Sample Size Calculations for Comparing Groups with Binary Outcomes

Xunan ZHANG¹, Jiangnan LYU², Justin TU³, Jinyuan LIU⁴, Xiang LU⁵*

Summary: Sample size is a critical parameter for clinical studies. However, to many biomedical and psychosocial investigators, power and sample size analysis seems like a magic trick of statisticians. In this paper, we continue to discuss power and sample size calculations by focusing on binary outcomes. We again emphasize the importance of close interactions between investigators and biostatisticians in setting up hypotheses and carrying out power analyses.

Key words: sample size, binary outcomes

[Shanghai Arch Psychiatry. 2017; 29(5): 316-324. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.217132]

1. Introduction
Sample size plays a critical role in clinical research studies. It provides information for optimal use of available resources to detect treatment differences. In the last article, we discussed sample size calculations for comparing means of continuous outcomes between two groups. In this report, we continue our discussion of this topic and turn our attention to extending our earlier considerations to binary outcomes.

Sample size is determined through power analysis. Unlike data analysis, power analysis is carried out at the design stage of a clinical study before any data is collected. Because of lack of data during power analysis, study investigators need to provide information about treatment differences, which not only allow biostatisticians to proceed with power analysis, but enable power analysis results to become meaningful and reliable.¹ Thus, power analysis is not a “trick” played by the statistician, but rather, an integrative process involving close interactions between study investigators and biostatisticians.

Note that editors of some medical journals sometimes ask authors of a manuscript to provide power analysis results of their study to support their findings. Such post-hoc power analysis generally makes no logical sense.² As most research studies are conducted based on a random sample from a study population of interest, results from power analysis become meaningless, as the random component in the study disappears once data are collected. Before the study begins, the study sample is unknown and outcomes of interest are random. Power analysis shows the probability, or likelihood, that a test statistic (function of data) will hypothesized difference between the two populations, such as the t statistic for comparing mean blood pressure levels between a hypertension and a normal population.³ Once the study is complete, we observe a sample, i.e., a particular group of subjects among many such groups from the study population, and data from this group of subjects become non-random.

In this article, we focus on comparing proportions of binary outcomes between two groups. As in our
previous article on power analysis for comparing two group means for continuous outcomes, we consider both independent and paired groups. We begin our discussion with a brief overview of the concept of power analysis within the context of one group. Although most studies involve comparing two or more treatment groups, the simplified setting of one group helps better illustrate the basic steps for sample size calculations.

2. Sample Size for One Group

Consider a binary outcome $X$ with values 0 and 1. In most clinical studies, the value 1 of $X$ generally denotes the occurrence of a disease or exposure of interest, such as depression or trauma. A binary outcome is generally modeled by the Bernoulli distribution with the probability $p$ of the occurrence of 1 in the outcome $X$, denoted by $X \sim \text{Bern}(p)$. Note that unlike the normal distribution for a continuous outcome, there is only one parameter for the Bernoulli distribution. This is because unlike the normal distribution, the variance of the Bernoulli is determined by $p$. Like the normal distribution, $p$ is also the mean of $X$, which is the proportion of 1’s. For example, if $X$ indicates the presence ($X = 1$) or absence ($X = 0$) of major depression in an individual from a population of interest, then $p$ is the prevalence of major depression in the population.

Consider testing the hypothesis,

$$H_0 : p = b \quad \text{vs.} \quad H_a : p \neq b,$$

(1)

where $b$ is a known constant, and $H_0$ and $H_a$ are known as the null and alternative hypotheses, respectively. Note that the above is known as a two-sided hypothesis, as no direction of effect is specified in the alternative hypothesis $H_a : p \neq b$. As two-sided alternatives are the most popular in clinical research, we only focus on such hypotheses in what follows unless stated otherwise.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $X \sim \text{Bern}(p)$ and $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean, which, within the current context of a binary outcome, is the percent, or proportion, of 1’s in the sample. If $\bar{X}$ indicates the presence or absence of major depression, $\bar{X}$ is the percent of major depression in the sample.

If the $n$ subjects are randomly sampled, $\bar{X}$ is an estimate of prevalence $p$ of major depression in the study population (for a non-random sample, $\bar{X}$ is still an estimate of prevalence, but for the general population because of potential selection bias). If the null $H_0$ is true, $\bar{X}$ has a high probability of being close to $b$. However, because $\bar{X}$ is random, it is still possible for $\bar{X}$ to be distant from $b$, although such probabilities are small, especially for large $n$. The type I error $\alpha$, a quantity introduced to indicate such an error rate, is the probability that measures the likelihood when $\bar{X}$ is far away from $b$ under $H_0$. This error rate is typically set at $\alpha = 0.05$ for most studies and at $\alpha = 0.01$ for studies with large sample sizes. Given $\alpha$, power is the probability that $H_0$ is rejected when it is false.

The decision to reject the null is based on some statistics that capture the difference between $\bar{X}$ and $b$ and the distribution of such a distance measure. The most popular and well-known measure is the $z$-score:

$$z = \frac{\sqrt{n}(\bar{X} - b)}{\sqrt{b(1-b)}},$$

(2)

which approximately follows the standard normal distribution for large sample size $n$. Thus, we reject $H_0$ as a type I error $\alpha$, if $|z| > z_{\alpha/2}$, where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution, i.e., $\Phi(z_{\alpha/2}) = 1 - \alpha/2$, with $\Phi$ denoting the cumulative distribution function of the standard normal distribution. For example, for $\alpha = 0.05$,$\, z_{\alpha/2} = 1.645$. If $H_0 : p = b$ is true, the probability of rejecting $H_0$, therefore committing a type I error $\alpha$, is given by:

$$P \left( \frac{\sqrt{n}(\bar{X} - b)}{\sqrt{b(1-b)}} \geq z_{\alpha/2} \mid H_0 \right) = \alpha.$$ 

(3)

Note that if we compare the $z$-score in (2) with the $z$-score for the continuous outcome in the previous article, we see that they only differ in the denominator:

$$z = \frac{\sqrt{n}(\bar{X} - b)}{\sqrt{b(1-b)}} \quad \text{for binary outcome}$$

vs.

$$z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \quad \text{for continuous outcome}.$$ 

This is because the standard deviation of the Bernoulli distribution is $\sqrt{b(1-b)}$, which is a function of $b$, rather than a different parameter as for continuous outcomes.

In clinical studies, we are really interested in the opposite, $H_0$ is set up as a straw man to help formulate and test our hypotheses against $H_0$. Statistical power allows us to quantify the likelihood of rejecting $H_0$ in favor of the alternative $H_a$, to support a different proportion, $d \neq b$, i.e.,

$$H_0 : p = b \quad \text{vs.} \quad H_a : p = d \neq b.$$ 

(4)

Without loss of generality, we assume $d > b$. For power analysis, we must also specify a known value $d$ for $P$ a priori, in addition to the value $b$ under $H_0$, in order to quantify our ability to reject $H_0$ in favor of $H_a$. Such explicit specification is not required for data analysis after data is observed.
Given a type I error $\alpha$ and a specific $d$ in $H_a$, we then calculate power, or the probability that (the absolute value of) the standardized difference in (2) exceeds the threshold $\frac{d}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, i.e.,

$$\text{Power}(n, \alpha, H_0, H_a) = P\left( \frac{\sqrt{n}(\bar{X} - b)}{\sqrt{b(1-b)}} \geq z_{\alpha/2} \mid H_a \right). \tag{5}$$

By comparing the above with (3), we see that the only difference in (5) is the change of condition from $H_a$ to $H^\alpha$. The probability in (5) is again readily evaluated to yield:

$$\text{Power}(n, \alpha, H_0, H_a) = 1 - \Phi\left( z_{\alpha/2} - \frac{\sqrt{n}(\bar{X} - b)}{\sqrt{b(1-b)}} \right) + \Phi\left( z_{\alpha/2} + \frac{\sqrt{n}(\bar{X} - b)}{\sqrt{b(1-b)}} \right). \tag{6}$$

Thus, as in the case of continuous outcomes, power $\text{Power}(n, \alpha, H_0, H_a)$ is a function of sample size $n$, type I error $\alpha$ and values of the parameter of the Bernoulli, $p$, specified in the null $H_0$ and alternative $H_a$.

Once $\alpha$ is selected, power is only a function of sample size $n$, and $b$ and $d$ specified in the null and alternative hypothesis. To determine sample size $n$, we must specify $b$ and $d$ reflect treatment effects, which are study specific and require investigators’ knowledge. As power is quite sensitive to these parameters, careful consideration and justification of these quantities is critical for calculated sample size to be meaningful, reliable and informative. Thus, power analysis is not merely an algebraic and computational exercise by biostatisticians, but is an integrative process involving critical input from content researchers.

Power increases as $n$ grows and approaches 1 as $n$ grows unbounded. Thus, by increasing sample size, we can have more power to reject the null, or ascertain treatment effect. However, we must be mindful about selecting an appropriate power level, as arbitrarily increasing sample size not only leads to waste of precious manpower and resources, but also increases the likelihood of failed studies due to logistic constraints, and diminishing interest and return due to rapid scientific progresses and discoveries and changing technologies. Power is generally set at some reasonable level such as 0.8 or 0.9. Also, small treatment effect may have little clinical relevance. Thus, it is critical that we specify treatment effects that correspond to clinically meaningful differences, which again require critical input from investigators specializing in the field of study.

Given a type I error $\alpha$, a pre-specified power, often denoted as $1 - \beta$, and $H_0$ and $H_a$, sample size is the smallest $n$ such that the test has the given power to reject $H_0$ under $H_a$.

$$\text{Power}(n, \alpha, H_0, H_a) \geq 1 - \beta. \tag{7}$$

Although it is generally difficult to find an analytical formula to compute the smallest $n$ satisfying (7), such an $n$ is readily obtained by using statistical packages. Note that power in the literature is typically denoted by $1 - \beta$, where $\beta$, known as “type II error rate”, denotes the probability that the null $H_0$ is accepted when in fact it is false.

For continuous outcomes, difference $\mu_i - \mu_0$ between $\mu_i$ under $H_a$ and $\mu_0$ under $H_0$ is generally expressed as an “effect size” to remove its dependence on the scale of $X$:

$$\delta = \frac{|\mu_i - \mu_0|}{\sigma}. \tag{8}$$

For binary outcomes, such standardization is not needed, as the variance of $X$ is completely determined by the parameter $p$ of the Bernoulli distribution. Thus, for binary outcomes, the difference $p_i - p_0$ between $p_i$ under $H_a$ and $p_0$ under $H_0$ is often interpreted as the effect size.

Note that for large sample size $n$, the $z$-score in (2) has approximately the standard normal distribution, which provides the basis for evaluating power using the expression in (6) when testing the hypothesis in (4). For moderate sample size, the normal approximation can still be used if $np \geq 5$ and $n(1-p) \geq 5$, where $P$ is either $p_0$ or $p_1$. If these conditions are not met, the $z$-score may deviate significantly from the normal distribution and the expression in (6) no longer provides reliable power estimates. Different methods must be used. For example, in exact inference, we use the binomial distribution of count of 1’s to derive the power function.[9] Exact methods work for both small and large sample size. However, for large sample size, it takes a long time to evaluate the power function, even with modern computing power. Thus, exact methods are usually used only in cases where $P$ or $n$ or both are small.

3. Sample Size for Two Independent Groups

Now consider two independent samples and let $X_i$ ( $i = 0,1; j = 1, K, n_i$) denote the random outcomes from the two samples. We assume that both group outcomes follow Bernoulli distribution, $X_i \sim \text{Bern}(p_i)$, with parameters $p_i$ ($i = 0,1$).

Considering testing the hypothesis,

$$H_0 : p_0 - p_1 = 0 \text{ vs. } H_a : p_0 - p_1 = \delta \neq 0 \tag{9}$$

Let $\bar{X}_i = n_i^{-1} \sum_{i=1}^{n_i} X_i$ denote the sample mean of the $i$th group ($i = 0,1$). As in the one-sample case, the difference between the two sample means $\bar{X}_1 - \bar{X}_0$ should be close to 0 if $H_0$ is true. Again, because $\bar{X}_1$ and $\bar{X}_0$ are random, it is still possible for $\bar{X}_1 - \bar{X}_0$ to be very different from 0, although such probabilities are small, especially for large $n$. As noted earlier, the level of such type I error $\alpha$ is set equal to 0.05 or 0.01 depending on sample size as discussed earlier.
In most clinical trials, groups have equal sample size, i.e., \( n_0 = n_1 \). Some studies may have a larger sample size for one of the groups. In our discussion below, we assume unequal sample sizes so that \( n_0 \neq n_1 \).

If \( H_0 : p_1 - p_0 = 0 \) is true, the probability of rejecting \( H_0 \) if therefore committing a type I error \( \alpha \), is:

\[
P\left(\frac{\overline{X}_1 - \overline{X}_0}{\sqrt{\frac{n_1(1-p_1)}{n_0} + \frac{n_1(1-p_1)}{n_1}}} \geq z_{\alpha/2} \mid H_0\right) = \alpha,
\]

where \( z_{\alpha/2} \) is the upper \( \alpha/2 \) quantile of the standard normal distribution. For power analysis, our goal is to reject the null \( H_0 \) in favor of the alternative \( H_a \). Without loss of generality, we assume \( \delta > 0 \). Given a significance level \( \alpha \), \( H_0 \) and \( H_a \), we then calculate power, or the probability that the absolutely value of the difference in (10) exceeds the threshold \( z_{\alpha/2} \), i.e.,

\[
\text{Power}(n, \alpha, H_0, H_a) = P\left(\frac{\overline{X}_1 - \overline{X}_0}{\sqrt{\frac{n_1(1-p_1)}{n_0} + \frac{n_1(1-p_1)}{n_1}}} \geq z_{\alpha/2} \mid H_a\right).
\]

Note that as the power function depends on the parameters of Bernoulli distributions for both groups, \( p_i \) and \( p_o \), not just the difference \( \delta = p_1 - p_0 \), we must specify both parameters, in addition to the difference \( \delta \) to compute power. This step is similar to specifying the standard deviations, \( \sigma_1 \) and \( \sigma_0 \), in addition to the mean difference, \( \mu_1 - \mu_0 \), when calculating power for continuous outcomes as discussed in the previous article. As the standard deviation of \( \overline{X}_1 \) (or \( \overline{X}_0 \)) is determined by \( p_i \) (or \( p_o \)), the parameter \( p_i \) (or \( p_o \)) plays the role of specifying the standard deviation for \( \overline{X}_1 \) (or \( \overline{X}_0 \)). Thus, we recast hypotheses (9) equivalently as:

\[
H_0 : p_0 = p_1 = b \quad \text{vs.} \quad H_a : p_0 = d_0, \quad p_1 = d_1, \quad d_0 \neq d_1.
\]

Given a type I error \( \alpha \), a power \( 1 - \beta \), and \( H_0 \) and \( H_a \), we can also readily determine numerically the smallest \( n \) such that the test has the given power to reject the null \( H_0 \) under \( H_a \), i.e.,

\[
\text{Power}(n, \alpha, H_0, H_a) \geq 1 - \beta.
\]

3. Sample Size for Paired Groups

In the last section, the two groups are assumed independent. This assumption is satisfied when the groups are formed by different subjects, such as male vs. female and depressed vs. healthy control subjects. In many studies, we may also be interested in changes before and after an intervention on the same individual. For example, suppose we are interested in the effect of a new antidepressant medication. We may give the drug to a group of depressed patients and measure their depression severity before and after taking the medication. Unlike groups formed by different subjects, the control (before taking the medication) and intervention (after the medication) groups are formed by the same individuals and outcomes generally become dependent between the two groups. For example, patients higher on depression severity before the medication likely remain so after the medication. As a result, the power function for testing two independent groups discussed earlier no longer applies to such dependent “paired” groups.

Consider a study with \( n \) pairs of observations and let \( (X_{ij}, y_{ij}) \) denote the two paired outcomes in the \( j \)th pair \((1 \leq j \leq n)\). For each pair, treatment difference is \( D_j = X_{ij} - X_{ij} \). If the difference \( D_j \) has a mean \( d = 0 \), then there is no treatment effect. Thus, we are interested in testing the hypothesis:

\[
H_0 : d = 0 \quad \text{vs.} \quad H_a : d \neq 0.
\]

For continuous outcomes \( X_{ij} \) and \( X_{ij} \), the difference \( D_j = X_{ij} - X_{ij} \) is also continuous. Thus, (13) becomes a hypothesis for testing whether the mean of \( D_j \) is 0 and sample size calculations can be carried out using the power function for the one group case as discussed in the previous article for power analysis for continuous outcomes. This approach, however, does not work within the current context of binary outcomes, since the difference \( D_j = X_{ij} - X_{ij} \) may take on the value \(-1\) in addition to 0 and 1 and thus no longer follows the Bernoulli distribution.

The most common approach for paired groups with binary outcomes is the McNemar’s test. By displaying \( n \) paired outcomes in a \( 2 \times 2 \) contingency table, we have:

|       | \( X_0 \) | \( X_1 \) | Marginal Total |
|-------|-----------|-----------|---------------|
| \( X_0 \) | \( n_{11} \) | \( n_{12} \) | \( n_{1+} \) |
| \( X_1 \) | \( n_{21} \) | \( n_{22} \) | \( n_{2+} \) |
| Marginal Total | \( n_{+1} \) | \( n_{+2} \) | \( n \) |

where the rows (columns) denote the two values of \( X_0 \) (or \( X_1 \)) and \( n_{ij} \) denotes the cell counts for the cell defined by \( X_0 = i \) and \( X_1 = j \) \((i, j = 0, 1)\). For data analysis, we can estimate the mean \( p_0 \) of \( X_{ij} \) and mean \( p_1 \) of \( X_{ij} \), and mean \( \delta = p_1 - p_0 \) of \( D_j = X_{ij} - X_{ij} \) by the cell counts:

\[
\hat{p}_1 = \frac{n_{+2}}{n}, \quad \hat{p}_0 = \frac{n_{+2}}{n}, \quad \hat{\delta} = \hat{p}_1 - \hat{p}_0,
\]

where \( \hat{p}_1 \), \( \hat{p}_0 \) and \( \hat{\delta} \) denote sample versions of the (population) parameters \( p_0 \), \( p_1 \) and \( \delta \), respectively.
For large \( n \), McNemar’s test statistic, \( \sqrt{\frac{n_{12} - n_{21}}{n_{12} + n_{21}}} \), has approximately a standard normal under the null \( H_0 \) in (13), i.e.,
\[
\frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \sim N(0,1),
\]
It is interesting to note that \( n_{12} - n_{21} \) in the numerator of McNemar’s statistic is the number of discordant pairs and large values of this statistic indicate evidence against the null \( H_0 \) in favor of the alternative \( H_a \). Thus, only the number of discordant pairs contributes information to testing the null hypothesis, which makes perfect sense, since concordant pairs indicate no change before and after the intervention. Note that McNemar’s test statistic is also often expressed in the form of an approximate chi-square distribution:
\[
\frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2_1,
\]
where \( \chi^2_1 \) denotes the chi-square distribution with one degree of freedom.

If \( n \) or any of the cell counts \( n_{ij} \) is small, the normal (or chi-square) distribution may have a poor approximation to the sampling distribution of McNemar’s statistic and other methods may be used to compute p-values.\(^5\) For example, in exact inference, we use an alternative form of McNemar’s statistic,\( \frac{n_{21}}{n_{21} + n_{12}} \), and determine the sampling distribution of this statistic to compute p-values for testing the null in (13). Both the normal (or chi-square) and exact method can be utilized to derive power functions for performing power analysis, with the exact method providing more reliable estimates for relatively small sample sizes.

4. Illustrations
In this section, we illustrate power and sample size calculations for comparing two independent and two paired groups. We continue to use G*Power in our examples, as it is free and easy to use. In all cases, we set power at 80% and two-sided alpha at \( \alpha = 0.05 \).

Example 1. A San Diego-based biopharmaceutical company plans to conduct a study to test the efficacy of an experimental Ebola drug. To determine the sample size, the investigators use their pilot data and obtain the following information concerning death rates between the company’s new drug and standard care:
- Death rate for new drug: 0.22
- Death rate for standard care: 0.38.

The problem is to estimate sample size for the study to detect the above difference in death rates between the two treatment conditions.

Let \( p_1 \) (\( p_0 \)) denote the percent of death for the new drug (standard care). We can express the corresponding statistical hypothesis as follows:
\[
H_0 : p_0 = p_1 = 0.38 \quad \text{vs.} \quad H_a : p_1 = 0.22 < p_0 = 0.38.
\]

Since subjects will be randomized to either the new drug or standard care, the study sample forms two independent groups. For convenience, we assume that the two treatment groups have the same number of subjects, i.e., \( n_0 = n_1 \).

To calculate sample size using the G*Power package, we enter the following information:

Test family > z tests
Statistical test > Proportions: Difference between two independent proportions
Type of power analysis > A priori: Compute required sample size – given \( \alpha \), power and effect size
Tails > Two
Proportion p2 > 0.22
Proportion p1 > 0.38
\( \alpha \) err prob > 0.05
Power (1 - \( \beta \) err prob) > 0.80
Allocation ratio N2/N1 > 1

By clicking on “Calculate”, we obtain a sample size of 128 for each group, or a total of 256 for both groups (see Figure 1).

The G*Power also offers an exact method to calculate sample size. In this case, we enter the following information:

Test family > Exact
Statistical test > Proportions: Inequality, two independent groups (Fisher’s exact test)
Type of power analysis > A priori: Compute required sample size – given \( \alpha \), power and effect size
Tails > Two
Proportion p2 > 0.22
Proportion p1 > 0.38
\( \alpha \) err prob > 0.05
Power (1 - \( \beta \) err prob) > 0.80
Allocation ratio N2/N1 > 1

By clicking on “Calculate”, we obtain a sample size of 139 for each group, or a total of 278 for both groups (see Figure 2). The estimated sample size using the exact method is slightly higher than the asymptotic method based on the standard normal distribution. Here the sample size is moderate and the discrepancy between the asymptotic and exact methods likely reflects the limited sample size. In general, if exact methods are used, we should go with sample size estimated from such methods. Fortunately, differences between
asymptotic and exact methods diminish as sample size increases. Thus, such difference generally does not have any major impact on real studies.

Example 2. A research team is interested in conducting research on sexual behaviors among the Botswana Defense Force. The team has learned from other similar studies that self-reported sexual behaviors based on a daily diary is more accurate than a retrospective survey. They have estimated that about 50% would report having sex with spouse within last two weeks by daily diary, while only 20% would report such events by retrospective recall. Before conducting the survey, the research team wants to confirm such discrepancy to justify their use of a daily diary for their study.

Let \( p_1 \) (\( p_0 \)) denote the percent of sex reported in a daily diary (retrospective recall). Then, the team’s interest can be stated in a hypothesis as:

\[ H_0 : p_0 = 0.3 \quad \text{vs.} \quad H_a : p_1 = 0.5 > p_0 = 0.3. \]

Since both daily diary and retrospective recall are completed by the same subject, the outcomes from the diary and retrospective recall are not independent. Thus, we use McNemar’s test for comparing sexual behaviors reported by the two assessment strategies and estimate sample size using the method for paired groups.

To use the G*Power, we need to enter the odds ratio and proportion of discordant pairs under \( H_a \). To compute these quantities, it is helpful to create the following 2x2 table indicating both the marginal probabilities \( p_1 \) (\( p_0 \)) (specified in the hypothesis) and joint probabilities (calculated from the marginal probabilities).
Table 2. Marginal and joint cell probabilities for the marginal and joint distribution of paired binary outcomes.

| $X_0$ | Marginal probability |
|-------|----------------------|
| 0     | $p_0$                |
| 1     | $p_0(1 - p_1)$       |
|       | $p_1$                |

Both the odds ratio and proportion of discordant pairs are readily computed from the above table:

Odds ratio: $$\theta = \frac{p_1 / (1 - p_1)}{p_0 / (1 - p_0)} = \frac{0.5 / (1 - 0.5)}{0.3 / (1 - 0.3)} = 2.333$$

Proportion of discordant pairs: $$\phi = p_1p_0(1 - p_1) + p_0p_1(1 - p_1) = 0.18$$

We then enter these quantities, along with some other information, into the G*Power:

Test family > Exact
Statistical test > Proportions: Inequality, two dependent groups (McNemar)
Type of power analysis > A priori: Compute required sample size – given $\alpha$, power and effect size
Tails > Two
Odds ratio > 2.333
$\alpha$ err prob > 0.05
Power (1 - $\beta$ err prob) > 0.80
Prop discordant pairs > 0.18

By clicking on “Calculate”, we obtain a sample of 273 subjects to detect the hypothesized difference in reporting sexual activities between daily diary and retrospective recall (see Figure 3).
4. Conclusion
Sample size estimation is an essential component of planning clinical research studies. It provides critical information for assessing feasibility of a planned study. For power analysis to be informative and useful, it requires reliable information on effect size, which can only be provided by biomedical and psychosocial investigators specializing in the field of the study. Thus, although power and sample size analysis relies on solid statistical theory, efficient computational methods and modern computing power, sample size estimates obtained from state-of-the-art methods and cutting-edge computing power are really useless without input from scientific investigators.

Conflicts of interest statement
The authors have no conflict of interest to declare.

Authors’ contributions
Michael Zheng, Justin Tu and Xiang Lu: Manuscript outline, structure and drafting.
Michael Zheng, Jinyuan Liu and Jinyuan Liu: Technical details of statistical tests, power functions.
Jiangnan Lyn and Jinyuan Liu: Computations of sample size for illustrative examples using statistical software.
Michael Zheng, Jiangnan Lyn, Justin Tu, Jinyuan Liu, Xiang Lu: Manuscript finalizing and proofreading.

Funding statement
This study received no external funding.
二分类结果组间比较的样本量计算

Zhang X, Lyn J, Tu J, Liu J, Lu X

概述：样本大小是临床研究的一个重要参数。然而，把握度和样本量分析对许多生物医学和社会心理调查者来说似乎是一个统计学家的魔术。在本文中，我们继续讨论二分类结果的把握度和样本量的计算。我们再次强调了在建立假设和进行把握度分析中调查者和生物统计学家之间密切联系的重要性。

关键词：样本量、二分类结果

References

1. Chow SC, Chang M. Adaptive design methods in clinical trials. New York: Chapman & Hall / CRC; 2007
2. Heonig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001; 55(1): 19--24. doi: https://doi.org/10.1198/000313001300339897
3. Kreyszig E. Advanced Engineering Mathematics (Fourth ed.). New York: Wiley; 1979
4. Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, et al. Prophylactic Implantation of a Defibrillator in Patients with Myocardial Infarction and Reduced Ejection Fraction. N Engl J Med. 2002; 346: 877--883. doi: https://doi.org/10.1056/NEJMoa013474
5. Tang W, He H, Tu XM. Applied Categorical and Count Data Analysis. FL: Chapman & Hall/CRC; 2012

Xunan Zhang is currently completing his BS in the department of Computer Science and Statistics, University of California - Berkeley, USA. His research interest is in the field of biostatistics.