*. 1997; 10:65–82.

Figure 1.

Observed versus proposed variance estimates by number of samples: 'shedding' refers to the shedding frequency of a single group as used in the cohort design (equation (1)), and 'crossover' refers to the change in shedding (equation (3)). Sample variances come from the testing half of the UW-VRC data while the estimates from the formula are based on parameters obtained from the estimation subset.

Figure 2.
Range of Beta distributions considered as possible prior distributions for outcome frequencies.

Figure 3.
Range of Beta distributions considered as possible prior distributions for AR1 correlation.

Figure 4.

Estimated parameter values versus true values, showing amount of allowable misspecification. Lines show the limits of an acceptable region between which variance estimates will fall within 25% of the true value.

Table 1

Power for the cohort study. μ = outcome frequency on treatment A. α , β , ϵ and ζ are hyperparameters. K = reduction on treatment B. m = samples per participant. n = participants per arm. Power(F) = formulaic power, Power(T) = simulated power using the t-test.

μ	α	β	ϵ	ζ	K	m	n	Power(F)	Power(T)
14%	1.0	6.2	2.1	1.8	0.2	10	38	80%	84%
14%	1.0	6.2	2.1	1.8	0.2	30	23	80%	87%
14%	1.0	6.2	2.1	1.8	0.2	60	17	80%	86%
14%	1.0	6.2	2.1	1.8	0.5	10	121	80%	81%
14%	1.0	6.2	2.1	1.8	0.5	30	71	80%	82%
14%	1.0	6.2	2.1	1.8	0.5	60	53	80%	81%
14%	1.0	6.2	2.1	1.8	0.8	10	911	80%	80%
14%	1.0	6.2	2.1	1.8	0.8	30	544	80%	80%
14%	1.0	6.2	2.1	1.8	0.8	60	412	80%	81%

Power for the crossover study. μ = outcome frequency on treatment A. α , β , ϵ and ζ are hyperparameters. K = reduction on treatment B. m = samples per participant. n = total participants. Power(F) = formulaic power, Power(T) = simulated power using the t-test.

Table 2

μ	α	β	ϵ	ζ	K	m	n	Power(F)	Power(T)
25%	0.7	2.1	0.9	1.0	0.2	10	19	80%	85%
25%	0.7	2.1	0.9	1.0	0.2	30	13	80%	85%
25%	0.7	2.1	0.9	1.0	0.2	60	10	80%	82%
25%	0.7	2.1	0.9	1.0	0.5	10	45	80%	80%
25%	0.7	2.1	0.9	1.0	0.5	30	26	80%	81%
25%	0.7	2.1	0.9	1.0	0.5	60	19	80%	84%
25%	0.7	2.1	0.9	1.0	0.8	10	281	80%	80%
25%	0.7	2.1	0.9	1.0	0.8	30	144	80%	79%
25%	0.7	2.1	0.9	1.0	0.8	60	91	80%	80%