DOI: 10.1534/genetics.110.118778

Statistical Analysis of Nondisjunction Assays in Drosophila

Yong Zeng,* Hua Li,† Nicole M. Schweppe,†,‡ R. Scott Hawley†,§ and William D. Gilliland†,,1**

**Department of Mathematics and Statistics, University of Missouri, Kansas City, Missouri 64110, †Stowers Institute for Medical Research, Kansas City, Missouri 64110, ‡Kansas City University of Medicine and Biosciences, Kansas City, Missouri 64119, §Department of Physiology, University of Kansas Medical Center, Kansas City, Kansas 66103-2918 and **Department of Biological Sciences, DePaul University, Chicago, Illinois 60614-3207*

Manuscript received May 11, 2010
Accepted for publication July 22, 2010

ABSTRACT

Many advances in the understanding of meiosis have been made by measuring how often errors in chromosome segregation occur. This process of nondisjunction can be studied by counting experimental progeny, but direct measurement of nondisjunction rates is complicated by not all classes of nondisjunctional progeny being viable. For *X* chromosome nondisjunction in Drosophila female meiosis, all of the normal progeny survive, while nondisjunctional eggs produce viable progeny only if fertilized by sperm that carry the appropriate sex chromosome. The rate of nondisjunction has traditionally been estimated by assuming a binomial process and doubling the number of observed nondisjunctional progeny, to account for the inviable classes. However, the correct way to derive statistics (such as confidence intervals or hypothesis testing) by this approach is far from clear. Instead, we use the multinomial-Poisson hierarchy model and demonstrate that the old estimator is in fact the maximum-likelihood estimator (MLE). Under more general assumptions, we derive asymptotic normality of this estimator and construct confidence interval and hypothesis testing formulae. Confidence intervals under this framework are always larger than under the binomial framework, and application to published data shows that use of the multinomial approach can avoid an apparent type 1 error made by use of the binomial assumption. The current study provides guidance for researchers designing genetic experiments on nondisjunction and improves several methods for the analysis of genetic data.

MEIOSIS is a specialized cell division, where a diploid cell undergoes a single round of replication followed by two rounds of segregation to produce four haploid gametes. During this segregation, chromosomes must correctly separate (or disjoin) from their homologs at meiosis I, followed by sister chromatids disjoining at meiosis II. When chromosomes fail to disjoin from their partners, the resultant nondisjunction produces aneuploid gametes with the wrong number of chromosomes. The study of meiotic nondisjunction in Drosophila has a long and distinguished history of publication in GENETICS, with the inaugural article published in this journal being Calvin Bridges' use of nondisjunction to prove the chromosome theory of heredity (BRIDGES 1916). The first study that screened variants isolated from natural populations used nondisjunction to identify meiotic mutants (SANDLER *et al.* 1968), as did the first EMS-induced mutant screen (BAKER and CARPENTER 1972). Subsequent screens using new mutagens or techniques have also relied on measuring nondisjunction to identify mutants of interest (SEKELSKY *et al.* 1999). Indeed, much of the progress that has been made in the study of meiosis would not have been possible without the use of nondisjunction to identify new mutations that are defective at some step in chromosome segregation.

However, one difficulty in estimating nondisjunction rates is that in most instances the resulting aneuploid progeny cannot survive. Fortunately, in Drosophila it is possible to design crosses to recover them. Sex determination in flies is based on the number of *X* chromosomes, rather than a masculinizing *Y* chromosome as in mammals. This means that *XO* flies are viable (but sterile) males, while *XXY* flies are viable females. Therefore, it is possible to recover both normal and nondisjunctional progeny, as a nullo-*X* egg fertilized by an *X*-bearing sperm will survive as an *XO* male, while a diplo-*X* egg fertilized by a sperm lacking an *X* will be female (*XXY*). By using visible markers on the sex chromosomes, these exceptional progeny are straightforward to identify. However, if those eggs are fertilized by the other class of sperm, the resulting *OY* or *XXX* progeny are inviable. Therefore, the nondisjunction rate that occurs during meiosis is not equal to the proportion of nondisjunctional progeny, as only 50% of nondisjunctional eggs receive sperm compatible with viability, while all normal eggs are viable.

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.118778/DC1.

[1] *Corresponding author:* Department of Biological Sciences, DePaul University, 2325 N. Clifton St., Chicago, IL 60614-3207.
E-mail: wgillila@depaul.edu

Given this experimental limitation, what is the correct method to calculate the error rate during meiosis? For this discussion, let N be the total number of progeny produced in an experiment, let X_1 be the number of inviable nondisjunctional progeny (*OY* and *XXX*), let X_2 be the number of viable nondisjunctional progeny (*XO* and *XXY)*, and let X_3 be the number of normal progeny (*XY* and *XX*), such that $N = X_1 + X_2 + X_3$. If all progeny could be counted, then the nondisjunction rate $\hat{p}$ would simply be $(X_1 + X_2)/N$.

However, only flies that survive to adulthood can be counted, and therefore both X_1 and N are unknown. As *X*- and *Y*-bearing sperm are produced in equal numbers, live and dead nondisjunctional progeny are also expected in equal numbers. Therefore, K.W. Cooper (Cooper 1948) proposed the widely used estimator for the *X* chromosome nondisjunction rate, where X_2 is substituted for X_1 in the above formula, giving the rate as:

$$\hat{p} = \frac{2X_2}{2X_2 + X_3}. \qquad (1)$$

While this estimator works, the statistical properties of this estimator are not clear. Instead of following the early literature to combine X_1 and X_2 and use a binomial distribution, we go back to the three original categories and model the process as a multinomial distribution with latent number of progeny N, considering all three possible phenotypes for each progeny (nondisjunctional dead, nondisjunctional living, and normal). Whether a nondisjunctional oocyte becomes a nondisjunctional dead or nondisjunctional living progeny depends on the sex chromosome content of the sperm that fertilized it. As *X*- and *Y*-bearing sperm are produced in equal numbers during male meiosis, the usual genetic expectation for the rates of nondisjunctional dead and living progeny will be $X_1 = X_2 = p/2$. However, even assuming that the rates of nondisjunctional dead and living progeny are different, with a Poisson assumption of N, we can derive the maximum-likelihood estimators (MLEs) for the nondisjunctional dead and nondisjunctional living rates. Under the usual genetic expectation of equality, the MLE of the nondisjunctional rate coincides with Cooper's estimator, and we furthermore derive the exact distribution of $\hat{p}$. Under another set of reasonable assumptions, we show the consistency and asymptotic normality of Cooper's estimator, and derive asymptotic results when comparing two nondisjunction rates. All these distributional results enable us to develop confidence interval and hypothesis testing related to p, or $p_x - p_y$ in the case of comparing two nondisjunction rates from populations *x* and *y*.

FORMULATION OF THE PROBLEM

Suppose an experiment produces a total of N oocytes. There are three possible cases for each oocyte: nondisjunctional dead, nondisjunctional living, and normal. These classes have the corresponding probabilities p_1, p_2, and $1 - p_1 - p_2$, where p_1 (p_2) is the nondisjunctional dead (living) rate. For the *i*th progeny, let X_{i1} be the indicator of the *i* nondisjunctional dead defined as $X_{i1} = 1$ if *i*th progeny is nondisjunctional dead, and $X_{i1} = 0$, otherwise. Similarly, we define X_{i2} and X_{i3} as the indicators of the *i*th nondisjunctional living and regular progeny. Then, $X_{i1} + X_{i2} + X_{i3} = 1$. For $j = 1, 2, 3$, $X_j = \sum_{i=1}^{N} X_{ij}$, $N = \sum_{j=1}^{3} X_j$, and X_1, X_2, and X_3 are the number of progeny in each of three categories.

Given $N = n$, the conditional distribution of (X_1, X_2, X_3) is a multinomial distribution with $(p_1, p_2, 1 - p_1 - p_2)$. The probability mass function (p.m.f.) is

$$f(X_1 = x_1, X_2 = x_2, X_3 = x_3 \mid N = n) = \frac{n!}{x_1!x_2!x_3!}(p_1)^{x_1}(p_2)^{x_2}(1 - p_1 - p_2)^{x_3}\mathbf{I}_{\{x_1+x_2+x_3=n\}}. \qquad (2)$$

THE EXACT DISTRIBUTION OF $\hat{P}$ UNDER POISSON ASSUMPTION

First, we make a Poisson assumption for N, which naturally comes from the most classical hierarchical model, known as binomial-Poisson hierarchy (see Casella and Berger 2001, Examples 4.4.1 and 4.4.2). We then derive $\hat{p}_1$ and $\hat{p}_2$, the maximum-likelihood estimators for p_1 and p_2. Under the usual genetic expectation that *X*- and *Y*-bearing sperm are produced in equal numbers (and therefore $p_1 = p_2$), and ignoring all other causes of mortality, we show that $\hat{p}_1 + \hat{p}_2$ is equal to Cooper's estimator of $\hat{p}$, and we further derive its exact distribution.

The likelihood function: To specify the likelihood function of the observed (X_2, X_3), we assume that the number of progeny, N, has a Poisson probability distribution: $P(N = n) = e^{-\lambda}\lambda^n/n!$. Then, the joint p.m.f. can be written as

$$f(x_1, x_2, x_3) = \left(e^{-\lambda p_1}\frac{(\lambda p_1)^{x_1}}{x_1!}\right)\left(e^{-\lambda p_2}\frac{(\lambda p_2)^{x_2}}{x_2!}\right)\left(e^{-\lambda(1-p_1-p_2)}\frac{(\lambda(1 - p_1 - p_2))^{x_3}}{x_3!}\right). \qquad (3)$$

This implies that under the Poisson progeny assumption, X_1, X_2, and X_3 are independent Poisson random variables with parameters λp_1, λp_2, and $\lambda(1 - p_1 - p_2)$, respectively. This desirable property with the observation that $\sum_{x_1=1}^{\infty} e^{-\lambda p_1}((\lambda p_1)^{x_1}/x_1!) = 1$ helps to obtain a simple likelihood of (p_1, p_2) by summing over x_1 as follows:

$$L(p_1, p_2) = L(p_1, p_2; x_2, x_3) = \left(e^{-\lambda p_2}\frac{(\lambda p_2)^{x_2}}{x_2!}\right)\left(e^{-\lambda(1-p_1-p_2)}\frac{(\lambda(1 - p_1 - p_2))^{x_3}}{x_3!}\right). \qquad (4)$$

Let $l(p_1, p_2) = \log L(p_1, p_2)$ be the log likelihood.

The maximum-likelihood estimators: Setting $(\partial/\partial p_1)l(p_1, p_2) = 0$ and $(\partial/\phi_2)l(p_1, p_2) = 0$, we obtain

$$\frac{\partial l}{\partial p_1} = \lambda - \frac{x_3}{1 - p_1 - p_2} = 0 \text{ and}$$
$$\frac{\partial l}{\partial p_2} = \frac{x_2}{p_2} - \frac{x_3}{1 - p_1 - p_2} = 0$$

with roots:

$$p_1 = \frac{\lambda - x_2 - x_3}{\lambda} \text{ and } p_2 = \frac{x_2}{\lambda}.$$

It can be checked that the second-order Jacobian matrix is nonpositive definite, ensuring that $(\hat{p}_1, \hat{p}_2)$ is the maximizer.

To realize the estimators of p_1 and p_2, we need to estimate λ. However, without further constraint on p_1 and p_2, λ can be any positive number larger than $x_2 + x_3$ because the given observations of x_2 and x_3 allow us to only estimate the ratio of p_2 and p_3. Further restricting $p_1 = kp_2$ for a positive k, a reasonable estimate for λ is $\lambda = (1/k+1)\,x_2 + x_3$. and then MLEs for p_1 and p_2 are

$$p_{1\mathrm{ML}} = \frac{x_2}{(k+1)x_2 + kx_3} \text{ and } p_{2\mathrm{ML}} = kp_{1\mathrm{ML}}$$
$$= \frac{kx_2}{(k+1)x_2 + kx_3}.$$

Of course, the usual genetic case is $k = 1$. In such a case, we obtain $\lambda = 2x_2 + x_3$ and the nondisjunctional rate $p = p_1 + p_2$. The invariance property of maximum-likelihood estimators implies that $p_{\mathrm{ML}} = \hat{p}_1 + \hat{p}_2$ and interestingly, p_{ML} turns out to be

$$p_{\mathrm{ML}} = \frac{2x_2}{2x_2 + x_3}, \tag{5}$$

which is exactly Cooper's estimator, $\hat{p}$ in (1).

The exact distribution of $\hat{p}$: Focusing on the case $p_1 = p_2$ and letting $p = p_1 + p_2$, we can rewrite (4) as

$$P(X_2 = x_2, X_3 = x_3) = \left(e^{-\lambda p/2}\frac{(\lambda p/2)^{x_2}}{x_2!}\right)\left(e^{-\lambda(1-p)}\frac{(\lambda(1-p))^{x_3}}{x_3!}\right), \tag{6}$$

with $\lambda = 2x_2 + x_3$. By defining a transformation as $y_2 = 2x_2 + x_3$ and $y_3 = 2x_2/(2x_2 + x_3)$, we can derive the joint p.m.f. of (Y_2, Y_3) using (6), and then get the marginal exact p.m.f. of Y_3

$$P(Y_3 = y_3) = \sum_{y_2 \text{ possible}} p(Y_2 = y_2, Y_3 = y_3), \tag{7}$$

which is the p.m.f. of $\hat{p}$. This distribution could be obtained numerically and an R script is available upon request.

ASYMPTOTIC RESULTS WITHOUT POISSON ASSUMPTION

For the asymptotic properties of $\hat{p}$, if $N = n$ is known (equivalently, X_1 is observed), it is the classical parameter estimation problem of multinomial distribution. It is well known that $X_2/n \to p/2$ in probability, and $(X_2/n - p/2)/\sqrt{[(p/2)(1-p/2)]/n} \Rightarrow N(0,1)$, where the ⇒ means convergence in distribution. However, in this framework X_1 is not observed and N is unknown. Hence, we cannot apply the existing results.

We study the asymptotic properties of $\hat{p}$ with more general assumptions, and the asymptotic properties of $\hat{p}_x - \hat{p}_y$, which allow the testing of differences between two nondisjunctional rates.

One nondisjunction rate: Let the number of progeny produced in an experiment, N_n, be a random variable taking only nonnegative integer values with a probability distribution $P(N_n = k)$. Each individual progeny can only have three possible outcomes (nondisjunctional dead, nondisjunctional living, and normal), and progeny are independent of each other. Let the probabilities of a progeny being in the three categories be $(p/2,\ p/2,\ 1-p)$. If X_i denotes the number of progeny resulting in outcome $i (i = 1, 2, 3)$, then the joint p.m.f. of $(X_1,\ X_2,\ X_3)$ given $N_n = k$ is the multinomial distribution $M(p/2,\ p/2,\ 1 - p;k)$, whose p.m.f. is given by Equation 2.

THEOREM 1. *Assume that $\{N_n\}$ is a sequence of random variables such that $E(N_n) = cn$ and $N_n/n \to c$ in probability for a constant c. Moreover, assume that as $n \to \infty$, $(N_n/(2X_2 + X_3) - 1)\sqrt{2X_2 + X_3} \to 0$ in probability. Then, Cooper's estimator $\hat{p}$ has the following property: (1) $\hat{p} \to p$ in probability, and (2) $(\hat{p} - p)/\sqrt{[\hat{p}(2 - \hat{p})]/(2X_2 + X_3)} \Rightarrow N(0,1)$.*

Remark 1. The assumptions of Theorem 1 are necessarily met by a Poisson distribution for N.

The proof of this remark as well as all the theorems are provided in the Appendix.

Similar to the usual normal approximation to the binomial, we require that $(2X_2 + X_3)\hat{p} = 2X_2 \geq 5$ and $(2X_2 + X_3)(1 - \hat{p}) = X_3 \geq 5$ to ensure a good approximation as our simulation demonstrates. On the basis of the above theorem, we can easily obtain the $(1 - \alpha)$ 100% confidence interval for p as $\hat{p} \pm z_{\alpha/2}\sqrt{[\hat{p}(2 - \hat{p})]/(2X_2 + X_3)}$. For hypothesis testing with H_0: $p = p_0$ *vs.* H_1: $p > p_0$ (for example), let $Z_1 = \hat{p} - p_0/\sqrt{p_0(2 - p_0)/(2X_2 + X_3)}$. Then, the decision rule at significance level α is to reject H_0 if $Z_1 > z_\alpha$.

The difference of two nondisjunction rates: Suppose that there are two progeny populations X and Y. We observed X_2, Y_2, X_3, Y_3 as the number of nondisjunctional living and regular normal progeny for both populations. We would like to assess whether the nondisjunction rates of two populations are statistically different from each other. Specifically, we are interested in testing: H_0: $p_x - p_y = \delta_0$ *vs.* H_1: $p_x - p_y \neq \delta_0$, for example, or in constructing the confidence interval of $p_x - p_y$ Similarly, let the number of progeny from the X population be N_n, and the number of progeny from the Y population be M_m, where both N_n and M_m are

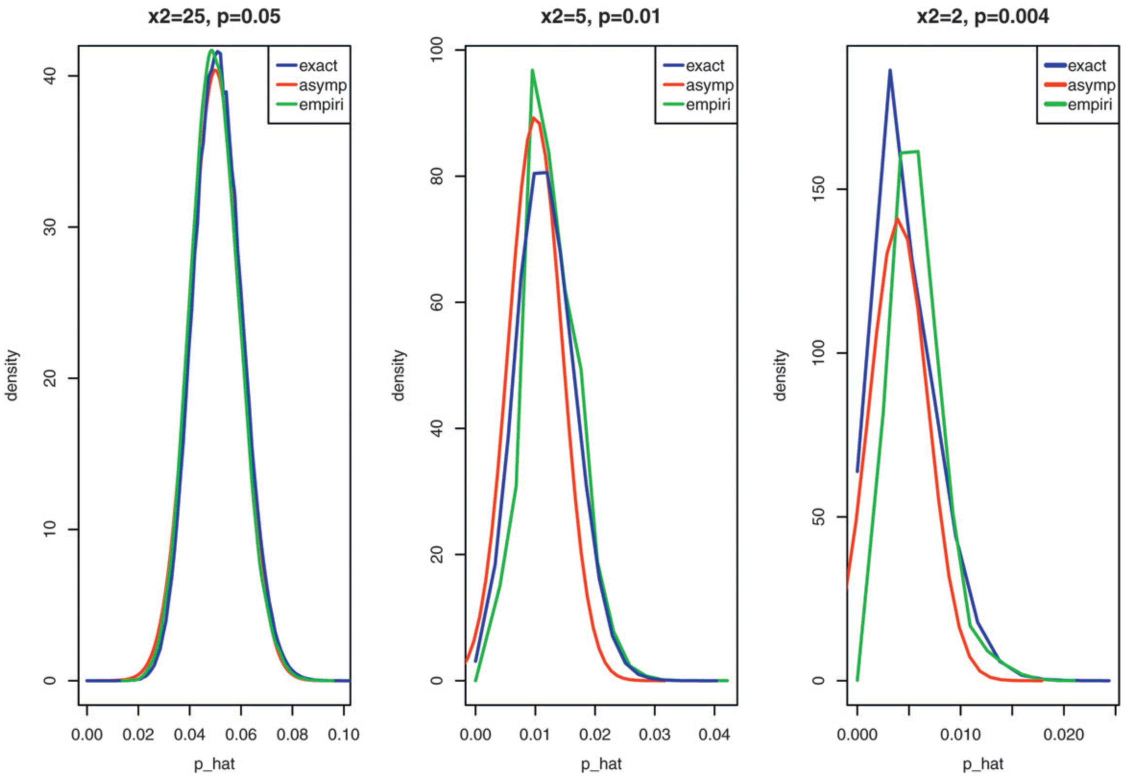

FIGURE 1.—Comparison of the exact distribution and the asymptotic distribution of the nondisjunction rate. Three cases are considered, the nondisjunction rates being 0.05, 0.01, 0.004, to show the difference of two distributions. The green curve is the empirical distribution of p with 50,000 simulations.

random variables. Let the probabilities of a progeny's outcome being in the three categories (X_1, X_2, X_3) be ($p_x/2$, $p_x/2$, $1 - p_x$) in the X population, and the probabilities of a progeny's outcome being in the three categories (Y_1, Y_2, Y_3) be ($p_y/2$, $p_y/2$, $1 - p_y$) in the Y population. We define

$$\hat{p}_x = \frac{2X_2}{2X_2 + X_3} \text{ and } \hat{p}_y = \frac{2Y_2}{2Y_2 + Y_3}.$$

THEOREM 2. *Assume that* $\{N_n\}$ *is a sequence of random variables such that* $E(N_n) = c_1 n$ *and* $N_n/n \to c_1$ *in probability for a constant* c_1. *Assume that* $\{M_m\}$ *is a sequence of random variables such that* $E(M_m) = c_2 m$ *and* $M_m/m \to c_2$ *in probability for a constant* c_2. *Moreover, assume that as*

$$n \to \infty, \left(\frac{N_n}{2X_2 + X_3} - 1\right)\sqrt{2X_2 + X_3} \to 0$$

in probability, and as

$$m \to \infty, \left(\frac{M_m}{2Y_2 + Y_3} - 1\right)\sqrt{2Y_2 + Y_3} \to 0 \text{ in probability,}$$

then: (1) $\hat{p}_x - \hat{p}_y \to p_x - p_y$ *in probability, and* (2)

$$\hat{p}_x - \hat{p}_y - (p_x - p_y) \Big/ \sqrt{\frac{\hat{p}_x(2 - \hat{p}_x)}{2X_2 + X_3} + \frac{\hat{p}_y(2 - \hat{p}_y)}{2Y_2 + Y_3}} \Rightarrow N(0, 1).$$

Similarly, the Poisson assumptions of N_n and M_m satisfy the assumptions of Theorem 2.

Again, we require that $(2X_2 + X_3)\hat{p}_x = 2X_2 \geq 5$ and $(2X_2 + X_3)(1 - \hat{p}_x) = X_3 \geq 5$ as well as $(2Y_2 + Y_3)\hat{p}_y = 2Y_2 \geq 5$ and $(2Y_2 + Y_3)(1 - \hat{p}_y) = Y_3 \geq 5$ to ensure a good approximation. On the basis of the above theorem, we can easily obtain the $(1 - \alpha)100\%$ confidence interval for $p_x - p_y$ as

$$\hat{p}_x - \hat{p}_y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(2 - \hat{p}_x)}{2X_2 + X_3} + \frac{\hat{p}_y(2 - \hat{p}_y)}{2Y_2 + Y_3}}.$$

For hypothesis testing with H_0: $p_x - p_y = \delta_0$ *vs.* H_1: $p_x - p_y \neq \delta_0$ (for example), let

$$Z_2 = (\hat{p}_x - \hat{p}_y - \delta_0) \Big/ \sqrt{\frac{\hat{p}_x(2 - \hat{p}_x)}{2X_2 + X_3} + \frac{\hat{p}_y(2 - \hat{p}_y)}{2Y_2 + Y_3}}.$$

Then, the decision rule at significance level α is to reject H_0 if $|Z_2| > z_{\alpha/2}$. Finally, for the future experiment with the expected difference as δ_0, the sample size can be calculated as $n = (p_x(2 - p_x) + p_y(2 - p_y))(z_{\alpha/2} + z_\beta)^2/\delta_0^2$ with power $1 - \beta$ and probability of type I error as α. For readers not interested in the derivation, the final equations are summarized in File S1.

COMPARISON OF THE EXACT AND THE ASYMPTOTIC DISTRIBUTIONS

In this study, we present two ways of getting the distribution of nondisjunction rate estimator $\hat{p}$. The exact distribution of $\hat{p}$ is derived with stronger assumptions, namely, the Poisson distribution for the total number of progeny (N) with its mean equal to $2x_2 + x_3$. The asymptotic results are derived with weaker assumptions and are applicable as long as N satisfies conditions in Theorem 1. The Poisson assumption of N is one special case where Theorem 1 can be applied. When the number of nondisjunctional living progeny (x_2) is not too small, usually $x_2 \geq 5$, the approximation is good. We demonstrate this by comparing the two distributions assuming there is a total of 1000 progenies for three cases: (1) $X_2 = 25$, $X_3 = 950$, then $p = 0.05$; (2) $X_2 = 5$, $X_3 = 990$, then $p = 0.01$; and (3) $X_2 = 2$, $X_3 = 996$, then $p = 0.004$.

We further generate the empirical distributions of p's under the three cases by simulations to see how our derived distributions matched the simulated ones. The detailed procedures are to first, simulate a N from a Poisson distribution with mean being 1000; second, simulate x_1, x_2, x_3 from a multinomial distribution with $(p/2, p/2, 1 - p)$; and third, calculate $\hat{p} = 2x_2/(2x_2 + x_3)$. The procedure is repeated 50, 000 times each, with p set to 0.05, 0.01, or 0.004, respectively, as shown in Figure 1. When X_2 is large, the assumptions for asymptotic results are well met and the three distributions (exact, asymptotic, and empirical) are almost identical (case 1 and 2). When X_2 is small ($2X_2 < 5$ and p is close to 0) (case 3), the asymptotic density deviates more from the exact distribution, but still in good agreement. These results show that the asymptotic normal distribution is a very good approximation of the exact distribution. In the extreme case that the data are not well modeled by the Poisson distribution, the asymptotic results are still valid. We suggest using the asymptotic results for constructing confidence intervals and doing hypothesis tests unless either $2X_2$ or X_3 is small (<5). As nondisjunction assays in Drosophila usually have sample sizes of at least several hundred, this condition is most likely to be violated in cases where the value of p is close to $1/N$.

ANALYSIS USING REAL DATA

Case study I: The common objectives for doing a nondisjunction assay include estimating the nondisjunction rate and testing if two genotypes have rates that are statistically significantly different. In the first example, we compare results of point estimation and hypothesis tests between the asymptotic results derived in this study and the asymptotic results assuming the traditional binomial distribution. As we discussed, most published literature has used the binomial distribution to model the nondisjunctional event as Binomial (N, p) assuming that N is observed and $N = 2X_2 + X_3$. With this assumption, the estimator turns out to be the same as one in this study, $\hat{p} = 2x_2/(2x_2 + x_3)$, but the standard deviation is calculated as $\sqrt{\hat{p}(1 - \hat{p})/2X_2 + X_3}$. This approximation ignores the fact that the number of nondisjunctional dead progeny is an unobserved random number. When this randomness is accounted for, as we do in this study, the standard deviation is calculated as $\sqrt{\hat{p}(\mathbf{2} - \hat{p})/2X_2 + X_3}$, which is at least 1.414 times as large as the one calculated with the binomial distribution (Figure 2). Unlike the binomial assumption that the standard deviation reaches to the largest when $p = 0.5$, under the multinomial assumption, the standard deviation of p increases as p increases. Therefore, as p gets larger, the ratio between these two standard deviations gets larger. We illustrate this using a published data set (Zhang and Hawley 1990). This study tested nondisjunction rates from a number of different mutant alleles of the gene *nod*. The estimated X nondisjunctional rate for these mutants is around 0.5 (Table 1). The standard deviation calculated using our asymptotic results is always larger (1.74–1.83 times as large) and the difference tends to increase as p gets larger.

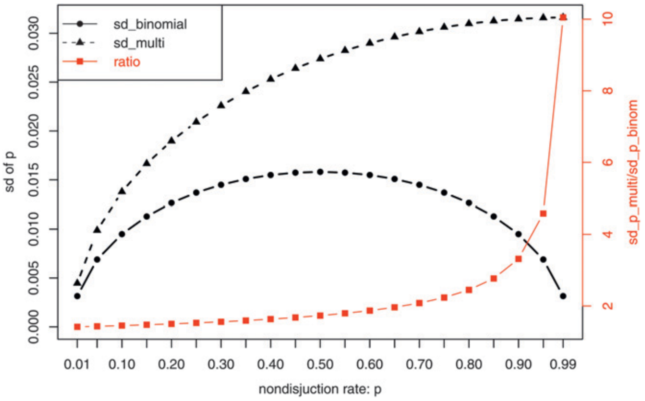

Figure 2.—Comparison of standard deviations calculated using asymptotic results assuming a binomial distribution $\left(\sqrt{\hat{p}(1 - \hat{p})/(2X_2 + X_3)}\right)$ and asymptotic results assuming a multinomial distribution as proposed in this study $\left(\sqrt{\hat{p}(2 - \hat{p})/(2X_2 + X_3)}\right)$. In the plot, we use $2X^2 + X^3 = 1000$.

Taking this randomness into consideration also has a large effect in terms of hypothesis tests. For comparing two nondisjunction rates p_x and p_y, our results show that

$$\hat{p}_x - \hat{p}_y - (p_x - p_y)\Bigg/\sqrt{\frac{\hat{p}_x(2 - \hat{p}_x)}{2X_2 + X_3} + \frac{\hat{p}_y(2 - \hat{p}_y)}{2Y_2 + Y_3}} \Rightarrow N(0, 1)$$

under the null hypothesis, which is different from

TABLE 1

Comparisons of standard deviation calculation between the asymptotic results derived in this study and the asymptotic results assuming the binomial distribution

	FM7a, nod^{b27}/nod^{a}	*FM7a*, nod^{b34}/nod^{a}	*FM7a*, nod^{b9}/nod^{a}	*FM7a*, nod^{b1}/nod^{a}	*FM7a*, nod^{b17}/nod^{a}	*FM7a*, nod^{b29}/nod^{a}	*FM7a*, nod^{bd}/nod^{a}
Regular	1167	639	844	897	2566	598	639
X NDJ	661	323	527	573	1319	400	378
Total	1828	962	1371	1470	3885	998	1017
Adj.total	2489	1285	1898	2043	5204	1398	1395
X NDJ rate	0.5311	0.5027	0.5553	0.5609	0.5069	0.5722	0.5419
std1 (asymp.normal)	0.0177	0.0242	0.0206	0.0199	0.0121	0.0242	0.0238
std2 (binomial)	0.0100	0.0139	0.0114	0.0110	0.0069	0.0132	0.0133
Ratio (std1/std2)	1.77	1.74	1.80	1.81	1.74	1.83	1.78

The data are taken from Zhang and Hawley (1990), which studied nondisjunction rates from a number of different mutant alleles of the gene *nod*.

$$\hat{p}_x - \hat{p}_y - (p_x - p_y) \Big/ \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{2X_2 + X_3} + \frac{\hat{p}_y(1-\hat{p}_y)}{2Y_2 + Y_3}} \Rightarrow N(0, 1),$$

when N is assumed to be observed. When we test if all seven mutants have the same nondisjunction rates by pairwise comparison (Table 1), we found that there are no statistically significant differences among them with the family-wise error rate ≤ 0.05 (Bonferroni multitest correction). This is consistent with the genetic analysis of these alleles, which appear to act as complete nulls that have lost all gene function. In contrast, using the same multitest correction method with asymptotic results derived from the traditional binomial distribution, the *b*34 and *b*17 alleles appear to be significantly different from *b*9, *b*1, and *b*29 (Table 2). This suggests that the genetic analysis is wrong and that these alleles retain some residual function. However, in light of our current analysis, the traditional binomial method would appear to yield false-positive results caused by ignoring the randomness in the number of nondisjunctional dead progeny.

Case study II: In the second data set, a collection of fly lines isolated from nature that had been used in a population genetics sequencing project for meiotic genes (Anderson *et al.* 2009) was assayed for their *X* nondisjunction rates. The nondisjunction rates observed among these lines were small (ranging from $p = 0$ to $p = 0.014$; Table 3). After multitest correction to control the FDR $\Leftarrow$ 0.05 (Benjamini and Hochberg 1995), the line MW9X showed a significant difference with several other lines (marked with * in Table 3, *P*-values = 0.05) with changes ranging between 6- and 20-fold. This result shows that while these lines do not carry alleles of large effect, such as those isolated by a screen of natural variation (Sandler *et al.* 1968), these assays have nonetheless successfully identified naturally occurring phenotypic variation in the trait of meiotic segregation. This is consistent with the genotypic variation identified in these same natural populations having phenotypic consequences as well. While these phenotypic differences are only just statistically significant at these sample sizes, at the population level these differences should clearly be subject to natural selection. This result also raises several experimental design considerations, such as when designing an assay to compare the nondisjunction rate for alleles of small effect, what sample size would be needed to reject H_0: $p_x - p_y = 0$ with 80% power? For example, if the values of p for two lines differ by 1% (*e.g.*, $p_x = 0.005$, $p_y = 0.015$), a sample size of 2338 per group is required to achieve a power of at least 0.8 with a two-sided significance level of 0.05. In Table 4, we list the sample size required for pairwise comparisons of a list of nondisjunction rates, ranging from 0.01 to 0.31. This table indicates that if the expected difference in rates is quite large (*e.g.*, 20% *vs.* 1%, as might be seen in comparing a mutant to a mutant plus rescue construct) then sample sizes of only a few hundred would be more than sufficient. Conversely, as the real rates under consideration become closer, the needed sample size becomes much larger and quickly becomes experimentally intractable. This indicates that any experimental outcome that hinges on nondisjunction rates being different by only 1% or 2% should be viewed with great skepticism.

DISCUSSION

The nondisjunction rate is an important parameter in the study of meiosis. We have studied the statistical properties of the currently widely used Cooper's estimator $\hat{p}$, which is $2X_2/(2X_2 + X_3)$. Under stringent assumptions, the estimator turns out the be the MLE and the exact distribution of $\hat{p}$ could be obtained numerically. When p is not too close to 0 and the observed nondisjunctional progeny (X_2) is not too small ($2X_2 \leq 5$), $\hat{p}$ is shown to have an asymptotic normal distribution (Theorem 1), and the asymptotic distribu-

TABLE 2

Comparisons of Bonferroni-adjusted *P*-values from the hypothesis tests between the asymptotic results derived in this study (conditional multinomial) and the asymptotic results assuming the binomial distribution using data in Table 1

genotype1	genotype2	adjp(Multinomial)	adjp(Binomial)
nod^{b34}	nod^{b9}	1	0.0737
nod^{b34}	nod^{b1}	1	0.0218
nod^{b34}	nod^{b29}	0.8879	0.0063
nod^{b17}	nod^{b9}	0.8879	0.0060
nod^{b17}	nod^{b1}	0.4232	0.0007
nod^{b17}	nod^{b29}	0.3276	0.0003

TABLE 3

***X* nondisjunctional rate estimates for lines from two natural populations (North American and Africa)**

Line	NonX	Normal	X nondis rate (std)
301	0	177	0 (–)
303	3	1905	0.0031(0.0018)
304*	2	3818	0.0010(0.0007)
306	0	1601	0 (–)
319	2	2295	0.0017(0.0012)
322	4	3826	0.0021(0.0010)
335	3	2784	0.0022(0.0012)
336	7	3168	0.0044(0.0017)
350	7	3843	0.0036(0.0014)
357	3	2658	0.0023(0.0013)
358	6	2908	0.0041(0.0017)
359	2	525	0.0076(0.0053)
361	3	3651	0.0016(0.0009)
375	6	2650	0.0045(0.0018)
390	5	1122	0.0088(0.0039)
397	0	664	0 (–)
399*	1	3053	0.0007(0.0007)
732	4	1845	0.00439(0.0022)
740*	1	2691	0.0007(0.0007)
774	2	1222	0.0033(0.0023)
MW11-3	0	218	0 (–)
MW25X	2	2909	0.0014(0.0010)
MW27X*	1	1937	0.0010(0.0010)
MW28X	0	159	0 (–)
MW28-5	0	148	0 (–)
MW38X	6	2155	0.0055(0.0023)
MW46-1	0	499	0 (–)
MW6-3II	1	1482	0.0013(0.0013)
MW6X	3	1526	0.0039(0.0023)
MW9-2	1	919	0.0022(0.0022)
MW9-4*	1	2162	0.0009(0.0009)
MW9X	14	2024	0.0136(0.0036)

Nondisjunction rates were measured by crossing wild-type females to *y cv v f car*/B^SY males under standard conditions, which allowed identification of nondisjunctional progeny as multiply-marked males (*XO*) or B^S females (*XXY*). The numbers in parentheses are the standard deviations of the *X* nondisjunction rates. The lines marked with * have significantly different nondisjunction rates when compared to MW9X.

tion approximates the exact distribution well when *p* is large. In the real data analysis, we suggest use of asymptotic results whenever possible because it requires no specific distribution on *N*. Unless both $2X_2$ and X_3 are small (<5), the asymptotic normal distribution is a good approximation of the exact distribution as shown in our simulation study. The use of the normal approximation also enables us to apply classical statistical tools to this problem. For example (as shown in Table 4), the power/sample size calculation can be carried out and this can provide experimental guidelines for designing nondisjunction assays. Statistical significance tests (*P*-value calculation) also can be carried out on the basis of Theorem 2. We provide a MS EXCEL file to do these calculations as supporting information material in File S2.

The analysis of nondisjunction data using this framework suggests several important conclusions. The first is that as nondisjunction rates approach zero, the number of nondisjunctional progeny expected approaches zero. It is in this region that the random number of progeny surviving fertilization has its greatest effect on the estimated rate. Second, even for cases where *p* is far from zero, the variance of this process is greater than that of a binomial. The practical impact of this is clearly seen in our analysis of the published *nod* nondisjunction data (Zhang and Hawley 1990). While the genetic analysis indicated that the *nod* alleles were complete nulls, the binomial approach finds that their nondisjunction rates are statistically significantly different from one another, suggesting that these alleles retain at least some residual function. When the increased variance due to lethal aneuploidy after fertilization is accounted for, the differences are no longer significant, which is consistent with the genetic analysis. This avoidance of an apparent false-positive result is a clear benefit to using the multinomial approach. Third, this suggests that differences in the nondisjunction rate of less than around 2% may simply not be amenable to direct experimental analysis, even with sample sizes of several thousand. This is a point of concern for population genetics, as variants that reduced nondisjunction by even a fraction of a percent should be advantageous and undergo positive selection in species as numerous as Drosophila. Our results suggest that any experimental program working with alleles of small effect should consider the use of sensitized assays, where the genetic background is weakened so that small genotypic differences are magnified to an experimentally tractable level (Zwick *et al.* 1999). Finally, while increasing sample sizes does decrease confidence intervals, sample size increases rapidly experience diminishing returns. As a rule of thumb, Table 4 appears to show that reasonable statistical payoffs (such as reduction of sizes of confidence intervals) in increasing sample sizes from ~100 to ~1000, but very little improvement in increasing sample sizes from ~3000 to > 10,000. The exact sample sizes aimed for in an experiment should be considered in light of the

TABLE 4

The expected sample size needed from each of two populations with the given nondisjunction rates (p_x and p_y) to declare a statistically significant difference with probability of type I error of $\alpha = 0.05$ and power of $\beta = 0.90$

	p_y					
p_x	0.05	0.1	0.15	0.2	0.25	0.3
0.01	771	273	160	111	84	67
0.06	22,476	2,013	511	256	162	115
0.11	892	41,810	3,188	737	346	209
0.16	341	1,414	60,092	4,298	950	432
0.21	195	492	1,908	77,325	5,342	1,150
0.26	132	264	634	2,372	93,506	6,321
0.31	97	171	329	768	2,807	108,637

data's intended purpose to meet research goals without wasted efforts.

In the current work, we have considered only estimating the rate of *X* nondisjunction in female meiosis. The small *4* chromosome can also be used in nondisjunction assays, as *triplo-4* progeny are viable and can therefore be observed. By mating experimental females to males bearing a *compound-4*, both normal and nondisjunctional oocytes have the same 50% chance of being fertilized by the type of sperm that results in viable progeny. This means that the rate of nondisjunction is expected to be equal to the proportion of nondisjunctional progeny observed, without the doubling used in Cooper's estimator for *X* chromosome nondisjunction. In light of our current results, it is clear that the use of a binomial model for *4* nondisjunction would also underestimate the true size of the confidence intervals. A preliminary examination of this process suggests that as random survival is applied to all progeny, instead of solely to the nondisjunctional classes, the increase in variance of estimates of *4* nondisjunction rates due to sperm chromosome content may be even greater than that for the *X*. This appears to be because in the *X*-only case the 50% chance of dying from fertilization by the wrong sperm is applied solely to nondisjunctional progeny, while all of the normal progeny are assumed to survive. In the *4*-only case, the same 50% chance of dying is applied to both nondisjunctional and normal progeny. Therefore, while the value of $\hat{p}$ is equal to the observed proportion of nondisjunctional progeny observed, the variance of *4*-only nondisjunction should be greater than that of the *X*-only case. Furthermore, in practice nondisjunction for the *X* and *4* are often scored simultaneously. This practice is biologically relevant, as it has revealed the intriguing observation that rates of *X* and *4* nondisjunction are often found in a 2:1 ratio across certain classes of mutants (Zitron and Hawley 1989; Sekelsky *et al.* 1999). In this case, as *X* nondisjunctional oocytes have only a 25% chance of being viable after fertilization, this should result in an even larger increase in the variance than that of the *X*-only case. Therefore, researchers should be aware that when *compound-4* is used to simultaneously measure *X* and *4* nondisjunction, our method for calculating confidence intervals for *X* nondisjunction rates will be an underestimate of the true interval. We are continuing to study the process of *X* and *4* nondisjunction and hope to be able to develop similar multinomial results for the *4*-only and *X/4* simultaneous cases in the future.

The authors thank Boris Rubinstein and Arcady Mushegian for helpful discussion and comments, the editor, and two anonymous reviewers for their helpful suggestions to improve the manuscript. This work was supported by a Stowers Summer Scholarship to N.M.S., an American Cancer Society Research Professorship to R.S.H., and an American Cancer Society Postdoctoral Fellowship to W.D.G.

LITERATURE CITED

Anderson, J. A., W. D. Gilliland and C. H. Langley, 2009 Molecular population genetics and evolution of Drosophila meiosis genes. Genetics **181:** 177–185.

Baker, B. S., and A. T. Carpenter, 1972 Genetic analysis of sex chromosomal meiotic mutants in *Drosophila melanogaster.* Genetics **71:** 255–286.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B Methodol. **57:** 289–300.

Bridges, C. B., 1916 Non-disjunction as proof of the chromosome theory of heredity. Genetics **1:** 1–52.

Casella, G., and L. B. Berger, 2001 *Statistical Inference*, Ed. 2. Duxbury Press, Pacific Grove, CA.

Chung, K. L., 1974 *A Course in Probability Theory*, Ed. 2. Academic Press, New York.

Cooper, K. W., 1948 A new theory of secondary non-disjunction in female *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **34:** 179–187.

Sandler, L., D. L. Lindsley, B. Nicoletti and G. Trippa, 1968 Mutants affecting meiosis in natural populations of *Drosophila melanogaster.* Genetics **60:** 525–558.

Sekelsky, J. J., K. S. McKim, L. Messina, R. L. French, W. D. Hurley *et al.*, 1999 Identification of novel Drosophila meiotic genes recovered in a P-element screen. Genetics **152:** 529–542.

Zhang, P., and R. S. Hawley, 1990 The genetic analysis of distributive segregation in *Drosophila melanogaster.* II. Further genetic analysis of the nod locus. Genetics **125:** 115–127.

Zitron, A. E., and R. S. Hawley, 1989 The genetic analysis of distributive segregation in *Drosophila melanogaster.* I. Isolation and characterization of Aberrant *X* segregation (AXS), a mutation defective in chromosome partner choice. Genetics **122:** 801–821.

Zwick, M. E., J. L. Salstrom and C. H. Langley, 1999 Genetic variation in rates of nondisjunction: association of two naturally occurring polymorphisms in the chromokinesin *nod* with increased rates of nondisjunction in *Drosophila melanogaster.* Genetics **152:** 1605–1614.

Communicating editor: I. Hoeschele

APPENDIX

Proof of Theorem 1: The key result to obtain the asymptotic properties of $\hat{p}$ is the following Chung's lemma, which is Theorem 7.3.2 (Chung 1974).

Lemma 1. *Suppose that* $\{X_i,\ i \geq 1\}$ *is a sequence of i.i.d. random variables with mean* 0 *and variance* 1. *Define* $S_n = \sum_{i=1}^{n} X_i$. *Let* $\{\gamma_n, n \geq 1\}$ *be a sequence of random variables taking only strictly positive integer values (can be relaxed to "taking only nonnegative integers") such that* $\gamma_k/k \to c$ *in probability, where c is a positive constant. Then,* $S_{\gamma_n}/\sqrt{\gamma_n} \Rightarrow N(0,1)$.

The proof of Theorem 1 also relies on the two lemmas below (their proofs are available upon request).

Lemma 2.

$$\frac{2X_2}{N_n} \to p, \text{ in } L^2 \text{ and in probability.}$$

Lemma 3.

$$\frac{N_n}{2X_2 + X_3} \to 1 \text{ in } L^2 \text{ and in probability.}$$

Lemma 4.

$$p\left(\frac{N_n}{2X_2 + X_3} - 1\right)\Bigg/\sqrt{\frac{p(2-p)}{2X_2 s + X_3}} \to 0 \text{ in } L^2 \text{and in probability.}$$

Proof. Since $\hat{p} = 2X_2/N_n/((2X_2 + X_3)/N_n)$, Lemmas 2 and 3 imply the consistency of $\hat{p}$, namely, $\hat{p} \to p$ in probability.

Observe that $\left\{(X_{i2} - p/2)\big/\sqrt{(p/2)(1-p/2)}\right\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1. Then,

$$S_n = \sum_{i=1}^{n} \frac{X_{i2} - p/2}{\sqrt{(p/2)(1-p/2)}} = \left(\frac{X_2}{n} - p/2\right)\Bigg/\frac{\sqrt{(p/2)(1-p/2)}}{n}.$$

So, Chung's lemma and the assumption imply

$$\frac{S_{N_n}}{\sqrt{N_n}} = \frac{(X_2/N_n) - (p/2)}{\sqrt{(p/2)(1-p/2)/N_n}} = \frac{(2X_2/N_n) - p}{\sqrt{p(2-p)/N_n}} \Rightarrow N(0,1). \tag{8}$$

Next, consider

$$\frac{(2X_2/(2X_2 + X_3)) - p}{\sqrt{p(2-p)/(2X_2 + X_3)}} = \left(\frac{(2X_2/N_n) - p}{\sqrt{p(2-p)/N_n}}\right)\sqrt{\frac{N_n}{2X_2 + X_3}} + \frac{p((N_n/(2X_2 + X_3)) - 1)}{\sqrt{p(2-p)/(2X_2 + X_3)}}. \tag{9}$$

Slutsky's theorem with (8) and Lemmas 3 and 4 imply

$$\frac{\hat{p} - p}{\sqrt{p(2-p)/(2X_2 + X_3)}} \Rightarrow N(0,1). \tag{10}$$

Finally, observe that $(\hat{p} - p)\big/\sqrt{\hat{p}(2-\hat{p})/(2X_2 + X_3)} = \left((\hat{p} - p)/\sqrt{p(2-p)/2X_2 + X_3}\right)\left(\sqrt{p(2-p)}/\sqrt{\hat{p}(2-\hat{p})}\right)$. The consistency of $\hat{p}$ implies $\sqrt{p(2-p)}/\sqrt{\hat{p}(2-\hat{p})} \to 1/\sqrt{p(2-p)}$ in probability. Together with (10), Slutsky's theorem gives the desired asymptotic normal result. ■

Proof of Remark 1: With the Poisson assumption, N_n becomes N_λ having a Poisson distribution with parameter λ. It is well known that $E(N_\lambda) = \lambda$ and $N_\lambda/\lambda \to 1$ in probability. The first assumption is satisfied. To check the second assumption, it suffices to show the L^1 convergence by Markov inequality. Observe that $E(2X_2 + X_3) = E(N_\lambda) = \lambda$. Applying Cauchy–Schwartz inequality, we have

$$\begin{aligned} E\left[\left|\frac{N_\lambda}{2X_2 + X_3} - 1\right|\sqrt{2X_2 + X_3}\right] &\leq \left[E\left(\frac{N_\lambda}{2X_2 + X_3} - 1\right)^2\right]^{1/2}[E(2X_2 + X_3)]^{1/2} \\ &= \left[E((N_\lambda - 2X_2 - X_3)^2\frac{1}{(2X_2 + X_3)^2}\right]^{1/2}\lambda^{1/2} \leq E\left[\frac{1}{(2X_2 + X_3)^6}\right]^{1/6}[E(N_\lambda - 2X_2 - X_3)^3]^{1/3}\lambda^{1/2}. \end{aligned}$$

The last inequality comes by applying Hölder's inequality with $p = 3$ and $q = \frac{3}{2}$. It goes to zero because it can be shown that given the first assumption of Theorem 1, $[E(2X_2 + X_3)^{-6}]^{1/6} = O(\lambda^{-1})$ and $[E(N_\lambda - 2X_2 - X_3)^3]^{1/3} = O(\lambda^{1/3})$.

Proof of Theorem 2: *Proof.* Observe that the two samples are independent. The consistency follows immediately. With the assumptions, we can apply Theorem 1 to each sample and obtain

$$\begin{pmatrix} \dfrac{\hat{p}_1 - p_1}{\sqrt{\hat{p}_1(2-\hat{p}_1)/(2X_2 + X_3)}} \\ \dfrac{\hat{p}_2 - p_2}{\sqrt{\hat{p}_2(2-\hat{p}_2)/(2Y_2 + Y_3)}} \end{pmatrix} \Rightarrow \mathbf{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{11}$$

Let $g(x, y) = x - y$ and $\mathbf{D} = (\partial g/\partial x, \partial g/\partial y)$. Observe that $g(\hat{p}_1, \hat{p}_2) = \hat{p}_1 - \hat{p}_2$ and $\mathbf{D} = (1, -1)$. Then, the asymptotic normality comes from applying the multivariate δ methods.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.118778/DC1

Statistical Analysis of Nondisjunction Assays in Drosophila

Yong Zeng, Hua Li, Nicole M. Schweppe, R. Scott Hawley and William D. Gilliland

DOI: 10.1534/genetics.110.118778

File S1

Inference for one nondisjuntion rate

Let X_2 be the number of observed nondisjunctional living progeny, X_3 be the number of observed regular progeny in a study measuring X nondisjunction, then the nondisjunction rate p is calculated as:

$$\hat{p} = \frac{2X_2}{2X_2 + X_3}$$

The 95% confidence interval for p is:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(2-\hat{p})}{2X_2 + X_3}}.$$

Comparison of two nondisjunction rates

Suppose that there are two progeny populations X and Y. We observed X_2, Y_2, X_3, Y_3 as the number of nondisjunctional living and regular normal progeny for both populations. Nondisjunctional rates for two populates are calculated as:

$$\hat{p_x} = \frac{2X_2}{2X_2 + X_3},$$

and

$$\hat{p_y} = \frac{2Y_2}{2Y_2 + Y_3}$$

For hypothsis testing with $H_0 : p_x - p_y = \delta_0$ vs. $H_1 : p_x - p_y \neq \delta_0$, let

$$Z = \frac{\hat{p}_x - \hat{p}_y - \delta_0}{\sqrt{\frac{\hat{p}_x(2-\hat{p}_x)}{2X_2+X_3} + \frac{\hat{p}_y(2-\hat{p}_y)}{2Y_2+Y_3}}}$$

Then, the decision rule at significance level $\alpha = 0.05$ is to reject H_0 if $|Z| > 1.96$. For the specific test of whether p_x and p_y are equal, set δ_0 to 0.

Calculation of sample size

For an experiment designed to identify a difference between two nondisjunction rates $p_x - p_y = \delta_0$, the sample size per population required to achieve the desired power of 90% while controlling the probability of type I error as $\alpha = 0.05$ is calculated as:

$$n = \frac{(p_x(2-p_x) + p_y(2-p_y))(1.96+1.28)^2}{\delta_0^2}$$

FILE S2

Nondisjunction Assay Calculator

File S2 is available for download as an Excel file at http://www.genetics.org/cgi/content/full/genetics.110.118778/DC1.