Reexamining the Pooled Sampling Approach for Estimating Prevalence of Infected Insect Vectors

T. A. EBERT,[1] R. BRLANSKY, AND M. ROGERS

Citrus Research and Education Center, University of Florida, 700 Experiment Station Road, Lake Alfred, FL 33850

Ann. Entomol. Soc. Am. 103(6): 827–837 (2010); DOI: 10.1603/AN09158

ABSTRACT Our goal was to estimate seasonal changes in the proportion of Asian citrus psyllid, *Diaphorina citri* Kuwayama (Hemiptera: Psyllidae), carrying *Candidatus* Liberibacter asiaticus. Our approach was to test Asian citrus psyllid by using pooled samples. The initial question was about pool size and the consequences of choosing poorly. Assuming no loss in sensitivity when diluting one infected individual with many healthy individuals, then it is recommend that a combination of all the published limits be used: keep the number of pools (n) above 20, the pool size (k) below 100, and the number of infected pools less than half the total number of pools. The most conservative approach to achieving the latter is to optimize pool size given an infection rate (p) such that $k = \ln(0.5)/\ln(1 - p)$. Exceeding these limits increases the probability that all the pools will be infected. If this occurs, then that particular sample will be discarded. Use of multiple pool sizes can be used to manage this risk, but this approach may not always be practical. PooledInfRate is a good program for estimating prevalence, and it is available for free from the Centers for Disease Control and Prevention (CDC). The program provides corrected confidence intervals for prevalence estimates using one or multiple pool sizes. We used a randomization test approach as a contrasting methodology. The bias corrected CDC 95% confidence interval is an upper bound to the "true" 95% confidence interval, and we provide an estimate of the magnitude of the remaining bias in the estimate.

KEY WORDS pooling, prevalence, confidence Intervals, vector, *Diaphorina*

Detecting a rare event or estimating the occurrence of such an event in a larger population is a problem encountered in working with disease vectors. The event may be so rare that thousands of individuals need to be tested to observe the event, and the cost of such testing exceeds the available budget and personnel. A methodology is required to achieve project goals within such constraints. The first published example of a pooling methodology involved a hypothetical example of trying to detect rare infection in military inductees. A significant cost savings was possible by first testing a sample created by mixing the blood from several people. If a sample tested positive, then the blood from all individuals in that pool would be retested (Dorfman 1943). This is usually referred to as "group testing" or Dorfman type testing, where the goal is to identify an infected individual. A related problem is to estimate the proportion of individuals that are positive. The strategy has been called "pooling" when applied to estimating a proportion (Hepworth 1999, 2004). This application was proposed thrice in the 1960s (Gibbs and Gower 1960, Chiang and Reeves 1962, Thompson 1962) and was used extensively in insect vector research. Some recent examples include beet leafhopper, *Circulifer tenellus* (Baker), transmitting a phytoplasma (Crosslin et al. 2005); potato purple top phytoplasma transmitted by leafhoppers (Munyaneza et al. 2007); black fly vector of onchocerciasis in West Africa (Yameogo et al. 1999); psyllid vector of phytoplasmas (Carraro et al. 2004, Garcia-Chapa et al. 2005); and mosquitoes vectoring a viral pathogen in deer (Andreadis et al. 2008). The technique also was used in medicine (Novack et al. 2008), animal health (Rovira et al. 2008), fisheries (Wallace et al. 2008), plant health (Geng et al. 1983, Coutts et al. 2009), and DNA mapping (Chi et al. 2009).

In practice, pooling is a simple process. If 30,000 mosquitoes are collected from the field, they could be tested one at a time for a viral pathogen. If each test takes 10 min and costs US$15, then this project will take 5,000 h and cost US$450,000. A shorter approach would be to smash 10 mosquitoes together and test this pooled sample. This approach would take 500 h and cost US$45,000. Even greater savings are achieved with larger pool sizes. However, there is an obvious problem. If two individuals are infected and pool size is 15,000, with one infected individual per pool, then one might conclude that all individuals were infected because both pools tested positive. Work backwards: three infected individuals and a pool size of 10,000; 10 infected individuals and a pool size of 3,000; and so

[1] Corresponding author, e-mail: tebert@ufl.edu.

Table 1. Probability of all pools being infected given optimal pool sizes as estimated by various studies

No. pools	Chaing and Reeves (1962)[a] $k = \frac{Log_e\left(\frac{1}{2}\right)}{Log_e(1-p)}$		Thompson (1962) $k = \frac{1.5936 - p}{p}$		Burrows (1987) $k \cong \frac{-1.44}{Log_e(1-p)}$	
	$p = 0.001$	$p = 0.100$	$p = 0.001$	$p = 0.100$	$p = 0.001$	$p = 0.100$
5	7.8E-6	0.039	0.321	0.312	N.A.	N.A.
10	6.1E-11	0.002	0.103	0.097	N.A.	N.A.
20	3.7E-21	2.2E-6	0.011	0.009	0.005	0.005
40	1.4E-41	5.0E-12	1.0E-4	8.9E-5	2.0E-5	2.1E-5

Number of individuals in a pool is k, and p is the true prevalence.
[a] The formula used by different authors for selecting optimal k is shown for reference.

forth. The conclusion is that one always overestimates the true infection rate whenever all the pooled samples test positive and that this is true for any pool size. It is also the case that for any pool size and any nonzero infection rate there will be a nonzero probability that all the pools will be infected. So, there are benefits and risks to the pooling methodology.

Our interest in this methodology stems from a problem in estimating seasonal changes in the proportion of Asian citrus psyllid, *Diaphorina citri* Kuwayama (Hemiptera: Psyllidae), carrying *Candidatus* Liberibacter asiaticus, the putative causal agent of Huanglongbing disease (HLB). For management purposes, an estimate of monthly Liberibacter asiaticus prevalence in Asian citrus psyllid was needed, but prevalence was very low. Consequently, sample size exceeded our resources. We first encountered pooling methodology as a footnote to a table in a paper on leafhopper transmission of maize chlorotic dwarf virus (Hunt et al. 1988) but later discovered that the basic approach was already published (Gibbs and Gower 1960, Chiang and Reeves 1962, Thompson 1962, Hauck 1991). Deriving the equations requires a basic understanding of probability, permutations, and combinations (Loyer 1983, Ross 1984) (see *Appendix*).

Many methods for analysis of pooled samples have been proposed. The most basic is the minimum infection level (=minimum infection rate, = minimum field infection rate) where one assumes that a positive pool is infected by a single individual and the infection rate is the number of positive pools divided by the total number of individuals tested (Orshan et al. 2008, Vitek et al. 2008). This is an easily applied method, but the assumption that each pool has only one positive cannot usually be proven. At low infection rates and small pool sizes, the probability that a pool has multiple positive individuals may be small, but the definition of "low" and "small" may differ between scientists, and the imprecision of this approach is not warranted. Better alternatives are the exact methods (Walter et al. 1980, Swallow 1985) or a slightly better approach (Burrows 1987). There are also asymptotic approximation methods. Some based on the Fisher Information statistic (Kline et al. 1989), a Logit function (Hepworth 1999), a complimentary log-log function as part of a generalized linear model (Farrington 1992), or by using moments (Bondell et al. 2007). These methods have been extended to include dealing with estimating prevalence with pools of different size (Walter et al. 1980, Hepworth 1999, Biggerstaff 2008, Hepworth and Watson 2009). These studies and studies cited therein developed estimates for confidence intervals. Complications arise if one has error in detecting the pathogen either due to dilution of an infected individual with many healthy individuals or errors in the assay technique (Tu et al. 1995). A very useful program that gives point estimates and confidence intervals is at http://www.cdc.gov/ncidod/dvbid/westnile/software.htm (Biggerstaff 2006).

One of the primary concerns in using these methods is in choosing the pooling size (k). Increasing pool size can decrease the cost of testing thousands of individuals, and pooling may improve accuracy if there is error in detecting the pathogen (Tu et al. 1995). However, pooling degrades the quality of the estimate, and if all pools are infected then the estimated infection rate is 100%. Obviously, one does not plan an experiment with this as the likely outcome, but field results are variable. It is possible that the base infection rate is 1%, but that one in 50 samples will have 10% of the individuals infected. Thus, a few samples may have the majority of pools testing positive. Even with good preliminary data, the optimal pool sizes that have been proposed leave a risk of having all pools infected (Table 1). The probability that all pools are infected is $[1 - (1 - p)^k]^n$, with n pools of size k and an infection rate p (Sterne 1934). It should be noted that all studies mentioned in Table 1 modified their recommendations for optimal pool size. Thompson (1962) recommended using a pool size smaller than the optimal value by using his equation (Table 1) but did not provide a numerical value for "smaller." Chaing and Reeves (1962) put an upper limit of $k = 100$. Burrows (1987) put a lower bound to the number of pools at 20.

One of the "risks" in pooling is the case when all pools test positive. With ten pools of size ten it is possible that one individual in each pool could be positive and therefore a 10% infection rate in the population results in an estimated 100% infection rate of the sample. Ideally, this never happens, but with pooling there is always the risk that it will happen. If this happens the outcome from that sample is typically ignored, and so for that sample one has wasted the time and resources expended in processing the sam-

ple. If one discards any result where all the pools are infected, then one might consider adjusting confidence intervals to account for this practice – essentially treating this case as a source of bias in an estimated confidence interval. We disagree with this approach and therefore decided to try a different technique to gain a better understanding of the bias in the estimated confidence intervals for estimated prevalence rates from pooled samples.

The first goal of this research was to outline the basic approach in the analysis of pooled samples. The methodology is essentially discrete, and it is possible to enumerate all possible outcomes for small sample sizes. Thus, the second goal is to enumerate all possible outcomes for specific cases to better understand the limitations of this approach. Finally, virtual populations are used to test the quality of 95% confidence limits (CL) by using the Centers for Disease Control and Prevention (CDC) program compared with a unique application of a randomization method. Our approach is graphically oriented rather than developing more equations, but we need to start with the basic equations.

Methods

Equations (an Exact Method). The basic approach assumes that the probability that any given individual is infected is independent of the infection status of all other individuals, and the probability of infection is the same for all individuals within the sample (independent and identically distributed, or IID). In many cases, this is a poor assumption based on the biology of the system being studied. However, the assumption can be satisfied using sampling methods that randomize the infected individuals with the sampled population (Hepworth 1996). Given that this assumption is true, the exact estimate for the infection rate is calculated based on the following equations (Thompson 1962).

Let p = proportion of the population testing positive. In standard statistical notation, characters with a circumflex ("^") symbol represent estimates of true values. Thus, $\hat{p}$ is an estimate of p, and likewise for other variables.

N = number of individuals;

n = number of pools;

k = number of individuals per pool, or pool size; and

π is the probability that a pool is infected, and thus

$$\hat{\pi} = 1 - \sum_{i=1}^{n} \frac{X_i}{n} \qquad [1]$$

where X_i is coded one for healthy and 0 for infected.

The probability of a pool not being infected is the probability that none of the individuals are infected to the power of the numbers of individuals in a pool. So, the number of uninfected pools is $(1 - p)^k$, where p is unknown and k is the number of individuals in a pool. However, we measure the number of uninfected pools as

$$\sum_{i=1}^{n} \frac{X_i}{n}.$$

Equating these two equations gives

$$\sum_{i=1}^{n} \frac{X_i}{n} = (1 - p)^k \text{ or } \hat{p} = 1 - \left(\sum_{i=1}^{n} \frac{X_i}{n} \right)^{\frac{1}{k}} \qquad [2]$$

The expected value for the infection rate is therefore

$$E(\hat{p}) = 1 - E\left(\left(\sum_{i=1}^{n} \frac{X_i}{n} \right)^{\frac{1}{k}} \right)$$

The expected value for this equation is the fraction of healthy individuals multiplied by the number of ways of ordering infected and healthy individuals multiplied by the probability of getting i healthy individuals in n pools. This can be written as

$$E(\hat{p}) = 1 - \sum_{i=0}^{n} \left(\frac{i}{n} \right)^{\frac{1}{k}} \binom{n}{i} \cdot [(1 - p)^k]^i [1 - (1 - p)^k]^{n-i} \qquad [3]$$

There is a well known bias in the estimator of p (Gibbs and Gower 1960, Thompson 1962, Hepworth and Watson 2009), and methods have been developed to correct this bias (Hepworth 2005, Biggerstaff 2008, Hepworth and Watson 2009).

Virtual Population. No biological data were used in this effort. Using biological data would complicate the presentation and results. In biological data, one can only estimate the true prevalence of a disease. Sampling errors and error in the testing procedures can skew and distort the prevalence estimate. Such issues are of great importance in understanding biological data. However, we need to understand the simple case first and then build a foundation for evaluating the effects of sampling and testing issues.

The figures were all produced either by enumerating all possible outcomes from the method, or by using a computer to generate a virtual population. With a virtual population, we can select the true infection rate in the population; therefore, we are always testing a population with a known predefined infection rate. In any sampling methodology (virtual or real), we hope that our sample is representative of the entire population. A sample of 6,000 random individuals might have a mean body weight of 1 g with a SD of 0.5 g. However, if another sample of 6,000 were taken, it might have a mean of 1.01 g and a SD of 0.59 g. If another sample were taken, it, too, would be a little different from the others. The notation for 600 of these samples would set size (S) of 600 with N = 6,000. One then calculates a SD for S to describe the reliability of the methodology in testing a population. Thus, there are S = 600 sets of N = 6,000 pooled into n = 600 pools of k = 10 individuals. The analysis of the pools is used

Table 2. An Excel spreadsheet with cell elements used to create a virtual population (column A), identify infected individuals at an infection rate of 15% (column B), and pool it with a pool size of 5 (column D)

	A	B	C	D
1		0.15		5
2	= rand()	= if(a2<b1,1,0)	1	= if(mod($c2,$d$1) = 1,if(sum($b2:$b6)>0.5,1,0),"")
3	= rand()	= if(a3<b1,1,0)	2	= if(mod($c3,$d$1) = 1,if(sum($b3:$b7)>0.5,1,0),"")
4	= rand()	= if(a4<b1,1,0)	3	= if(mod($c4,$d$1) = 1,if(sum($b4:$b8)>0.5,1,0),"")
...	...	...	...	...

to estimate prevalence and get confidence intervals, but the analysis of sets is used to test the methodology.

We created virtual populations of 6,000 individuals and fixed specific infection rates within these populations. We created these populations as a column of random numbers in an Excel spreadsheet using the following methodology. The first cell in our Excel spreadsheet was A1 (Table 2). Column A had a random population created using a random number generator, and we filled cells A2:A6001 with this function. The true infection rate (p) was put into cell B1 (0.15 in Table 2). We then filled all cells B2:B6001 with a one if the corresponding cell in A2:A6001 was less than B1. Column C has positive integers 1 through 6,000. Cell D1 has the pool size. The remaining cells in D sample B and if any individual (A) was positive (B) the pool (D) is positive (Table 2). We edited the end of each column to make sure that we did not have any partly filled pools (e.g., $k = 17$, $n = 6{,}000/17 = 352.9$, so the final sample had 16 individuals and was deleted before proceeding). This gave us bounds of $k < 71$, $n > 90$, which were within the limits described for Table 1 ($k < 100$, $n > 20$).

We used the CDC methodology to calculate prevalence rates, and a 95% confidence interval. To contrast with this approach, we used a randomization technique to calculate a 95% confidence interval (Ebert et al. 1998).

Randomization Approach. To contrast with the CDC approach, we designed an alternative methodology for calculating confidence intervals. There are too many possible combinations of infection rates, pool sizes, and numbers of pools. For our application, an infection rate <0.5% is essentially zero because it is impractical to process sufficient samples to improve our resolution of the infection rate. In this exercise, we chose a lower infection rate because we need to make sure that our results here are applicable to our data. So, we selected a minimum infection rate of 0.1% and a pool size of 20 ($k = 20$). We then estimated prevalence at progressively increasing true prevalence rates until there was no further change in the response. In this case that was $p = 42.1\%$. We used the following process for each prevalence level from 0.001 through 0.421.

1. Create a virtual population of 6,000 individuals, and group in pools of 20 ($k = 20$).
2. Use equation 2 to calculate prevalence.
3. Repeat 1,000 times.
4. Use the CDC approach to get a confidence interval about each of these 1,000 prevalence estimates.
5. Use a randomization test to create a similar confidence interval.
   a. Take the 1,000 repeats in step 3, and take sets of observations. With a set size of four, one would take the first four observations and process them as outlined below, then take the next four observations and process them, and then the next four, and so forth. Then, one does the same process with five, six, . . . through a set size of 10.
   b. For each set, run a randomization test to get a 95% confidence interval about each set. Following the approach by Ebert et al. 1998, take a set of five numbers designated ABCDE, make a new set of AABBCCDDEE, randomize it and break it into two parts. One outcome might be AABDE and BCCDE. Sum the numbers in each set, and take the absolute value of the difference. Do this 10,000 times and sort in ascending order. The 9,750th observation when added and subtracted from the mean will be the two-tailed 95% confidence interval.
   c. This process was done at least 30 times for each set size. We stopped when the resolution in the graphics used in this manuscript was insufficient to show the change caused by the addition of more data.
   d. Set sizes of two and three have so few possible outcomes that we used the maximum possible difference, and processed all 1,000 of the repeats from step 3.
6. We expected that increasing set size would decrease the confidence interval. Using linear regression, we could model the rate of decrease and extrapolate to a set size of 1.
7. Contrast the outcome of step 4 with that in step 6.

Additional Information. These methods used discrete mathematics and are therefore discontinuous functions. One can collect one or two individuals but not 1.5 individuals. A sample of ten individuals with pool size three would have three pools and the remaining individual could be discarded. One might consider two pools of three and four pools of one individual, but this is not equivalent to having a pool size of 1.67. There is considerable value to using pools of different size, but this adds another level of complexity that we prefer to avoid at this time.

Because the number of individuals and pool size are integer, the outcomes from the CDC method or using equation 3 are discrete. For any number of individuals (N) divided into k groups, there will be prevalence estimates that are numerically impossible to achieve. All figures plot discontinuous outcomes regardless of the method of presentation.

In producing the figures, set size was not predetermined. Initial set size was between 50 and 100. However, it was difficult to see a pattern in the outcomes based on such small set sizes. Doubling set size made the figures clearer, but only relative to the previous graphs. In a figure such as Fig. 5, $S = 400$ because the resulting pattern was clear. There was a peak for each value of p and that peak changed as p interacts with k. Irregularities such as those between $k = 55$ and 65 for $p = 0.15$ were unimportant. In a figure such as Fig. 4, S needed to be larger to clearly show where peak variability occurs. As S increases, the curves all become more regular, but it takes more time to generate and process the data.

Using these methods, we examined the limits to using pooled samples especially at critical limits to the method such as the difference in estimated prevalence when all but one pool tests positive versus when all pools test positive. We then suggest some additional guidelines when trying to choose an optimal pool size. We finish by contrasting the confidence interval estimated using an asymptotic method as implemented in the program published by the CDC against an equivalent interval estimated using a randomization approach. The randomization approach coupled with linear regression avoids the bias inherent in the standard equations estimating the variance in prevalence calculated using pooled samples. We used multiple sets for the randomization procedure, but the goal was to estimate the confidence interval expected with only one sample. A regression analysis was used to estimate this case and give a 95% confidence interval.

Results and Discussion

100% Positive. Given any sample from an infected population, there is always a nonzero probability that all the pools will be infected. We calculated this probability for fixed number of pools (n) with pool size based on the equations by the three sets of studies (referred to by author name[s]): Chaing and Reeves (1962), Thompson (1962), and Burrows (1987) (Table 1). These equations estimate k as a real number, but in practice k is integer. Converting from real to integer causes a sawtoothed pattern in the estimated probability of having all pools infected (Fig. 1). For any prevalence, the optimal k by Chaing and Reeves (1962) gives the lowest probability that all pools are infected, and the order is always Chaing and Reeves $\ll$ Burrows $<$ Thompson (Fig. 1A). The probability of getting all pools positive decreases with an increasing number of pools (Fig. 1B and 1C). However, even for the more conservative Chaing and Reeves method there may be a fair chance of having all pools positive with a small number of pools (Fig. 1C). From these observations, we recommended at least 20 pools for each sample. Under these conditions, the equation for optimal pool size proposed by Chaing and Reeves (1962) will keep the risk of all pools testing positive at $<1\%$.

Enumerating Outcomes. If processed as individuals, a sample with $N - 1$ infected individuals has a prevalence of $1 - (1/N)$. Each additional uninfected individual will decrease the estimated prevalence by $1/N$. However, the relationship between number infected and estimated prevalence was not linear in pooled samples ($k > 1$), and the relationship departed from linearity more quickly as pool size increased (Fig. 2). With $k = 1$, there was a proportional change in prevalence for each additional tested individual. However, in pooled samples the addition of another infected pool does not increase prevalence by a fixed proportion. Each additional infected pool increases the estimated prevalence a little more than what it was increased by the addition of the previous infected pool (Fig. 2). Thus, the probability of getting $(\hat{p}) < p$ is greater than getting $(\hat{p}) > p$, and the difference in probability between getting $(\hat{p}) < p$ versus $(\hat{p}) > p$ increases with increasing p. Therefore, there will be a

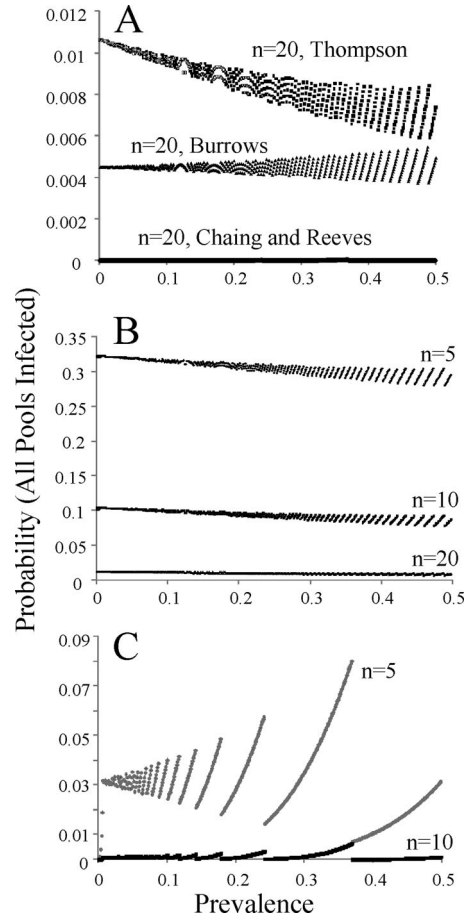

Fig. 1. Probability of all pools being infected given prevalence ranges from 0 to 0.5 in increments of 0.004. (A) Pool size (k) changes based on the optimal k as calculated by the three different authors with the equations given in Table 1. (B) Optimal pool size based on equation by Thompson for 5, 10, and 20 pools. (C) Optimal pool size based on the equation by Chaing and Reeves (see Table 1) changes for 5 or 10 pools. The sawtoothed pattern is caused by converting the estimated k from the equation into an integer value.

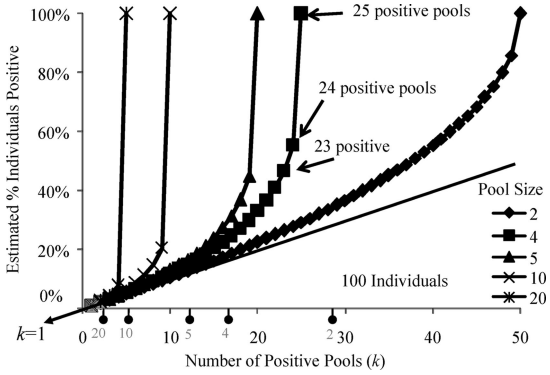

Fig. 2. Estimated prevalence for 100 individuals given pool sizes (k) 2, 4, 5, 10, and 20. For $N = 100$, $k = 4$, there are 25 pools. The graph plots prevalence if 1, 2, 3, . . . 25 of the pools are positive. For each k, the graph plots all possible outcomes given $N = 100$. Pushpins on the x-axis mark optimal p for each pool size as calculated using Chaing and Reeves (1962) equation solved for p. The diagonal line plots the outcome if one does not use pooling.

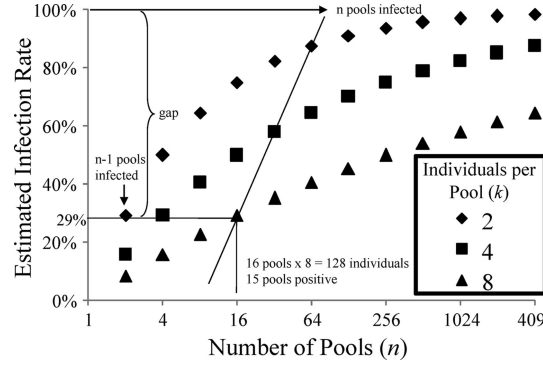

Fig. 3. The gap in prevalence estimates between n and $n - 1$ infected pools for different pool sizes (k) and different number of pools (n). If n pools are infected, the estimated prevalence will be 100%. Diagonal line connects points with equal number of individuals (N).

greater chance that a sample underestimates the true prevalence rather than over estimates it. This causes repeated prevalence estimates to be skewed and also results in a small group of outliers that are separated from the bulk of the estimates. Finally, there is a relatively large gap in estimated prevalence between all pools infected and all but one pool infected, and the size of this gap increases as the pool size increases (Fig. 2). Thus, if one tries to estimate a true prevalence of 60% by using 100 individuals and a pool size of two, it is likely that the estimate will be close. However, with $k = 5$ the methodology will result in either $(\hat{p}) = 100\%$ or $(\hat{p}) = 0\%$. This is a degradation in the quality of the prevalence estimate associated with the pooling methodology.

How do sample size and pool size interact in determining the prevalence estimate given that $n - 1$ pools are infected? Plotting three pool sizes using up to 4,096 individuals we noted that increasing the number of pools reduces the gap in estimated prevalence between $n - 1$ and n pools infected (Fig. 3). However, decreasing pool size reduces the size of this gap much more than increasing sample size (increasing the number of pools). Thus, if given a choice between collecting more samples at fixed pool size versus using the existing sample and reducing k, it is clear that reducing pool size has a greater effect on the maximum possible estimated prevalence.

Mapping π to p. This section is mostly about Fig. 4, which was created using a set size (S) of 600 repetitions of each different pool size ($k = 5, 10, 20$). Each repetition used 6,000 individuals, so it took 10.8 million individuals to generate the entire figure. The standard deviation is the variability within each set, and not the variability in N. Every increase in pool size (k) caused a shift to the left in the curves in these graphs. Thus, the curve for average p for $k = 5$ is a fairly straight line. At a pool size of 10, there was a marked increase in variability in estimated prevalence at higher prevalence levels, and at $k = 20$ this variability had a peak and subsequent decline. Similar plots with even larger pool sizes showed a shift left in all the graphs along with a steeper transition from 0 to 100%.

In situations without pooling ($k = 1$), variability in the percentage of infected individuals equals the variability in the percentage of infected pools. Pooling decouples this relationship as shown in Fig. 4, where the curve for the standard deviation of $\hat{p}$ (left column) differs from that of $\hat{\pi}$ (right column). One way to think of pooling is as a dose–response relationship where pool size is dose (e.g., "average" and "infected pools/n" [Fig. 4, $k = 20$]). As with any dose–response type relationship, variability is greatest at the 50% response level. As expected, the peak variability in the number of infected pools occurs when half the pools are infected (Fig. 4, π), and the maximum variability in the estimated prevalence rate occurs when half the time all the pools are positive (Fig. 4, $k = 20$).

Increasing the pool size increases the sharpness of the peaks in estimated standard deviation. Peak height decreases with increasing prevalence, and the peak keeps shifting to the left (Fig. 5). The rapid increase in standard deviation starts when sometimes all the pools are infected. At this point, there is also a corresponding increase in rate of change in the estimated prevalence (Fig. 6). We note that this mark is well to the left of the point where the probability that all pools are infected has much influence on estimated prevalence. This gives a margin of safety—where the method will continue to work even if there is an unusual sample with many times the typical infection rate. In this study, this safety margin remained proportionately constant at ≈6.2 times the optimal estimated pool size based on the equation by Chaing and Reeves (1962).

The reason for limiting pool size to 100 is shown in Fig. 6. The shape of each plot in Fig. 6 is driven by the probability of getting a sample where all pools test positive. The probability of having all pools test positive is a function of pool size and prevalence. Thus, the figure can be used to visualize the outcome of

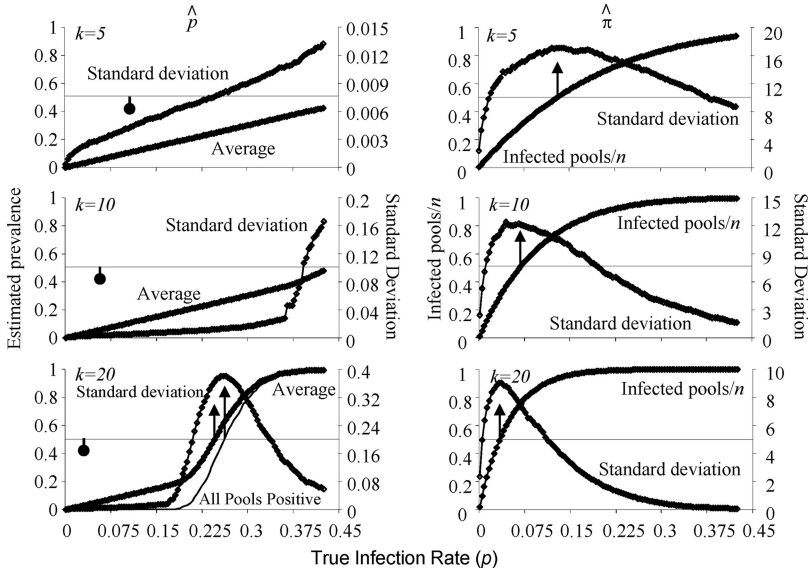

Fig. 4. Mapping the estimated prevalence (p), standard deviation in the estimated prevalence, fraction of infected pools (π), and the SD in the fraction of infected pools using equation 3 at true prevalence ranging from 0.001 to 0.421 in increments of 0.004 with pool sizes (k) of 5, 10, and 20. For p at $k = 20$, we include a line plot of the probability that all pools are infected. Pushpins mark optimal p for each pool size as calculated using the Chaing and Reeves (1962) equation solved for p. S = 600.

testing fixed pool sizes over a range of infection rates. As pool size increases, the prevalence range over which the method is useful decreases. Although the limit is a bit arbitrary, it provides a numerical value to a more general recommendation to keep pool size as small as possible if one lacks preliminary data sufficient to justify a more precise numerical answer.

Confidence Interval Estimates. How does pooling affect the 95% confidence interval estimate for p? The average confidence interval using the CDC approach provides the expected increase in confidence interval with an increase in prevalence given a fixed pool size (Fig. 7). A part of the increase in the interval width occurs because of the processes already discussed. Furthermore, the sample size decreased with increasing prevalence because samples were discarded when all pools were infected, and this occurs more frequently as prevalence increased.

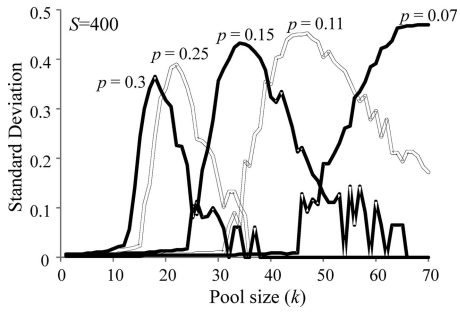

Fig. 5. Effect of pool size on the SD in estimated prevalence at several true prevalence values. There is a critical pool size beyond which the standard deviation increases rapidly.

In contrast to the gradual widening of the confidence interval in the CDC approach (Fig. 7), the confidence interval width using the randomization method increased slowly with increasing prevalence so long as true prevalence was below ≈0.17 (Fig. 8). It then increased rapidly up until half the time all pools were infected, at which point the interval declined until near zero because at that point it was almost certain that all pools were infected.

At set sizes of two or three, the confidence interval width was low relative to the set size of four, five and six (Fig. 8B). As set sizes of two or three, the number of sets required to observe rare events is so large that the rare events have less effect on the overall result. Within the range 0.001–0.17, the response was greatest

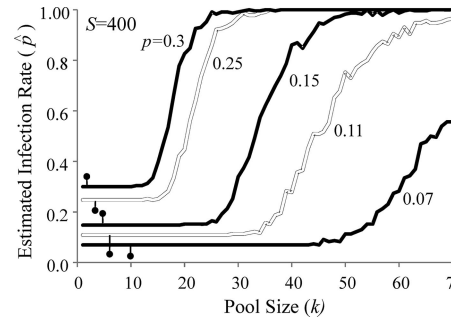

Fig. 6. Effect of pool size on estimated prevalence at different true prevalence rates. The values are the average of a set of 400. The pins on each line mark the point where half the pools are infected. The estimated prevalence starts inflating when one or more of the 400 sets has all pools infected. Pushpins mark optimal pool size as estimated using the equation by Chaing and Reeves (1962).

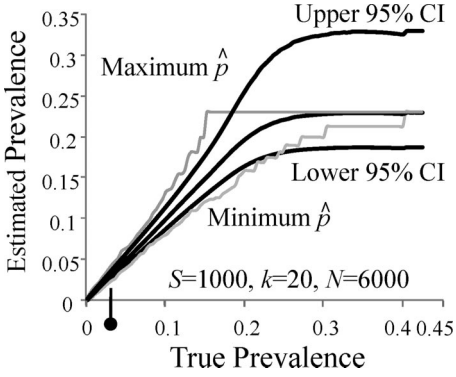

Fig. 7. Average and average 95% confidence interval for the CDC method run on 1,000 sets of 6,000 individuals tested using pool size of 20 individuals. The maximum and minimum estimates within those 1,000 sets also are shown. The push pin in the x-axis is where one half the pools are infected. Pushpin marks optimal prevalence for pool size of 20, as calculated using the Chaing and Reeves (1962) equation solved for p.

at set size of four, and declined thereafter (Fig. 8). Because the proposed limits for optimal pool size would not allow for $k = 20$ at $p > 0.17$, we used this limit in regression analyses. All regression equations for estimating the confidence interval width went from four to ten, and extrapolated back to a sample size of one. This approach always yielded a confidence interval that was smaller than that produced by the CDC methodology (Fig. 9). On average the regression approach had a confidence interval width 24% smaller than the CDC method. The upper 95%

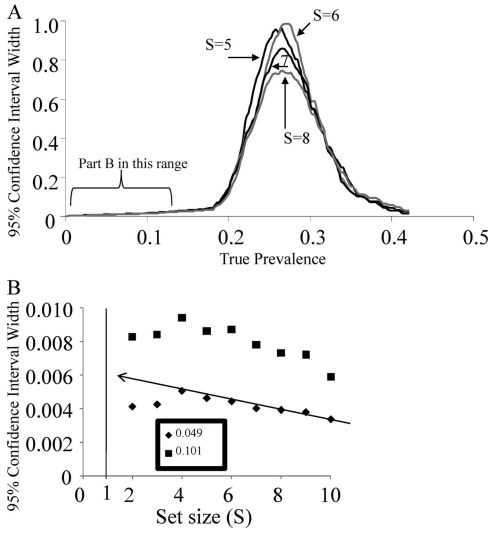

Fig. 8. (A) The 95% confidence interval width for prevalence estimates using equation 3 with set sizes (S) from two to eight. (B) Point estimates at different Set sizes for true prevalence rates of 0.049 and 0.101. The diagonal line is a regression line pointing to one set because that is the outcome we are estimating. The point estimates are the result of averaging 30 to 60 sets.

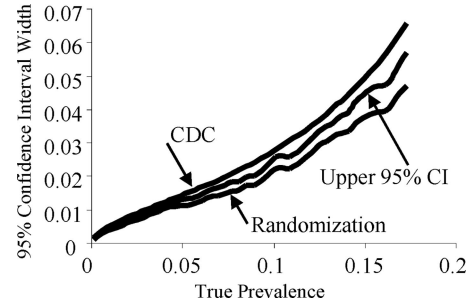

Fig. 9. Contrast in the width of the 95% confidence interval for the CDC method versus the randomization-regression method. The expected value for the randomization method is plotted as well as the upper 95% confidence interval for the randomization.

confidence interval estimate from the randomization approach was 12% smaller than that from the CDC method.

There is one final issue that is more about sample size rather than pooling methodology. In pooling we were concerned with the probability that all the pools were positive, and for any sample size where $Np > n$ there was a nonzero probability that all pools would be infected. At the other extreme, there is a nonzero probability that for any sample size and $p > 0$, that none of the individuals will be positive. So, at an infection rate of 0.1% there is a $(1 - 0.001)^{6,000} = 0.2\%$ probability that none of the 6,000 individuals in our virtual population would be infected. We can use this approach to select a required sample size. For example: We need to set a quarantine area to prevent further spread of a pathogen. A red area has high prevalence, yellow areas are where the pathogen in invading, and green areas are clean. However, we know that "clean" only means below detection levels. So we decide that infection rates below 0.5% are green areas. If we collect 600 individuals, there is still a 5% chance that we will miss an infection at the 0.005 prevalence level (Table 3). The task becomes more difficult if there are regulatory differences between green, yellow, and red. If yellow is between 0.005 and 0.01 prevalence, we need to decide how accurately we

Table 3. Probability of failing to find any infected individuals given specific sample sizes at a prevalence of 0.5%, and the upper 95% confidence limit to estimating a true prevalence of 0.005 without pooling given that no infected individuals were recovered

Sample size	Probability of zero	CDC upper 95% CI
100	0.605770	0.03699
200	0.366958	0.01885
300	0.222292	0.01264
400	0.134658	0.00951
500	0.081572	0.00762
600	0.049414	0.00636
700	0.029933	0.00546
800	0.018133	0.00478
900	0.010984	0.00425
1,000	0.006654	0.00383
2,000	0.000044	0.00192
3,000	0.0000003	0.00128

need to define the borders. At a sample size of 800, the upper 95% confidence interval no longer includes the 0.5% threshold for green status if no infected individuals are recovered. If eight infected individuals are recovered, the estimated infection rate for 3,000 individuals is 0.27%, with an upper 95% confidence level of 0.5%—just barely in the green zone given rounding error (actual value, 0.504%). If we now process these 3,000 samples using pooling methods, we will get wider confidence intervals and our green zone sample might be reclassified as yellow.

We are all concerned with what defines "large" sample sizes. In college Statistics 101, there were t-distribution tables where "infinite" was at a sample size of 121 or greater. A sample size of 10 or so would be a large number of replicates in field pesticide efficacy trials. However, in estimating prevalence the definition of "large sample size" is dependent on the lowest prevalence of interest. For a given application, even a sample size of 3,000 may still be too small. We had these cautions in mind when we chose a sample size of $N = 6{,}000$.

Pooling is a powerful tool for detection of rare events given limited time and funding. However, it is important to collect a sufficient sample size to make detection of the event probable. In processing the sample one must make a choice about pool size. Too small a pool size will expend additional resources, but that is preferable to having pools too large. Several authors have recommended different criteria for selecting pool size. We suggest that all are used because in combination they minimize the risk of having all pools test positive. The recommended limits are to have the number of pools (n) >20, pool size (k) <100, and fewer than half the number of pools testing positive. Within these limits, we recommend using the formula by Chaing and Reeves (1962) because it provides the most conservative optimal pool size. Information on bioassay sensitivity and biological variability could be used to modify these limits. Data from pooled samples can be analyzed using a program available through the CDC. The CDC program calculates confidence intervals that will facilitate reader interpretation of figures. The bias in estimating these confidence intervals means that the CDC program calculates an interval width at least 11% wider than necessary when there was a single pool size, and large sample size (N). However, the confidence intervals calculated by the CDC program should still be used until something better is developed.

Acknowledgments

This project was funded by grants from the Florida Citrus Production Research Advisory Council and the USDA, APHIS, PPQ, CPHST program.

References Cited

Andreadis, T. G., J. F. Anderson, P. M. Armstrong, and A. J. Main. 2008. Isolations of Jamestown Canyon Virus (Bunyaviridae: *Orthobunyavirus*) from field-collected mosquitoes (Diptera: Culicidae) in Connecticut, USA: a ten-year analysis 1997–2006. Vector-Borne Zoonotic Dis. 8: 175–188.

Biggerstaff, B. J. 2006. PooledInfRate, version 3.0: a Microsoft® Excel® add-in to compute prevalence estimates from pooled samples. CDC, Fort Collins, CO.

Biggerstaff, B. J. 2008. Confidence Intervals for the difference of two proportions estimated from pooled samples. J. Agric. Biol. Environ. Stat. 13: 478–496.

Bondell, H. D., A. Liu, and E. F. Schisterman. 2007. Statistical inference based on pooled data: a moment-based estimating equation approach. J. Appl. Stat. 34: 129–140.

Burrows, P. M. 1987. Improved estimation of pathogen transmission rates by group testing. Phytopathology 77: 363–365.

Carraro, L., F. Ferrini, G. Labonne, P. Ermacora, and N. Loi. 2004. Seasonal infectivity of *Cacopsylla pruni*, vector of European stone fruit yellows phytoplasma. Ann. Appl. Biol. 144: 191–195.

Chi, X.-F., X.-Y. Lou, M.C.K. Yang, and Q.-Y. Shu. 2009. An optimal DNA pooling strategy for progressive fine mapping. Genetica 135: 267–281.

Chiang, C. L., and W. C. Reeves. 1962. Statistical estimation of virus infection rates in mosquito vector populations. Am. J. Hyg. 75: 377–391.

Coutts, B. A., R. T. Prince, and R.A.C. Jones. 2009. Quantifying effects of seedborne inoculum on virus spread, yield losses, and seed infection in the *Pea seed-borne mosaic virus*–field pea pathosystem. Phytopathology 99: 1156–1167.

Crosslin, J. M., J. E. Munyaneza, A. Jensen, and B. P. Hamm. 2005. Association of beet leafhopper (Hemiptera: Cicadellidae) with a clover proliferation group phytoplasma in Columbia basin of Washington and Oregon. J. Econ. Entomol. 98: 279–283.

Dorfman, R. 1943. The detection of defective members in large populations. Ann. Math. Stat. 14: 436–440.

Ebert, T. A., W. S. Fargo, B. Cartwright, and F. R. Hall. 1998. Randomization tests: an example using morphological differences in *Aphis gossypii* (Hemiptera: Aphididae). Ann. Entomol. Soc. Am. 91: 761–770.

Farrington, C. P. 1992. Estimating prevalence by group-testing using generalized linear-models. Stat. Med. 11: 1591–1597.

Garcia-Chapa, M., J. Sabate, A. Lavina, and A. Batlle. 2005. Role of *Cacopsylla pyri* in the epidemiology of pear decline in Spain. Eur. J. Plant Pathol. 111: 9–17.

Geng, S., R. N. Campbell, M. Carter, and F. J. HIlls. 1983. Quality-control programs for seedborne pathogens. Plant Dis. Rep. 67: 236–242.

Gibbs, A. J., and J. C. Gower. 1960. The use of a multiple-transfer method in plant virus transmission studies—some statistical points arising in the analysis of results. Ann. Appl. Biol. 48: 75–83.

Hauck, W. W. 1991. Confidence intervals for seroprevalence determined from pooled data. Ann. Epidemiol. 1: 277–281.

Hepworth, G. 1996. Exact confidence intervals for proportions estimated by group testing. Biometrics 52: 1134–1146.

Hepworth, G. 1999. Estimation of proportions by group testing, pp. 232. Mathematics and Statistics, University of Melbourne, Melbourne, Australia.

Hepworth, G. 2004. Mid-P confidence intervals based on the likelihood ratio for proportions estimated by group testing. Aust. N Z J. Stat. 46: 391–405.

Hepworth, G. 2005. Confidence intervals for proportions estimated by group testing with groups of unequal size. J. Agric. Biol. Environ. Stat. 10: 478–497.

Hepworth, G., and R. Watson. 2009. Debiased estimation of proportions in group testing. J. R. Stat. Soc. Ser C Appl. Stat. 58: 105–121.

Hunt, R. E., L. R. Nault, and R. E. Gingery. 1988. Evidence for infectivity of maize chlorotic dwarf virus and for a helper component in its leafhopper transmission. Phytopathology 78: 499–504.

Kline, R. L., T. A. Brothers, R. Brookmeyer, S. Zeger, and T. C. Quinn. 1989. Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. J. Clin. Microbiol. 27: 1449–1452.

Loyer, M. W. 1983. Bad probability, good statistics, and group testing for binomial estimation. Am. Stat. 37: 57–59.

Munyaneza, J. E., J. M. Crosslin, and I.-M. Lee. 2007. Phytoplasma disease and insect vectors in potatoes of the Pacific northwest of the United States. Bull. Insectol. 60: 181–182.

Novack, L., B. Sarov, R. Goldman-Levi, V. Yahalom, J. Safi, H. Soliman, M. Orgel, A. Yaari, J. S. Pliskin, and E. Shinar. 2008. Impact of pooling on accuracy of hepatitis B virus surface antigen screening of blood donations. Trans. R. Soc. Trop. Med. Hyg. 102: 787–792.

Orshan, L., H. Bin, H. Schnur, A. Kaufman, A. Valinsky, L. Shulman, L. Weiss, E. Mendelson, and H. Pener. 2008. Mosquito vectors of West Nile fever in Israel. J. Med. Entomol. 45: 939–947.

Ross, S. 1984. A first course in probability. Macmillan Publishing Company, New York.

Rovira, A., J. P. Cano, and C. Munoz-Zanzi. 2008. Feasibility of pooled-sample testing for the detection of porcine reproductive and respiratory syndrome virus antibodies on serum samples by ELISA. Vet. Microbiol. 130: 60–68.

Sterne, T. E. 1934. Some remarks on confidence of fiducial limits. Biometrika 41: 275–278.

Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology 75: 882–889.

Thompson, K. H. 1962. Estimation of the proportion of vectors in a natural population of insects. Biometrics 18: 568–578.

Tu, X. M., E. Litvak, and M. Pagano. 1995. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. Biometrika 82: 287–297.

Vitek, C. J., S. L. Richards, C. N. Mores, J. F. Day, and C. C. Lord. 2008. Arbovirus transmission by *Culex nigripalpus* in Florida, 2005. J. Med. Entomol. 45: 483–493.

Wallace, I. S., A. Gregory, A. G. Murray, E. S. Munro, and R. S. Raynard. 2008. Distribution of infectious pancreatic necrosis virus (IPNV) in wild marine fish from Scottish waters with respect to clinically infected aquaculture sites producing Atlantic salmon, *Salmo salar* L. J. Fish Dis. 31: 177–186.

Walter, S. D., S. W. Hildreth, and B.J.D. Beaty. 1980. Estimation of infection rates in populations of organisms using pools of variable size. Am. J. Epidemiol. 112: 124–128.

Yameogo, L., L. Toe, J.-M. Hougard, B. A. Boatin, and T. R. Unnasch. 1999. Pool screen polymerase chain reaction for estimating the prevalence of *Onchocerca volvulus* infection in *Simulium damnosum sensu lato:* results of a field trial in an area subject to successful vector control. Am. J. Trop. Med. Hyg. 60: 124–128.

Received 30 October 2009; accepted 29 July 2010.

Appendix

Basic Probability

1. Probabilities range from 0 to 1, where 0 indicates an event will never happen and one indicates the event will always happen.
2. The sum of the probabilities of all possible outcomes must total exactly 1.
3. The probability that multiple independent events will occur is the product of the individual probabilities.

A coin has two sides, heads and tails. Toss a fair coin into the air and half the time it will land heads, so the probability of getting heads is 0.5 for one toss. There are only two outcomes, and the probability of getting heads plus the probability of not getting heads is 0.5 + 0.5 = 1. The probability of getting three heads in a row is the product of the individual probabilities, or $0.5^3 = 0.125$. We check this answer by listing all possibilities. There are two possible outcomes for each toss of the coin. So there are two outcomes for the first toss, then two outcomes for the second toss, and two more for the third. Thus, there are $2 \times 2 \times 2 = 8$ possible outcomes with three tosses. With h = heads, and t = tails, these are hhh, hht, hth, htt, thh, tht, tth, or ttt. The probability of hhh plus the probability of hht, plus . . . plus the probability of ttt must total 1, so the probability of each event must be $1/8 = 0.125 = 0.5^3$.

What is the probability of three heads in five tosses of the coin? There are $2^5 = 32$ possible outcomes (permutations), but some of these permutations are identical to others, e.g., hhhtt and hthth. We need to figure out how many unique combinations there are in these 32 permutations. The simple answer is (probability of heads)^3 $\times$ (probability of not heads)^2 = 0.03125. There are six possible combinations: hhhhh, hhhht, hhhtt, hhttt, htttt, and ttttt. Therefore, the total probability will be $6 \times 0.03125 = 0.19$. So, the simple answer is wrong because the total must equal 1. We will start with a new problem and come back to this one at the end.

The letters abc can be written abc, bac, bca, cba, cab, acb—there are six ways. Alternatively, if there are three letters to be placed into three spaces, one could place the first letter into any space. One would then put the second letter into any of the remaining two spaces. 3x2x1 = 3! = 6. In general, for A objects placed into A spaces there are A! ways to arrange them: or Ax(A-1)x(A-2)x(A-3)x . . . x(A-(A-1)). But sometimes some of the objects are identical. For example, how many ways can one arrange the letters in the

word $B_1E_1D_1D_2E_2D_3$? If all the letters are in some way unique there would be $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 6! = 720$ permutations. However, if one D looks like another we need to reduce the number of permutations by the number that look identical. So, we would get $\frac{6!}{3!\,2!}$ because there are three Ds and two Es. For any set of letters the equations would be $\frac{L!}{L_1!L_2!\ldots L_r!}$. Now, consider a case where one wants three letters from the list A B C D E. There are five choices for the first letter, four for the second, and three for the third. However, it does not matter if one gets ABC or ACB or any of the other similar combinations. Therefore, there will be $\frac{5x4x3}{3x2x1}$ possible groups. In general, with n items to place into r groups there will be $\frac{n(n-1)(n-2)\ldots.(n-r+1)}{r!}$ outcomes. A shorthand notation for this is $\binom{n}{r}$ or n choose r. This is sometimes phrased as the number of possible combinations of n objects taken r at a time.

To answer the three heads in five tosses, we noted that the simple approach was wrong. However, we can now include the number of ways of arranging three heads in five tosses of the coin. There is one way of getting five heads and one way of getting no heads. The probability of each event is 3%. There are five ways of getting four heads and five ways of getting only one head. The probability of each event is 16%. There are 10 ways of getting three heads or two heads, and the probability of each is 31%. Summing these probabilities gives $0.03 + 0.03 + 0.16 + 0.16 + 0.31 + 0.31 = 1$. One last problem: what is the probability of at least one head? One could work out all the probabilities, but there is only one outcome with no heads. The probability this will occur is 0.03, so the answer must be $1 - 0.03 = 0.97$. In our pooling problem, we can work with either the number of infected pools or with the number of healthy pools. A pooled sample will not be infected only if all individuals within that pool are not infected.