Chapter 2
Classifying Individual Samples into One of Two Categories

2.1 Introduction

Testing of groups for subsequent identification of group members possessing a trait was initiated by Dorfman (1943) to identify US servicemen infected with syphilis. Other reported applications of composite sampling include screening for pollutants (Schaeffer et al., 1982; Rajagopal and Williams, 1989), testing for leaking containers (Sobel and Groll, 1959; Thomas et al., 1973), identifying faulty components in a flow test (Hwang, 1984), identifying active users in a communications system (Hayes, 1978; Berger et al., 1984; Wolf, 1985; Garg and Mohan, 1987), recognizing the pattern of a binary code, and screening experimental factors affecting yield (Hwang, 1984). If the trait is relatively rare, then initially testing composite samples and subsequently only those individual samples that belong to a composite that has tested positive can greatly reduce the required number of tests (see, for instance, Feller, 1968; Garner et al., 1986).

Group testing was originally developed for a binary response variable, where the trait is either present or absent. However, the method can also be used for a (non-negative) continuous response variable where the trait is defined as the exceedance of some specified criterion level c. This latter case requires separate treatment since compositing results in an averaging of the individual values and, consequently, a measurement on the composite can fall below c even while some of the individual values exceed c. By contrast, in the binary response case, absence of the trait in the composite implies absence for all the individual samples. In either case, however, considerable savings can be realized by compositing if the trait is relatively rare.

Individual samples are first collected and prepared for laboratory procedures. Composite samples are then formed and measured. If a composite sample tests positive, then further testing of some or all constituent individual samples must be undertaken in order to classify the individual samples. Testing individual samples, either separately or in the form of composite subsamples of the original composite sample, is called *retesting*. Retesting may thus involve forming further composites as well as making additional laboratory measurements. When the cost of forming composites is negligible compared to the laboratory costs, then the effectiveness

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_2, © Springer Science+Business Media, LLC 2011

of compositing can be characterized by the *relative cost* which is defined as the number of measurements per individual sample classified. Exhaustive testing of all individual samples results in a relative cost of one measurement per sample. In order for composite sampling techniques to be cost-effective, their relative cost must be smaller than 1. Considerable cost savings can be realized when the trait is relatively rare. For example, when the prevalence p of the trait is 0.01, then a simple compositing strategy can result in a relative cost of 20%, or a savings of 80%, in the required number of measurements compared with exhaustive testing of all individual samples.

Consider the case of a continuous response variable. For a measurement such as the concentration of some pollutant in a composite sample, the composite sample measurement is the average of the individual sample values plus a measurement error, if present. For a measurement such as the total pollutant present in a composite sample, the composite sample measurement is the sum of the individual sample values plus a possible measurement error. These two cases are really the same since the concentration in a sample can be obtained from the total amount of the pollutant and the volume of the sample. We will use, as an illustrative formulation, the testing of water wells for the concentration of pollutants. Assume that the analytical measurement is accurate. The measurement on a composite sample is then the average of the individual sample values. Let c be the criterion value or action level. That is, the water from any well with a concentration exceeding c is not potable. Further, assume a detection limit of d. That is, if the concentration of the pollutant in any sample, either individual or composite, does not exceed d, then the laboratory procedure will return a measurement of 0 or an imprecise measurement. If water from one well with a concentration level of c and from $k-1$ other wells with no pollution is mixed to form a composite sample, then the composite sample value will be c/k. In order not to misclassify the polluted well as not polluted, it is necessary that $c/k \geq d$. This implies that the composite sample size k should satisfy $k \leq c/d$. In any application of composite sampling, it is accordingly necessary to place an upper limit on the composite sample size in order to avoid detection limit difficulties. If the criterion level is not at least twice as large as the detection level, then composite sample techniques cannot be used for classification. Conversely, if composite sample techniques are used, then implicitly there is a criterion level $c = kd$ below which classification is undependable. The detection limit for a composite sample of size k is also d, but due to dilution a polluted well with concentration between c and kd may escape detection and be misclassified as unpolluted.

For continuous measurements made without error, if any one individual sample value exceeds c, then the measurement on a composite sample of k individual samples will exceed c/k. Of course, it is possible that none of the individual sample values exceeds c and the composite sample measurement still exceeds c/k. If the composite sample measurement exceeds c/k, then further testing would be undertaken to identify individual sample values that exceed c, even though there may be none. On the other hand, if the composite measurement does not exceed c/k, then none of the individual samples needs be tested further, for none exceeds the value of c.

Section 2.2 discusses the presence/absence case, and Section 2.3 discusses the case of a continuous response variable.

2.2 Presence/Absence Measurements

During World War II, it was feared that some of the US servicemen were infected with syphilis-causing bacteria (*Treponema pallidum*). While the proportion of infected servicemen was not expected to be high, it was necessary to be certain that no infected individual remained undetected. The most commonly used laboratory procedures for the detection of syphilis are carried out on a sample of blood serum (serological tests for syphilis, or STS). The STS are based on detection of one of two substances that appear in blood serum soon after the onset of the disease: syphilis reagin and *treponemal* antibody. It was clear that a large proportion of laboratory procedures would return a negative response, but it was essential to make certain that the blood sample of every individual was subjected to laboratory procedures.

Dorfman (1943) came up with an apparently simple and yet cost-efficient procedure to identify the infected servicemen. The basic argument that Dorfman developed was as follows: Fix a positive integer k, and pool blood samples of k servicemen to be subjected to the STS as a single specimen. Assuming independence among servicemen, the probability that the syphilis bacteria are absent in a pooled sample from k servicemen is $g = 1 - p^k$, where p is the proportion of infected servicemen. Since p is small, g is large and negatively testing composites occur frequently, in which case a single test allows us to correctly classify k servicemen as uninfected.

The cost of collecting and preparing samples for laboratory procedures is constant since the number of individual samples is predetermined. The relative cost of classification can therefore be defined as the expected number of tests divided by the number of samples classified. When the cost of measurement is large and the cost of forming composite samples is relatively small, the relative cost can be reduced by the use of composite sample techniques. In Sections 2.2.1 through 2.2.6, it is assumed that the hierarchical processing of a single composite sample results in the classification of all individual samples eventually making up that composite. The composite sample size is the number of individual samples used to form the composite sample. If the composite sample size does not divide the total number of individual samples to be classified evenly, then near the end of the classification procedure, smaller composite sample sizes must be used. For a relatively small number of samples this "remainder" composite sample size must be taken into account. On the other hand, if the total number of samples is large, then the remainder effect can be ignored. In the limiting case, as the number of individual samples increases indefinitely, the relative cost is the same as the asymptotic relative cost.

The asymptotic relative cost is derived for the retesting procedures discussed in Sections 2.2.1 through 2.2.7. The formulation of Section 2.2.7 begins with the finite case and deduces the asymptotic case. In the finite case, an optimal partition of all

individual samples of various sizes is desired. Bush et al. (1984) and Gilstein (1985) discuss optimal partitions in the finite case for the exhaustive retesting procedure. This problem is not discussed in this monograph, as the emphasis here is in the asymptotic relative cost of classification.

2.2.1 Exhaustive Retesting

The original composite sampling procedure of exhaustive retesting is due to Dorfman (1943) and is often referred to as the Dorfman procedure in the literature (see, for instance, Johnson et al., 1991). Exhaustive retesting utilizes two stages of testing. The first stage consists of testing only the composite samples, while the second stage consists of testing all members of positive testing composite samples. Figure 2.1 shows the two stages of the exhaustive retesting procedure.

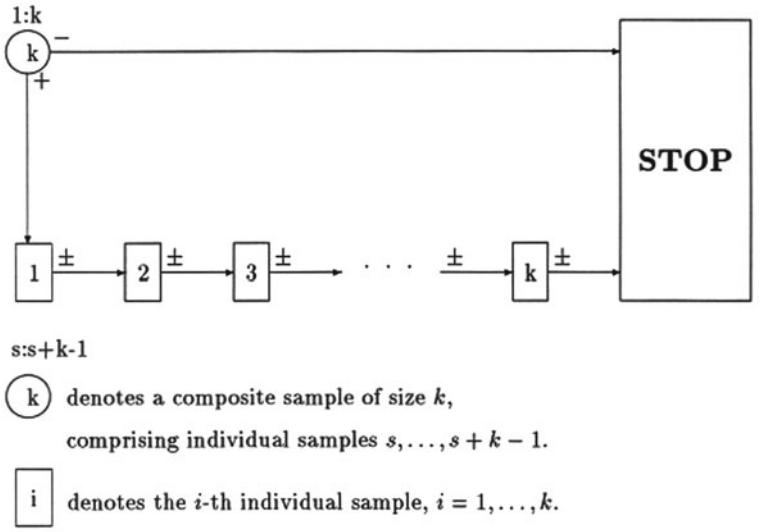

Fig. 2.1 Exhaustive retesting

Our analysis of the method employs a binomial model for the occurrence of the trait. Thus the individual samples are treated as independent trials with a constant probability, p, of possessing the trait. Let $I_1, I_2, \ldots, I_k$ be the k individual sample values, each taking a value of 0 (trait not present) or 1 (trait present). Now $I_1, \ldots, I_k$ are independent and identically distributed with $\Pr[I_i = 1] = p$ which is small for a rare trait. A composite sample formed from these individual samples will have a test result $I = 0$ or $I = 1$, with respective probabilities

$$\Pr[I = 0] = \Pr[I_1 = 0, I_2 = 0, \ldots, I_k = 0] = q^k,$$
$$\Pr[I = 1] = 1 - \Pr[I = 0] = 1 - q^k,$$

where $q = 1 - p$. If the composite sample tests negative, then all the k constituent individual samples are classified as not possessing the trait without any further testing. That is, the classification is achieved with only one test. On the other hand, if the composite sample tests positive, then every constituent individual sample is tested and classified separately. The total number of tests in this case is $k + 1$, with one test for the composite sample and k tests for the k individual samples. In this way, for classifying the k individual samples in a single composite sample, the number of tests, denoted by T_k, is either 1 or $k + 1$. The expected number of tests required is

$$E[T_k] = 1 \cdot q^k + (k+1) \cdot [1 - q^k] = (k+1) - kq^k. \tag{2.1}$$

The ratio of $E[T_k]$, the expected number of tests, to k, the number of individual samples classified, is defined to be the (asymptotic) relative cost, RC, which in this case is given by

$$\text{RC} = 1 + 1/k - q^k. \tag{2.2}$$

For given p, the optimal composite size k can be obtained by minimizing (2.2). Samuels (1978) showed that for all $p \leq 1 - (1/3)^{\frac{1}{3}} \cong 0.307$, the optimal composite sample size is the choice of $1 + [p^{-\frac{1}{2}}]$ or $2 + [p^{-\frac{1}{2}}]$, whichever minimizes the relative cost, where $[x]$ represents the integer part of x (see Table 2.1). The expected number of tests per individual sample classified, when an optimal composite sample size is used, is shown in Fig. 2.2 for selected values of p. Note that p must be sufficiently small to gain most of the benefits of composite sampling (see Table 2.1). For instance, for exhaustive retesting with the optimal k to be twice as efficient as the conventional method of testing individual samples, p must be less than 0.07.

The performance of compositing can also be assessed by the relative savings defined as $\text{RS} = 1 - \text{RC}$. The relative savings is often expressed as a percentage and represents the number of tests per classified item that can be saved, on average, by using compositing instead of individually testing each item. For Dorfman's method, the relative savings becomes

$$\text{RS} = (1 - p)^k - \frac{1}{k}.$$

Notice that the relative savings can be negative indicating that conventional testing would outperform the compositing method under consideration.

Identification of an optimal composite sample size k can yield performance benefits but does require accurate prior information for the value of p. If this prior information is inaccurate then the relative savings can be substantially suboptimal and may even be negative. For this reason it is useful to have some rules of thumb for choosing k when knowledge of p is limited. Ideally the rule would produce near-optimal savings with only a small risk of negative savings. One such rule is called the "1/4/12 rule" for Dorfman's method and divides the possible values of p into

Table 2.1 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for exhaustive retesting

p	k_{opt}	RC[a]	p	k_{opt}	RC
(0.30663, 1.00000]	1	1.000000	(0.00049, 0.00051]	45	(0.044033, 0.044916]
(0.12394, 0.30663]	3	(0.660973, 0.999987]	(0.00047, 0.00049]	46	(0.043129, 0.044033]
(0.06558, 0.12394]	4	(0.487622, 0.660973]	(0.00045, 0.00047]	47	(0.042207, 0.043129]
(0.04112, 0.06558]	5	(0.389372, 0.487622]	(0.00043, 0.00045]	48	(0.041262, 0.042207]
(0.02828, 0.04112]	6	(0.324792, 0.389372]	(0.00041, 0.00043]	49	(0.040296, 0.041262]
(0.02066, 0.02828]	7	(0.278810, 0.324792]	(0.00040, 0.00041]	50	(0.039806, 0.040296]
(0.01577, 0.02066]	8	(0.244410, 0.278810]	(0.00038, 0.00040]	51	(0.038799, 0.039806]
(0.01243, 0.01577]	9	(0.217573, 0.244410]	(0.00036, 0.00038]	52	(0.037771, 0.038799]
(0.01005, 0.01243]	10	(0.196068, 0.217573]	(0.00035, 0.00036]	53	(0.037244, 0.037771]
(0.00830, 0.01005]	11	(0.178510, 0.196068]	(0.00034, 0.00035]	54	(0.036710, 0.037244]
(0.00697, 0.00830]	12	(0.163839, 0.178510]	(0.00033, 0.00034]	55	(0.036169, 0.036710]
(0.00593, 0.00697]	13	(0.151323, 0.163839]	(0.00031, 0.00033]	56	(0.035062, 0.036169]
(0.00511, 0.00593]	14	(0.140635, 0.151323]	(0.00030, 0.00031]	57	(0.034493, 0.035062]
(0.00445, 0.00511]	15	(0.131373, 0.140635]	(0.00029, 0.00030]	58	(0.033915, 0.034493]
(0.00391, 0.00445]	16	(0.123255, 0.131373]	(0.00028, 0.00029]	59	(0.033330, 0.033915]
(0.00346, 0.00391]	17	(0.116037, 0.123255]	(0.00027, 0.00028]	60	(0.032731, 0.033330]
(0.00309, 0.00346]	18	(0.109738, 0.116037]	(0.00026, 0.00027]	61	(0.032121, 0.032731]
(0.00277, 0.00309]	19	(0.103966, 0.109738]	(0.00025, 0.00026]	63	(0.031498, 0.032121]
(0.00250, 0.00277]	20	(0.098827, 0.103966]	(0.00024, 0.00025]	64	(0.030867, 0.031498]
(0.00227, 0.00250]	21	(0.094222, 0.098827]	(0.00023, 0.00024]	65	(0.030219, 0.030867]
(0.00206, 0.00227]	22	(0.089800, 0.094222]	(0.00022, 0.00023]	66	(0.029556, 0.030219]
(0.00189, 0.00206]	23	(0.086054, 0.089800]	(0.00021, 0.00022]	68	(0.028879, 0.029556]
(0.00173, 0.00189]	24	(0.082364, 0.086054]	(0.00020, 0.00021]	70	(0.028184, 0.028879]
(0.00160, 0.00173]	25	(0.079241, 0.082364]	(0.00019, 0.00020]	71	(0.027476, 0.028184]
(0.00148, 0.00160]	26	(0.076237, 0.079241]	(0.00018, 0.00019]	73	(0.026744, 0.027476]
(0.00137, 0.00148]	27	(0.073373, 0.076237]	(0.00017, 0.00018]	75	(0.025992, 0.026744]
(0.00127, 0.00137]	28	(0.070665, 0.073373]	(0.00016, 0.00017]	77	(0.025218, 0.025992]
(0.00118, 0.00127]	29	(0.068134, 0.070665]	(0.00015, 0.00016]	80	(0.024423, 0.025218]
(0.00111, 0.00118]	30	(0.066102, 0.068134]	(0.00014, 0.00015]	82	(0.023596, 0.024423]
(0.00104, 0.00111]	31	(0.063999, 0.066102]	(0.00013, 0.00014]	85	(0.022739, 0.023596]
(0.00097, 0.00104]	32	(0.061821, 0.063999]	(0.00012, 0.00013]	88	(0.021848, 0.022739]
(0.00091, 0.00097]	33	(0.059891, 0.061821]	(0.00011, 0.00012]	92	(0.020919, 0.021848]
(0.00086, 0.00091]	34	(0.058235, 0.059891]	(0.00010, 0.00011]	96	(0.019952, 0.020919]
(0.00081, 0.00086]	35	(0.056529, 0.058235]	(0.00009, 0.00010]	100	(0.018929, 0.019952]
(0.00077, 0.00081]	36	(0.055125, 0.056529]	(0.00008, 0.00009]	106	(0.017848, 0.018929]
(0.00073, 0.00077]	37	(0.053684, 0.055125]	(0.00007, 0.00008]	112	(0.016696, 0.017848]
(0.00069, 0.00073]	38	(0.052201, 0.053684]	(0.00006, 0.00007]	120	(0.015465, 0.016696]
(0.00065, 0.00069]	39	(0.050673, 0.052201]	(0.00005, 0.00006]	130	(0.014118, 0.015465]
(0.00062, 0.00065]	40	(0.049498, 0.050673]	(0.00004, 0.00005]	142	(0.012628, 0.014118]
(0.00059, 0.00062]	41	(0.048293, 0.049498]	(0.00003, 0.00004]	158	(0.010936, 0.012628]
(0.00056, 0.00059]	42	(0.047054, 0.048293]	(0.00002, 0.00003]	183	(0.008940, 0.010936]
(0.00054, 0.00056]	43	(0.046214, 0.047054]	(0.00001, 0.00002]	224	(0.006500, 0.008940]
(0.00051, 0.00054]	44	(0.044916, 0.046214]			

[a]RC: Relative cost

three categories, corresponding to "frequent," "infrequent," and "rare" occurrence of the trait in question. A different value of k is used for each case as follows:

Frequent. If $0.29 < p < 1$, use $k = 1$ (conventional testing).
Infrequent. If $0.01 < p < 0.29$, use $k = 4$.
Rare. If $0 < p < 0.01$, use $k = 12$.

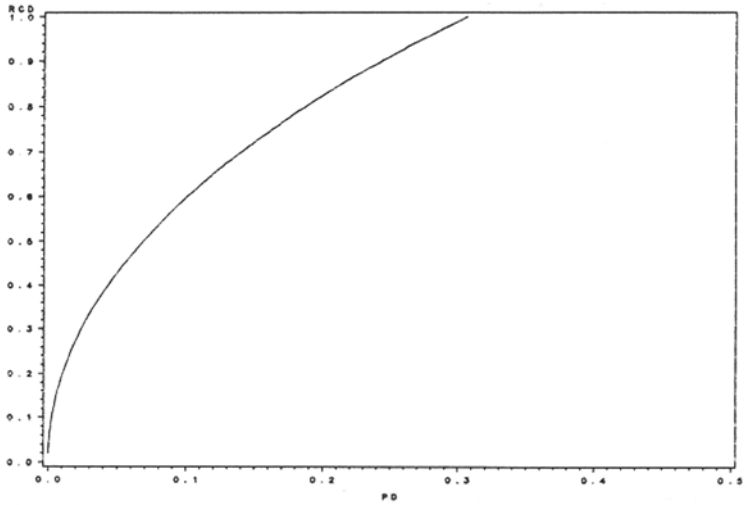

Fig. 2.2 Relative cost for exhaustive retesting using the optimal composite sample size

The following tabulation compares the relative savings achieved by the 1/4/12 rule with the optimal relative savings for several values of p. Relative savings is expressed as percentages.

p	RS (1/4/12)	RS (optimal)
0.275	3	5
0.25	7	9
0.20	16	18
0.15	27	28
0.10	41	41
0.05	56	57
0.025	65	69
0.01+	71	80
0.01−	80	80
0.005	86	86
0.001	90	94

The tabulation supposes that p is correctly classified into the frequent, infrequent, or rare categories. Classification error can occur when p is near a category boundary. In this case, the achieved savings can be somewhat more or somewhat less than indicated in the tabulation.

2.2.2 Sequential Retesting

Sterrett (1957) suggested a modification to exhaustive retesting, and hence this modified procedure is often referred to as the Sterrett procedure. The modification is

motivated by the following observations. When a composite tests positive, then the test result tells us that at least one of the k constituent individual samples possesses the trait. The Dorfman procedure determines which samples have the trait by individually testing each item. Here, Sterrett notes that as soon as an individual sample is found with the trait, then there is no information on whether any of the remaining (untested and therefore unclassified) samples from that composite has the trait. Moreover, the prevalence of the trait among these unclassified samples is still p. It is then natural to argue that compositing the unclassified samples should be more economical than individual testing, even at this stage. This was the observation that led Sterrett (1957) to propose the following: When a composite tests positive, its constituent individual samples are tested sequentially until a positively testing sample is identified. At this stage, all the remaining individual samples (from among the k that formed the original composite) are used to form a new composite sample for testing. If this composite tests negative, then all its constituent individual samples are classified as not possessing the trait, and no more testing is necessary. On the other hand, if this composite tests positive, then the same procedure is repeated, beginning with sequential testing of constituent individual samples until an individual sample is identified as possessing the trait. This procedure continues until all the k individual samples comprising the original composite sample have been classified. Figure 2.3 displays the sequential retesting procedure.

Let T_k be the number of tests required to classify the k individual samples that constitute a composite sample. For small values of k, the expected number of tests can be calculated directly, giving, for instance,

$$E[T_1] = 1 \ \ (T_1 \equiv 1), \ \ E[T_2] = 3 - 2q^2, \ \text{and} \ \ E[T_3] = 5 - q - 2q^2 - q^3.$$

A recurrence formula can be found by conditioning on the number of retests J to find the first positively testing item. Thus, $J = 0$ if the composite sample tests negative; $J = 1$ if the first item retested tests positive, etc. Now

$$\begin{aligned} E[T_k] &= E[E\,(T_k|J)] = \sum_{j=0}^{k} E[T_k|J = j]P[J = j] \\ &= E[T_k|J = 0]q^k + \sum_{j=1}^{k} E[T_k|J = j]q^{j-1}p, \ \ k = 2, 3, \ldots. \end{aligned}$$

Given that the composite sample tests positive, that the first $j - 1$ individual samples test negative, and that the jth individual sample tests positive, the remaining $k - j$ individual sample values are independent Bernoulli random variables with parameter p. We therefore have

$$E[T_k|J = j] = j + 1 + E[T_{k-j}], \quad j > 0,$$

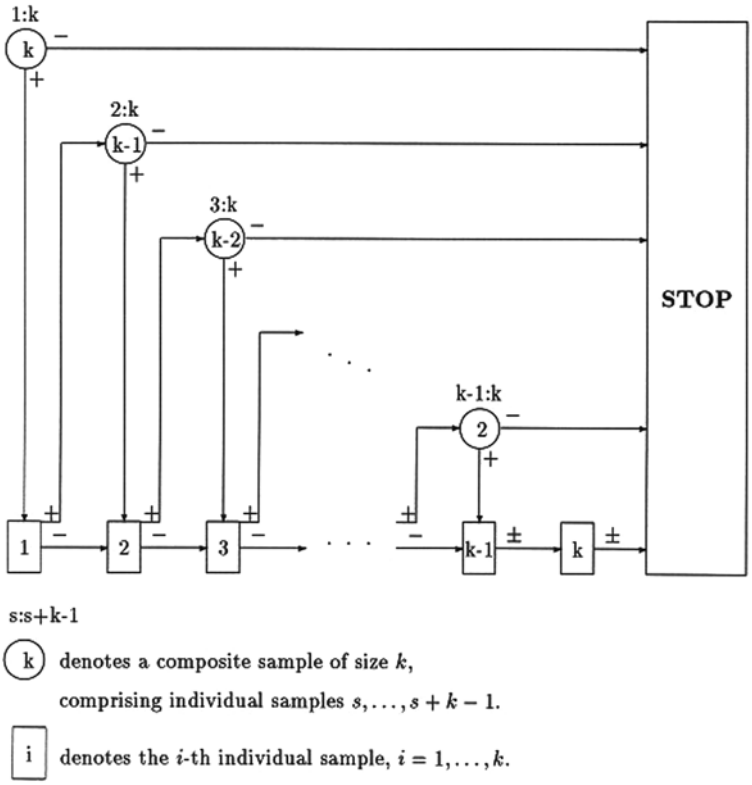

Fig. 2.3 Sequential retesting

where $T_0 = 0$, and hence

$$E[T_k] = q^k + \sum_{j=1}^{k} \left\{ j + 1 + E[T_{k-j}] \right\} q^{j-1} p.$$

So

$$E[T_k] = q^k + \sum_{j=1}^{k} (j+1) q^{j-1} p + \sum_{j=1}^{k} E[T_{k-j}] q^{j-1} p$$

$$= q^k + \sum_{j=0}^{k-1} (j+2) q^j p + \sum_{j=0}^{k-1} E[T_j] q^{k-j-1} p.$$

Also

$$q\,E[T_{k-1}] = q^k + \sum_{j=1}^{k-1} (j+1) q^j p + \sum_{j=0}^{k-2} E[T_j] q^{k-j-1} p, \quad k = 3, 4, \ldots.$$

Then

$$E[T_k] - qE[T_{k-1}] = \left[2p + \sum_{j=1}^{k-1} q^j p\right] + E[T_{k-1}]p.$$

Thus

$$E[T_k] - E[T_{k-1}] = 2p + \frac{qp\left(1 - q^{k-1}\right)}{1-q} = 2p + q - q^k.$$

Further

$$\begin{aligned} E[T_k] - E[T_2] &= \sum_{j=3}^{k} \left\{E[T_j] - E[T_{j-1}]\right\} \\ &= (k-2)(2p+q) - \sum_{j=3}^{k} q^j \\ &= (k-2)(2p+q) - \frac{q^3\left(1-q^{k-2}\right)}{1-q}, \quad k = 2, 3, \ldots. \end{aligned}$$

Substituting the value of $E[T_2]$ gives

$$\begin{aligned} E[T_k] &= 3 - 2q^2 + (k-2)(2p+q) + \frac{1-q^3}{p} - \frac{1-q^{k+1}}{p} \\ &= 2k - (k-3)q - q^2 - \frac{1-q^{k+1}}{p}, \quad k = 2, 3, \ldots. \end{aligned}$$

The (asymptotic) relative cost is

$$\mathrm{RC} = E[T_k]/k = 2 - q + \frac{1}{k}\left[3q - q^2 - \frac{1-q^{k+1}}{p}\right]. \tag{2.3}$$

The optimal composite sample size is tabulated in Table 2.2, and the relative cost using the optimal composite sample size is shown in Fig. 2.4 for selected values of p.

2.2.3 Binary Split Retesting

The exhaustive and sequential retesting procedures presented above have the following limitation to their cost-efficiency. The prevalence of individual samples that possess the trait is sufficiently small to justify the use of composite sampling. However,

Table 2.2 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for sequential retesting

p	k_{opt}	RC	p	k_{opt}	RC
(0.30437, 1.00000]	1	1.000000	(0.00064, 0.00066]	56	(0.036512, 0.037090]
(0.21253, 0.30437]	3	(0.827997, 0.999984]	(0.00061, 0.00064]	57	(0.035628, 0.036512]
(0.12774, 0.21253]	4	(0.622819, 0.827997]	(0.00059, 0.00061]	58	(0.035030, 0.035628]
(0.08283, 0.12774]	5	(0.487586, 0.622819]	(0.00057, 0.00059]	59	(0.034418, 0.035030]
(0.05759, 0.08283]	6	(0.397398, 0.487586]	(0.00056, 0.00057]	60	(0.034111, 0.034418]
(0.04224, 0.05759]	7	(0.334221, 0.397398]	(0.00054, 0.00056]	61	(0.033485, 0.034111]
(0.03226, 0.04224]	8	(0.287849, 0.334221]	(0.00052, 0.00054]	62	(0.032847, 0.033485]
(0.02543, 0.03226]	9	(0.252541, 0.287849]	(0.00050, 0.00052]	63	(0.032198, 0.032847]
(0.02055, 0.02543]	10	(0.224784, 0.252541]	(0.00049, 0.00050]	64	(0.031869, 0.032198]
(0.01695, 0.02055]	11	(0.202456, 0.224784]	(0.00047, 0.00049]	65	(0.031200, 0.031869]
(0.01422, 0.01695]	12	(0.184127, 0.202456]	(0.00045, 0.00047]	66	(0.030519, 0.031200]
(0.01210, 0.01422]	13	(0.168814, 0.184127]	(0.00043, 0.00045]	68	(0.029821, 0.030519]
(0.01042, 0.01210]	14	(0.155825, 0.168814]	(0.00042, 0.00043]	69	(0.029466, 0.029821]
(0.00906, 0.01042]	15	(0.144618, 0.155825]	(0.00040, 0.00042]	70	(0.028746, 0.029466]
(0.00796, 0.00906]	16	(0.134999, 0.144618]	(0.00038, 0.00040]	72	(0.028006, 0.028746]
(0.00704, 0.00796]	17	(0.126487, 0.134999]	(0.00036, 0.00038]	74	(0.027249, 0.028006]
(0.00627, 0.00704]	18	(0.118972, 0.126487]	(0.00035, 0.00036]	75	(0.026861, 0.027249]
(0.00562, 0.00627]	19	(0.112299, 0.118972]	(0.00034, 0.00035]	76	(0.026469, 0.026861]
(0.00507, 0.00562]	20	(0.106376, 0.112299]	(0.00033, 0.00034]	78	(0.026070, 0.026469]
(0.00460, 0.00507]	21	(0.101079, 0.106376]	(0.00032, 0.00033]	79	(0.025667, 0.026070]
(0.00419, 0.00460]	22	(0.096254, 0.101079]	(0.00031, 0.00032]	80	(0.025259, 0.025667]
(0.00383, 0.00419]	23	(0.091835, 0.096254]	(0.00030, 0.00031]	81	(0.024840, 0.025259]
(0.00351, 0.00383]	24	(0.087744, 0.091835]	(0.00029, 0.00030]	83	(0.024417, 0.024840]
(0.00323, 0.00351]	25	(0.084020, 0.087744]	(0.00028, 0.00029]	84	(0.023989, 0.024417]
(0.00299, 0.00323]	26	(0.080711, 0.084020]	(0.00027, 0.00028]	85	(0.023550, 0.023989]
(0.00277, 0.00299]	27	(0.077567, 0.080711]	(0.00026, 0.00027]	87	(0.023104, 0.023550]
(0.00258, 0.00277]	28	(0.074758, 0.077567]	(0.00025, 0.00026]	88	(0.022649, 0.023104]
(0.00240, 0.00258]	29	(0.072004, 0.074758]	(0.00024, 0.00025]	89	(0.022187, 0.022649]
(0.00224, 0.00240]	30	(0.069476, 0.072004]	(0.00023, 0.00024]	92	(0.021714, 0.022187]
(0.00210, 0.00224]	31	(0.067194, 0.069476]	(0.00022, 0.00023]	94	(0.021230, 0.021714]
(0.00197, 0.00210]	32	(0.065010, 0.067194]	(0.00021, 0.00022]	95	(0.020735, 0.021230]
(0.00185, 0.00197]	33	(0.062933, 0.065010]	(0.00020, 0.00021]	98	(0.020230, 0.020735]
(0.00174, 0.00185]	34	(0.060973, 0.062933]	(0.00019, 0.00020]	101	(0.019714, 0.020230]
(0.00164, 0.00174]	35	(0.059140, 0.060973]	(0.00018, 0.00019]	104	(0.019182, 0.019714]
(0.00155, 0.00164]	36	(0.057445, 0.059140]	(0.00017, 0.00018]	105	(0.018635, 0.019182]
(0.00147, 0.00155]	37	(0.055900, 0.057445]	(0.00016, 0.00017]	108	(0.018072, 0.018635]
(0.00139, 0.00147]	38	(0.054312, 0.055900]	(0.00015, 0.00016]	113	(0.017495, 0.018072]
(0.00132, 0.00139]	39	(0.052889, 0.054312]	(0.00014, 0.00015]	115	(0.016895, 0.017495]
(0.00125, 0.00132]	40	(0.051428, 0.052889]	(0.00013, 0.00014]	119	(0.016275, 0.016895]
(0.00119, 0.00125]	41	(0.050145, 0.051428]	(0.00012, 0.00013]	126	(0.015629, 0.016275]
(0.00114, 0.00119]	42	(0.049054, 0.050145]	(0.00011, 0.00012]	131	(0.014956, 0.015629]
(0.00108, 0.00114]	43	(0.047710, 0.049054]	(0.00010, 0.00011]	135	(0.014258, 0.014956]
(0.00104, 0.00108]	44	(0.046797, 0.047710]	(0.00009, 0.00010]	141	(0.013520, 0.014258]
(0.00099, 0.00104]	45	(0.045630, 0.046797]	(0.00008, 0.00009]	150	(0.012742, 0.013520]
(0.00094, 0.00099]	46	(0.044435, 0.045630]	(0.00007, 0.00008]	156	(0.011909, 0.012742]
(0.00091, 0.00094]	47	(0.043704, 0.044435]	(0.00006, 0.00007]	169	(0.011026, 0.011909]
(0.00087, 0.00091]	48	(0.042710, 0.043704]	(0.00005, 0.00006]	186	(0.010057, 0.011026]
(0.00083, 0.00087]	49	(0.041694, 0.042710]	(0.00004, 0.00005]	202	(0.008988, 0.010057]
(0.00080, 0.00083]	50	(0.040917, 0.041694]	(0.00003, 0.00004]	228	(0.007782, 0.008988]
(0.00077, 0.00080]	51	(0.040126, 0.040917]	(0.00002, 0.00003]	248	(0.006532, 0.007782]
(0.00074, 0.00077]	52	(0.039321, 0.040126]	(0.00001, 0.00002]	250	(0.005269, 0.006532]
(0.00071, 0.00074]	53	(0.038499, 0.039321]			
(0.00069, 0.00071]	54	(0.037941, 0.038499]			
(0.00066, 0.00069]	55	(0.037090, 0.037941]			

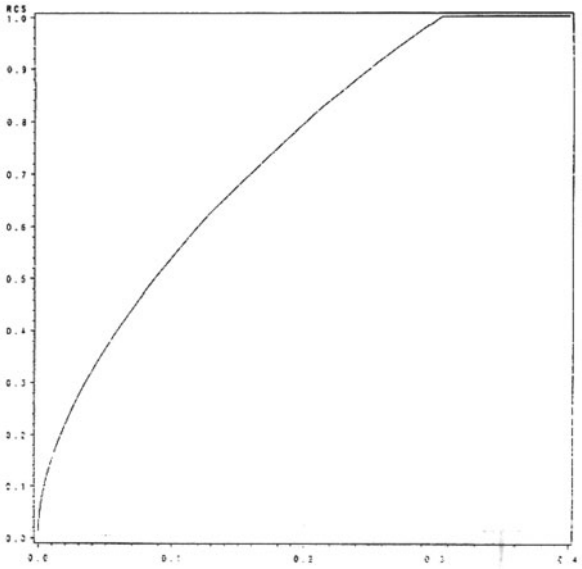

Fig. 2.4 Relative cost for sequential retesting using the optimal composite sample size

after a composite sample tests positive, both of these retesting procedures recommend exhaustive testing of the constituent individual samples (at least initially), which may not be very cost-effective. Also, note that the conditional prevalence of the trait among individual samples that comprise a composite that tests positive is $p^+ = \frac{p}{1-q^k}$, where p is the prevalence of the trait in the population and k is the composite sample size. It is clear that $p^+ > p$, and hence the optimal composite sample size corresponding to the prevalence p^+ will clearly not be larger than k, which is optimal corresponding to the prevalence p. It may therefore be more reasonable to form subcomposites of a positive testing composite rather than resort to exhaustive testing. With this modification and assuming a binomial model, Gill and Goltlieb (1974) proposed that positive testing composites be divided into two subcomposites of as equal sizes as possible, and positive testing subcomposites be recursively tested after dividing into two further subcomposites. Figure 2.5 describes the binary split retesting procedure of Gill and Gottlieb. Examples of the binary split retesting procedure for $k = 4$ and $k = 8$ are also presented in Figs. 2.6 and 2.7.

Let T_k be the number of tests required for classifying the k individual samples in a composite. For small values of k, the expectations can be calculated directly, giving

$$E[T_1] = 1, \quad E[T_2] = 3 - 2q^2, \text{ and } E[T_3] = 5 - 2q^2 - 2q^3.$$

When retesting is required, aliquots of the k individual samples are composited into two groups of sizes k_1 and $k_2 = k - k_1$, where $k_1 = k_2 = k/2$ if k is even and

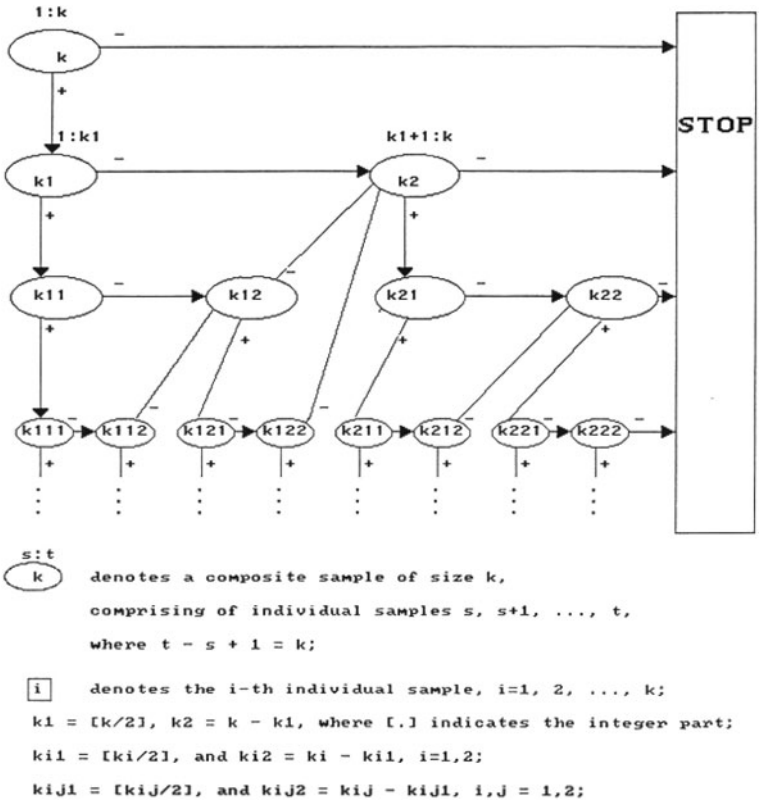

Fig. 2.5 Binary split retesting

where $k_1 = (k-1)/2$ and $k_2 = (k+1)/2$ if k is odd. Let "split" indicate that the composite of size k tests positive. Note that "split" means at least one of the k items has the trait. Now

$$\begin{aligned} E[T_k] &= E\left[T_k \mid \text{split}\right] + E\left[T_k \mid \text{not split}\right] \\ &= \left(1 + E\left[T_{k_1} + T_{k_2} \mid \text{split}\right]\right) \Pr\left[\text{split}\right] + 1 \cdot \Pr\left[\text{not split}\right] \\ &= E\left[T_{k_1} + T_{k_2} \mid \text{split}\right] \Pr\left[\text{split}\right] + 1. \end{aligned}$$

Now consider two separate composites of sizes k_1 and k_2, and let "split" indicate that at least one of the composites tests positive. Note that "split" means that at least one of the $k_1 + k_2 = k$ items tests positive. Then

$$\begin{aligned} E\left[T_{k_1} + T_{k_2}\right] = {} & E\left[T_{k_1} + t_{k_2} \mid \text{split}\right] \Pr\left[\text{split}\right] \\ & + E\left[T_{k_1} + T_{k_2} \mid \text{not split}\right] \Pr\left[\text{not split}\right]. \end{aligned}$$

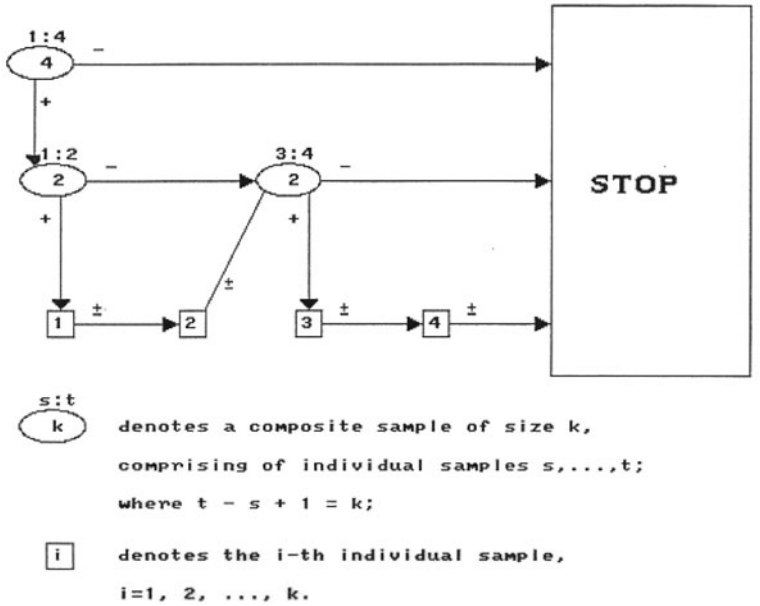

Fig. 2.6 An example of binary split retesting with $k = 4$

That is,

$$E\left[T_{k_1}\right] + E\left[T_{k_2}\right] = E\left[T_{k_1} + t_{k_2} \mid \text{split}\right] \Pr\left[\text{split}\right] + 2q^k.$$

Subtracting from the above expression for $E\,[T_k]$ now yields

$$E\,[T_k] = E\left[T_{k_1}\right] + E\left[T_{k_2}\right] + 1 - 2q^k. \tag{2.4}$$

This formula can be used to recursively generate $E\,[T_k]$ for any given k. For example, $E[T_2] = 2E[T_1] + 1 - 2q^2$ and $E[T_3] = E[T_1] + E[T_2] + 1 - 2q^3$. In this way, we obtain

$$\begin{aligned}
E[T_2] &= 3 - 2q^2, \\
E[T_3] &= 5 - 2q^2 - 2q^3, \\
E[T_4] &= 7 - 4q^2 - 2q^4, \\
E[T_5] &= 9 - 4q^2 - 2q^3 - 2q^5, \\
E[T_6] &= 11 - 4q^2 - 4q^3 - 2q^6, \\
E[T_7] &= 13 - 6q^2 - 2q^3 - 2q^4 - 2q^7, \\
E[T_8] &= 15 - 8q^2 - 4q^4 - 2q^8, \\
E[T_9] &= 17 - 8q^2 - 2q^3 - 2q^4 - 2q^5 - 2q^9,
\end{aligned}$$

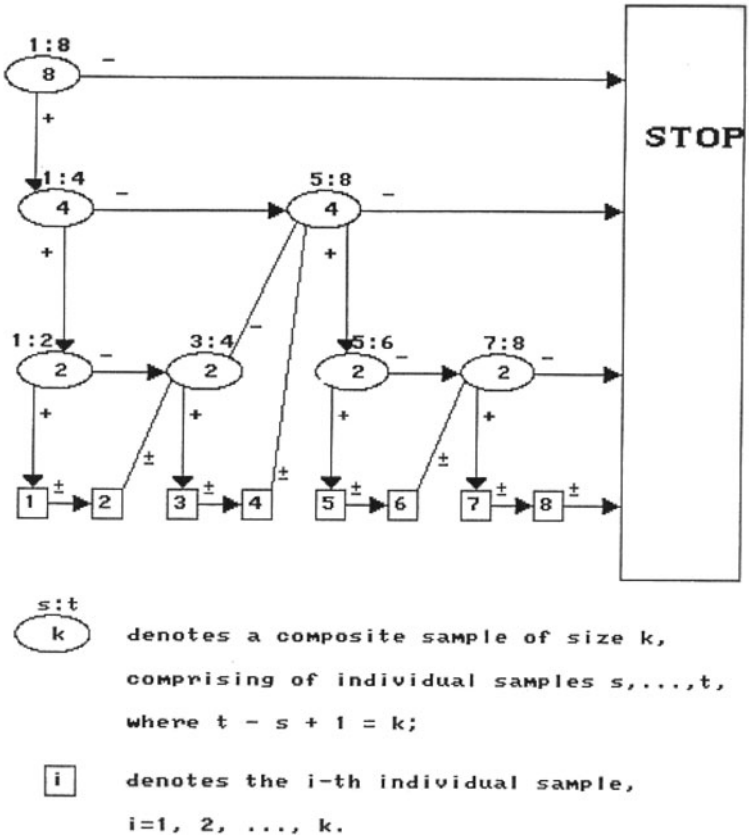

Fig. 2.7 An example of binary split retesting with $k = 8$

$$E[T_{10}] = 19 - 8q^2 - 4q^3 - 4q^5 - 2q^{10},$$
$$E[T_{11}] = 21 - 8q^2 - 6q^3 - 2q^5 - 2q^6 - 2q^{10},$$
$$E[T_{12}] = 23 - 8q^2 - 8q^3 - 4q^6 - 2q^{12}.$$

The asymptotic relative cost can be calculated by dividing the expected number of tests by the corresponding composite sample size. The optimal composite sample size is tabulated in Table 2.3, and the relative cost of classification with the binary split procedure using the optimal composite sample size is shown in Fig. 2.8 for selected values of p.

2.2.4 Curtailed Exhaustive Retesting

Many of the retesting strategies can be improved upon when the laboratory procedure is free of testing error. It is also necessary to have some freedom to schedule

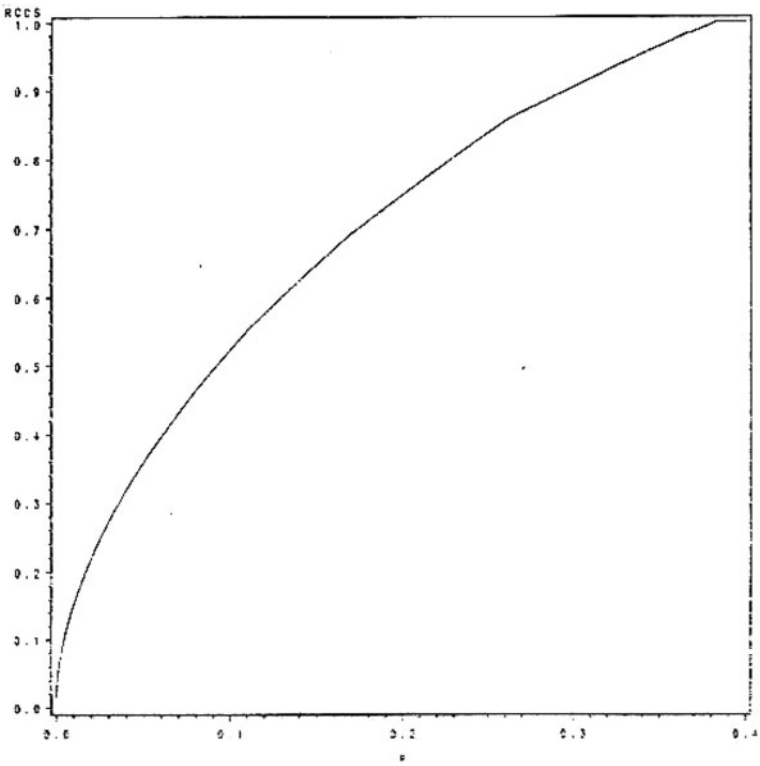

Fig. 2.8 Relative cost for binary split retesting using the optimal composite sample size

Table 2.3 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for binary split retesting

p	k_{opt}	RC
(0.29289, 1.00000]	1	1.00
(0.15910, 0.29289]	2	(0.792883, 0.999995]
(0.08299, 0.15910]	4	(0.555524, 0.792883]
(0.04239, 0.08299]	8	(0.360729, 0.555524]
(0.02142, 0.04239]	16	(0.222719, 0.360729]
(0.01077, 0.02142]	32	(0.132800, 0.222719]
(0.00540, 0.01077]	64	(0.077175, 0.132800]
(0.00267, 0.00540]	128	(0.043552, 0.077175]

the laboratory tests sequentially. In general, whenever a composite tests positive, the items comprising that composite must be subjected to retesting, either individually or in groups. Now suppose the items making up the composite can be partitioned into two subsets and we know, from our retesting, that none of the items in one of the subsets has the trait. Then, without testing, we know that at least one of the items in the second subset does have the trait. The avoidance of the test on this second subset is referred to as *curtailment.* Curtailment comes at a price when testing errors

(specifically false positives) are possible. When retesting is not curtailed, the items in positively testing groups undergo retesting which reduces the false-positive rate. This advantage of compositing is lost with curtailed retesting. The effects of testing error are examined in greater detail in Section 2.2.8.

Now consider the exhaustive retesting (Dorfman) procedure and suppose the individual item values in a positively testing composite are denoted by $X_1, \ldots, X_k$. Let these items be retested in sequential order and suppose after the first $k-1$ retests that $X_1 = X_2 = \cdots = X_{k-1} = 0$. Then we know that the last item must have the trait ($X_k = 1$) without testing and the items can be completely classified with only $k - 1$ instead of k retests (see Fig. 2.9). Let T_k be the number of tests required to completely classify a composite of size k using curtailed exhaustive retesting. Writing $q = 1 - p$ as above, T_k takes three possible values: 1, k, and $k + 1$. But

$$\begin{aligned} \Pr[T_k = 1] &= q^k, \\ \Pr[T_k = k] &= q^{k-1}p, \\ \Pr[T_k = k+1] &= 1 - q^k - q^{k-1}p = 1 - q^{k-1}. \end{aligned}$$

Thus

$$\begin{aligned} E[T_k] &= q^k + kq^{k-1}p + (k+1)(1 - q^{k-1}) \\ &= k + 1 - kq^k - q^{k-1}p. \end{aligned} \tag{2.5}$$

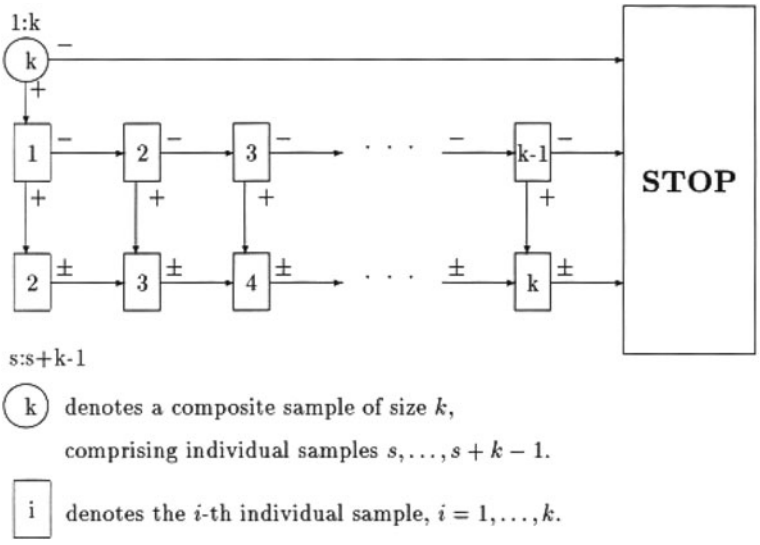

Fig. 2.9 Curtailed exhaustive retesting

The (asymptotic) relative cost is $\mathrm{RC} = E\left[T_k\right]/k = 1 - q^k + \frac{1}{k}\left[1 - q^{k-1}p\right]$. For selected values of p, the optimal composite sample size for the curtailed exhaustive retesting procedure is tabulated in Table 2.4 and the corresponding relative cost is shown in Fig. 2.10.

Table 2.4 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for curtailed exhaustive retesting

p	k_{opt}	RC	p	k_{opt}	RC
(0.38196, 1.00000]	1	1.000000	(0.00051, 0.00054]	44	(0.044905, 0.046202]
(0.20261, 0.38196]	2	(0.783386, 0.999993]	(0.00049, 0.00051]	45	(0.044022, 0.044905]
(0.10291, 0.20261]	3	(0.583770, 0.783386]	(0.00047, 0.00049]	46	(0.043119, 0.044022]
(0.06010, 0.10291]	4	(0.457106, 0.583770]	(0.00045, 0.00047]	47	(0.042198, 0.043119]
(0.03906, 0.06010]	5	(0.373965, 0.457106]	(0.00043, 0.00045]	48	(0.041253, 0.042198]
(0.02733, 0.03906]	6	(0.315871, 0.373965]	(0.00041, 0.00043]	49	(0.040288, 0.041253]
(0.02017, 0.02733]	7	(0.273231, 0.315871]	(0.00039, 0.00041]	50	(0.039297, 0.040288]
(0.01549, 0.02017]	8	(0.240669, 0.273231]	(0.00038, 0.00039]	51	(0.038792, 0.039297]
(0.01226, 0.01549]	9	(0.214956, 0.240669]	(0.00036, 0.00038]	52	(0.037764, 0.038792]
(0.00994, 0.01226]	10	(0.194156, 0.214956]	(0.00035, 0.00036]	53	(0.037238, 0.037764]
(0.00822, 0.00994]	11	(0.177008, 0.194156]	(0.00034, 0.00035]	54	(0.036704, 0.037238]
(0.00691, 0.00822]	12	(0.162633, 0.177008]	(0.00033, 0.00034]	55	(0.036163, 0.036704]
(0.00589, 0.00691]	13	(0.150415, 0.162633]	(0.00031, 0.00033]	56	(0.035056, 0.036163]
(0.00508, 0.00589]	14	(0.139900, 0.150415]	(0.00030, 0.00031]	57	(0.034488, 0.035056]
(0.00443, 0.00508]	15	(0.130814, 0.139900]	(0.00029, 0.00030]	58	(0.033910, 0.034488]
(0.00389, 0.00443]	16	(0.122720, 0.130814]	(0.00028, 0.00029]	59	(0.033325, 0.033910]
(0.00345, 0.00389]	17	(0.115686, 0.122720]	(0.00027, 0.00028]	60	(0.032727, 0.033325]
(0.00308, 0.00345]	18	(0.109404, 0.115686]	(0.00026, 0.00027]	61	(0.032117, 0.032727]
(0.00276, 0.00308]	19	(0.103645, 0.109404]	(0.00025, 0.00026]	63	(0.031495, 0.032117]
(0.00249, 0.00276]	20	(0.098514, 0.103645]	(0.00024, 0.00025]	64	(0.030864, 0.031495]
(0.00226, 0.00249]	21	(0.093915, 0.098514]	(0.00023, 0.00024]	65	(0.030216, 0.030864]
(0.00206, 0.00226]	22	(0.089714, 0.093915]	(0.00022, 0.00023]	66	(0.029553, 0.030216]
(0.00188, 0.00206]	23	(0.085749, 0.089714]	(0.00021, 0.00022]	68	(0.028876, 0.029553]
(0.00173, 0.00188]	24	(0.082298, 0.085749]	(0.00020, 0.00021]	70	(0.028181, 0.028876]
(0.00159, 0.00173]	25	(0.078932, 0.082298]	(0.00019, 0.00020]	71	(0.027473, 0.028181]
(0.00147, 0.00159]	26	(0.075926, 0.078932]	(0.00018, 0.00019]	73	(0.026742, 0.027473]
(0.00137, 0.00147]	27	(0.073326, 0.075926]	(0.00017, 0.00018]	75	(0.025990, 0.026742]
(0.00127, 0.00137]	28	(0.070623, 0.073326]	(0.00016, 0.00017]	77	(0.025216, 0.025990]
(0.00118, 0.00127]	29	(0.068096, 0.070623]	(0.00015, 0.00016]	80	(0.024421, 0.025216]
(0.00111, 0.00118]	30	(0.066067, 0.068096]	(0.00014, 0.00015]	82	(0.023594, 0.024421]
(0.00104, 0.00111]	31	(0.063967, 0.066067]	(0.00013, 0.00014]	85	(0.022738, 0.023594]
(0.00097, 0.00104]	32	(0.061793, 0.063967]	(0.00012, 0.00013]	88	(0.021847, 0.022738]
(0.00091, 0.00097]	33	(0.059865, 0.061793]	(0.00011, 0.00012]	92	(0.020918, 0.021847]
(0.00086, 0.00091]	34	(0.058211, 0.059865]	(0.00010, 0.00011]	96	(0.019951, 0.020918]
(0.00081, 0.00086]	35	(0.056507, 0.058211]	(0.00009, 0.00010]	100	(0.018929, 0.019951]
(0.00077, 0.00081]	36	(0.055104, 0.056507]	(0.00008, 0.00009]	106	(0.017847, 0.018929]
(0.00073, 0.00077]	37	(0.053665, 0.055104]	(0.00007, 0.00008]	112	(0.016695, 0.017847]
(0.00069, 0.00073]	38	(0.052183, 0.053665]	(0.00006, 0.00007]	120	(0.015465, 0.016695]
(0.00065, 0.00069]	39	(0.050657, 0.052183]	(0.00005, 0.00006]	130	(0.014118, 0.015465]
(0.00062, 0.00065]	40	(0.049483, 0.050657]	(0.00004, 0.00005]	142	(0.012628, 0.014118]
(0.00059, 0.00062]	41	(0.048280, 0.049483]	(0.00003, 0.00004]	158	(0.010936, 0.012628]
(0.00056, 0.00059]	42	(0.047041, 0.048280]	(0.00002, 0.00003]	183	(0.008940, 0.010936]
(0.00054, 0.00056]	43	(0.046202, 0.047041]	(0.00001, 0.00002]	224	(0.006500, 0.008940]

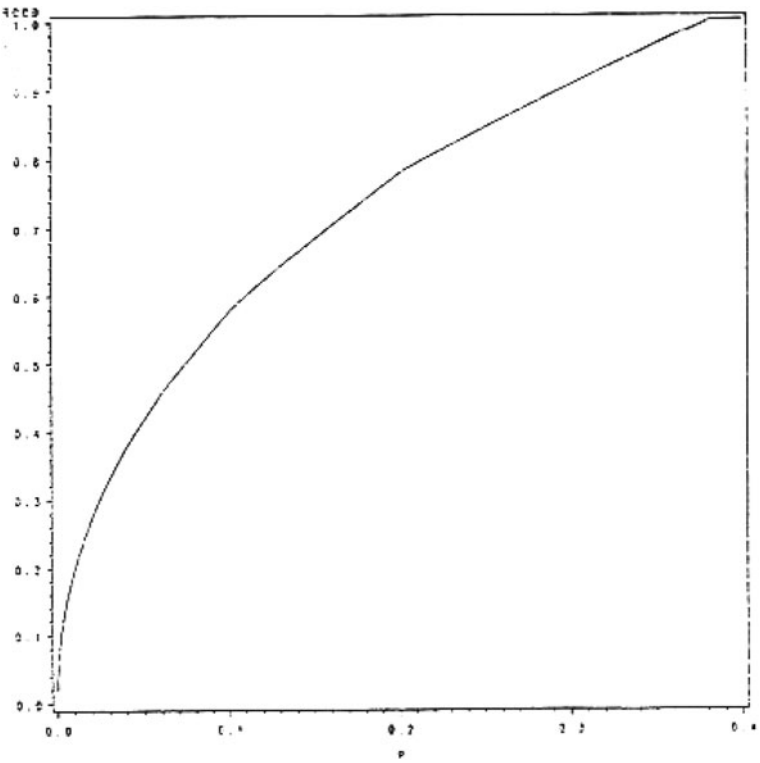

Fig. 2.10 Relative cost for curtailed exhaustive retesting using the optimal composite sample size

2.2.5 Curtailed Sequential Retesting

The sequential testing method can also be curtailed by avoiding a test of the last item whenever the first $k-1$ items of a composite test negative (see Fig. 2.11). The expected number of tests can be obtained by modifying the argument of Section 2.2.2. Direct calculation gives $E[T_1] = 1$; $E[T_2] = 3 - q - q^2$; and $E[T_3] = 5 - 2q - q^2 - q^3$. Let J be defined as in Section 2.2.2. Then

$$\begin{aligned} E[T_k] &= \sum_{j=0}^{k} E\left[T_k | J = j\right] P\left[J = j\right] \\ &= q^k + \sum_{j=1}^{k-1} E\left[T_k | J = j\right] q^{j-1} p + k q^{k-1} p \\ &= q^k + k q^{k-1} p + \sum_{j=1}^{k-1} \left\{j + 1 + E\left[T_{k-j}\right]\right\} q^{j-1} p \end{aligned}$$

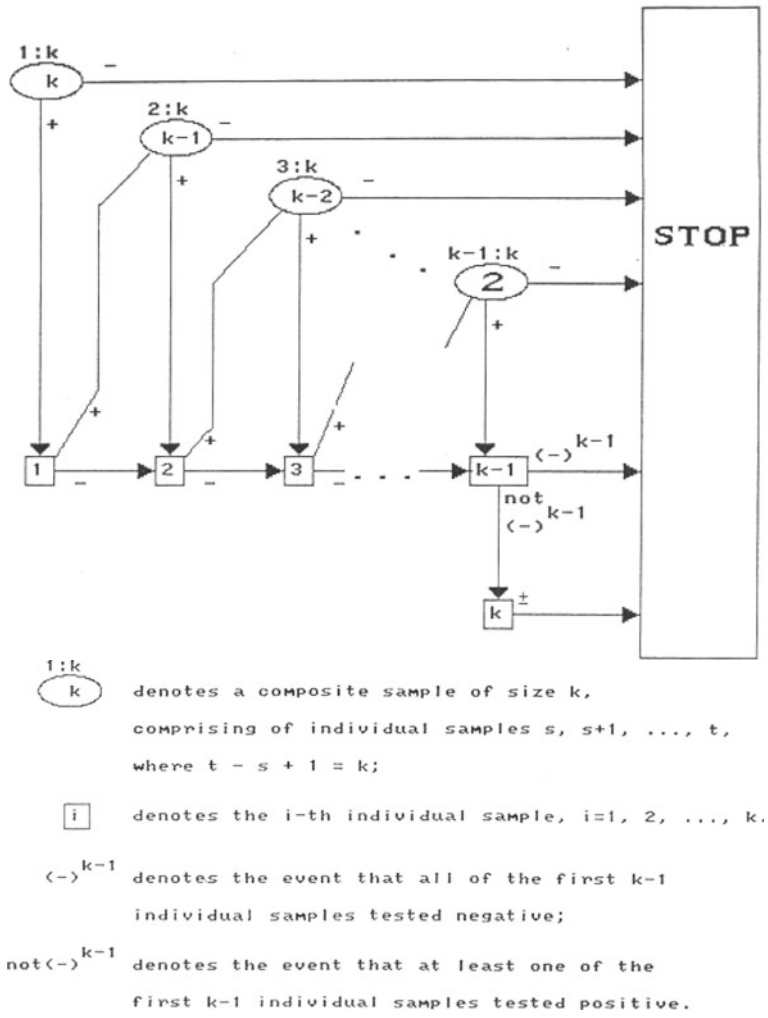

Fig. 2.11 Curtailed sequential retesting

$$= q^k + kq^{k-1}p + \sum_{j=0}^{k-2}(j+2)q^j p + \sum_{j=1}^{k-1} E\left[T_j\right] q^{k-j-1}p.$$

Similarly,

$$qE\left[T_{k-1}\right] = q^k + (k-1)q^{k-1}p + \sum_{j=0}^{k-3}(j+2)q^{j+1}p$$

$$+ \sum_{j=1}^{k-2} E\left[T_j\right] q^{k-j-1}p$$

$$= q^k + (k-1)q^{k-1}p + \sum_{j=1}^{k-2}(j+1)q^j p$$
$$+ \sum_{j=1}^{k-2} E\left[T_j\right] q^{k-j-1}p, \quad k = 3, 4, \ldots.$$

Thus, again

$$E\left[T_k\right] - qE\left[T_{k-1}\right] = q^{k-1}p + 2p + \sum_{j=1}^{k-2} q^j p + E\left[T_{k-1}\right] p$$

or

$$E\left[T_k\right] - E\left[T_{k-1}\right] = q^{k-1}p + 2p + pq\left(1 - q^{k-2}\right)/(1-q)$$
$$= q^{k-1}p + 2p + q\left(1 - q^{k-2}\right), \quad k = 3, 4, \ldots.$$

The number of tests takes any value (except 2) from 1 to $(2k-1)$. The average number of tests for classifying k individual samples is therefore given by

$$E\left[T_k\right] = k(2-q) + 2q - \left(1 - q^{k+1}\right)/p \tag{2.6}$$

and

$$\mathrm{RC} = 2 - q + \frac{1}{k}\left[2q - \frac{1-q^{k+1}}{p}\right]. \tag{2.7}$$

For example, if $k = 5$ and $p = 0.15$, then the expected number of tests is 3.30 and the relative cost is 0.66. So, sequential retesting would require only 66% as many tests as conventional testing of all individual items. By comparison, curtailed exhaustive retesting gives RC = 0.74 for $k = 5$ and $p = 0.15$. For these parameters, the curtailed sequential retesting procedure is more efficient than the curtailed exhaustive retesting procedure.

The optimal composite sample size for the curtailed sequential retesting procedure is tabulated for selected values of p in Table 2.5. Figure 2.12 shows the relative cost of classification when the optimal composite sample size is used for the curtailed sequential retesting procedure.

Table 2.5 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for curtailed sequential retesting

p	k_{opt}	RC	p	k_{opt}	RC
(0.38196, 1.00000]	1	1.000000	(0.00064, 0.00066]	56	(0.036501, 0.037078]
(0.26101, 0.38196]	2	(0.857449, 0.999993]	(0.00061, 0.00064]	57	(0.035618, 0.036501]
(0.16832, 0.26101]	3	(0.689891, 0.857449]	(0.00059, 0.00061]	58	(0.035020, 0.035618]
(0.10986, 0.16832]	4	(0.551025, 0.689891]	(0.00057, 0.00059]	59	(0.034409, 0.035020]
(0.07506, 0.10986]	5	(0.448910, 0.551025]	(0.00056, 0.00057]	60	(0.034102, 0.034409]
(0.05381, 0.07506]	6	(0.374861, 0.448910]	(0.00054, 0.00056]	61	(0.033477, 0.034102]
(0.04022, 0.05381]	7	(0.320154, 0.374861]	(0.00052, 0.00054]	62	(0.032839, 0.033477]
(0.03109, 0.04022]	8	(0.278530, 0.320154]	(0.00050, 0.00052]	63	(0.032191, 0.032839]
(0.02471, 0.03109]	9	(0.246079, 0.278530]	(0.00048, 0.00050]	64	(0.031529, 0.032191]
(0.02008, 0.02471]	10	(0.220107, 0.246079]	(0.00047, 0.00048]	65	(0.031193, 0.031529]
(0.01664, 0.02008]	11	(0.199027, 0.220107]	(0.00045, 0.00047]	66	(0.030513, 0.031193]
(0.01400, 0.01664]	12	(0.181486, 0.199027]	(0.00044, 0.00045]	67	(0.030166, 0.030513]
(0.01194, 0.01400]	13	(0.166741, 0.181486]	(0.00043, 0.00044]	68	(0.029815, 0.030166]
(0.01030, 0.01194]	14	(0.154162, 0.166741]	(0.00042, 0.00043]	69	(0.029460, 0.029815]
(0.00897, 0.01030]	15	(0.143279, 0.154162]	(0.00040, 0.00042]	70	(0.028740, 0.029460]
(0.00789, 0.00897]	16	(0.133894, 0.143279]	(0.00038, 0.00040]	72	(0.028001, 0.028740]
(0.00699, 0.00789]	17	(0.125616, 0.133894]	(0.00036, 0.00038]	74	(0.027244, 0.028001]
(0.00623, 0.00699]	18	(0.118237, 0.125616]	(0.00035, 0.00036]	75	(0.026856, 0.027244]
(0.00559, 0.00623]	19	(0.111699, 0.118237]	(0.00034, 0.00035]	76	(0.026464, 0.026856]
(0.00504, 0.00559]	20	(0.105800, 0.111699]	(0.00033, 0.00034]	78	(0.026066, 0.026464]
(0.00457, 0.00504]	21	(0.100521, 0.105800]	(0.00032, 0.00033]	79	(0.025663, 0.026066]
(0.00417, 0.00457]	22	(0.095828, 0.100521]	(0.00031, 0.00032]	80	(0.025255, 0.025663]
(0.00381, 0.00417]	23	(0.091421, 0.095828]	(0.00030, 0.00031]	81	(0.024837, 0.025255]
(0.00350, 0.00381]	24	(0.087471, 0.091421]	(0.00029, 0.00030]	83	(0.024413, 0.024837]
(0.00323, 0.00350]	25	(0.083896, 0.087471]	(0.00028, 0.00029]	84	(0.023986, 0.024413]
(0.00298, 0.00323]	26	(0.080459, 0.083896]	(0.00027, 0.00028]	85	(0.023547, 0.023986]
(0.00276, 0.00298]	27	(0.077320, 0.080459]	(0.00026, 0.00027]	87	(0.023101, 0.023547]
(0.00257, 0.00276]	28	(0.074516, 0.077320]	(0.00025, 0.00026]	88	(0.022647, 0.023101]
(0.00239, 0.00257]	29	(0.071768, 0.074516]	(0.00024, 0.00025]	89	(0.022185, 0.022647]
(0.00224, 0.00239]	30	(0.069404, 0.071768]	(0.00023, 0.00024]	92	(0.021711, 0.022185]
(0.00209, 0.00224]	31	(0.066961, 0.069404]	(0.00022, 0.00023]	94	(0.021228, 0.021711]
(0.00196, 0.00209]	32	(0.064778, 0.066961]	(0.00021, 0.00022]	95	(0.020733, 0.021228]
(0.00185, 0.00196]	33	(0.062879, 0.064778]	(0.00020, 0.00021]	98	(0.020228, 0.020733]
(0.00174, 0.00185]	34	(0.060923, 0.062879]	(0.00019, 0.00020]	101	(0.019713, 0.020228]
(0.00164, 0.00174]	35	(0.059094, 0.060923]	(0.00018, 0.00019]	104	(0.019180, 0.019713]
(0.00155, 0.00164]	36	(0.057403, 0.059094]	(0.00017, 0.00018]	105	(0.018633, 0.019180]
(0.00147, 0.00155]	37	(0.055862, 0.057403]	(0.00016, 0.00017]	108	(0.018071, 0.018633]
(0.00139, 0.00147]	38	(0.054276, 0.055862]	(0.00015, 0.00016]	113	(0.017494, 0.018071]
(0.00132, 0.00139]	39	(0.052856, 0.054276]	(0.00014, 0.00015]	115	(0.016894, 0.017494]
(0.00125, 0.00132]	40	(0.051398, 0.052856]	(0.00013, 0.00014]	119	(0.016274, 0.016894]
(0.00119, 0.00125]	41	(0.050117, 0.051398]	(0.00012, 0.00013]	126	(0.015628, 0.016274]
(0.00113, 0.00119]	42	(0.048805, 0.050117]	(0.00011, 0.00012]	131	(0.014955, 0.015628]
(0.00108, 0.00113]	43	(0.047686, 0.048805]	(0.00010, 0.00011]	135	(0.014258, 0.014955]
(0.00104, 0.00108]	44	(0.046774, 0.047686]	(0.00009, 0.00010]	141	(0.013519, 0.014258]
(0.00099, 0.00104]	45	(0.045608, 0.046774]	(0.00008, 0.00009]	150	(0.012742, 0.013519]
(0.00094, 0.00099]	46	(0.044415, 0.045608]	(0.00007, 0.00008]	156	(0.011909, 0.012742]
(0.00091, 0.00094]	47	(0.043685, 0.044415]	(0.00006, 0.00007]	169	(0.011026, 0.011909]
(0.00087, 0.00091]	48	(0.042692, 0.043685]	(0.00005, 0.00006]	186	(0.010057, 0.011026]
(0.00083, 0.00087]	49	(0.041677, 0.042692]	(0.00004, 0.00005]	202	(0.008988, 0.010057]
(0.00080, 0.00083]	50	(0.040902, 0.041677]	(0.00003, 0.00004]	228	(0.007782, 0.008988]
(0.00077, 0.00080]	51	(0.040111, 0.040902]	(0.00002, 0.00003]	248	(0.006532, 0.007782]
(0.00074, 0.00077]	52	(0.039307, 0.040111]	(0.00001, 0.00002]	250	(0.005269, 0.006532]
(0.00071, 0.00074]	53	(0.038486, 0.039307]			
(0.00069, 0.00071]	54	(0.037929, 0.038486]			
(0.00066, 0.00069]	55	(0.037078, 0.037929]			

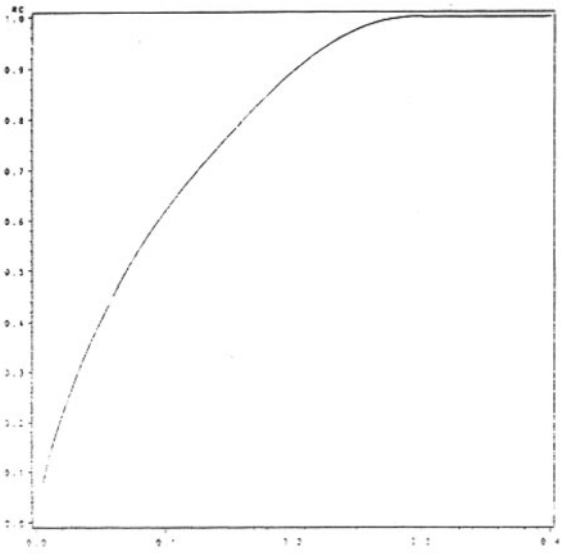

Fig. 2.12 Relative cost for curtailed sequential retesting using the optimal composite sample size

2.2.6 Curtailed Binary Split Retesting

The binary split retesting procedure of Gill and Gottlieb (2.4) can be curtailed so that only one of the two composite subsamples needs to be tested after certain binary splits. If the first of the two subcomposites tests negative, then the other subcomposite would test positive and the test need not be carried out. That is, the second composite is tested only when the first one tests positive. This can save up to half of all retesting efforts and costs. The following recursion formula can be obtained in the same way as that obtained for the uncurtailed procedure:

$$E\left[T_k\right] = E\left[T_{k_1}\right] + E\left[T_{k_2}\right] + 1 - q^{k-1} - q^k. \tag{2.8}$$

When k is even, then $k_1 = k_2 = k/2$. If k is odd, then the expected number of tests is smaller with $k_1 = (k-1)/2$ rather than with $k_1 = (k+1)/2$. With this choice for the value of k_1, the recursion formula in (2.8) can be used iteratively to obtain $E[T_k]$ for any given k. For example,

$$\begin{aligned}
E[T_2] &= 2E[T_1] + 1 - q - q^2 = 3 - q - q^2,\\
E[T_3] &= E[T_1] + E[T_2] + 1 - q - q^3 = 5 - 2q - q^2 - q^3,\\
E[T_4] &= 7 - 2q - 3q^2 - q^4,\\
E[T_5] &= 9 - 3q - 3q^2 - q^3 - q^5,\\
E[T_6] &= 11 - 4q - 2q^2 - 3q^3 - q^6,
\end{aligned}$$

$$
\begin{aligned}
E[T_7] &= 13 - 4q - 4q^2 - 2q^3 - q^4 - q^7, \\
E[T_8] &= 15 - 4q - 6q^2 - 3q^4 - q^8, \\
E[T_9] &= 17 - 5q - 6q^2 - q^3 - 2q^4 - q^5 - q^9, \\
E[T_{10}] &= 19 - 6q - 6q^2 - 2q^3 - 3q^5 - q^{10}, \\
E[T_{11}] &= 21 - 7q - 5q^2 - 4q^3 - 3q^5 - q^{11}, \\
E[T_{12}] &= 23 - 8q - 4q^2 - 6q^3 - 2q^5 - q^6 - q^{12}.
\end{aligned}
$$

The (asymptotic) relative cost can be calculated by dividing the expected number of tests by the corresponding composite sample size. Table 2.6 shows the optimal composite sample size corresponding to selected values of p for the curtailed binary split procedure, while Fig. 2.13 shows the relative cost of classification for this procedure when used with the optimal composite sample size.

Table 2.6 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for curtailed binary split retesting

p	k_{opt}	RC
(0.38196, 1.00000]	1	1.00
(0.26101, 0.38196]	2	(0.857449, 0.999993]
(0.16582, 0.26101]	3	(0.685100, 0.857449]
(0.10106, 0.16582]	5	(0.513090, 0.685100]
(0.08439, 0.10106]	7	(0.458100, 0.513090]
(0.06817, 0.08439]	9	(0.397968, 0.458100]
(0.06640, 0.06817]	10	(0.391045, 0.397968]
(0.05054, 0.06640]	11	(0.325611, 0.391045]
(0.04316, 0.05054]	13	(0.292382, 0.325611]
(0.03346, 0.04316]	19	(0.243240, 0.292382]
(0.02563, 0.03346]	21	(0.200881, 0.243240]
(0.02170, 0.02563]	27	(0.177679, 0.200881]
(0.01709, 0.02170]	37	(0.148125, 0.177679]
(0.01689, 0.01709]	38	(0.146811, 0.148125]
(0.01659, 0.01689]	42	(0.144806, 0.146811]
(0.01282, 0.01659]	43	(0.119011, 0.144806]
(0.01090, 0.01282]	53	(0.105004, 0.119011]
(0.00842, 0.01090]	75	(0.085534, 0.105004]
(0.00643, 0.00842]	85	(0.069102, 0.085534]
(0.00546, 0.00643]	107	(0.060614, 0.069102]
(0.00426, 0.00546]	149	(0.049464, 0.060614]
(0.00422, 0.00426]	150	(0.049086, 0.049464]
(0.00414, 0.00422]	170	(0.048319, 0.049086]
(0.00321, 0.00414]	171	(0.039262, 0.048319]
(0.00249, 0.00321]	213	(0.031876, 0.039262]
(0.00244, 0.00249]	214	(0.031355, 0.031876]
(0.00238, 0.00244]	234	(0.030722, 0.031355]
(0.00186, 0.00238]	235	(0.025168, 0.030722]
(0.00151, 0.00186]	245	(0.021335, 0.025168]
(0.00148, 0.00151]	246	(0.021003, 0.021335]

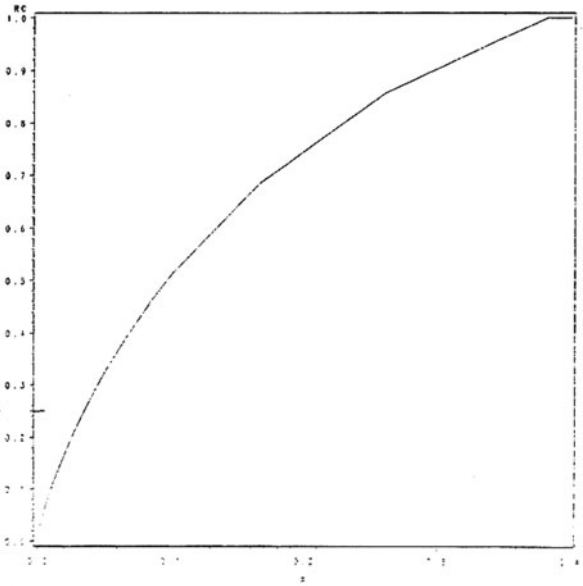

Fig. 2.13 Relative cost for curtailed binary split retesting

2.2.7 Entropy-Based Retesting

A source of inefficiency with hierarchical models such as binary splitting is that we may be required to test subcomposites of smaller and smaller sizes; but these sizes may be substantially less than the optimal size. In binary splitting, for example, suppose an original composite of size k has tested positive and has been split into two subcomposites of sizes k_1 and k_2. When the first subcomposite also tests positive, then we have no information regarding the status of the second composite of size $k_2 < k$. Instead of testing this subcomposite, we would be better off to return its k_2 items to the pool of unclassified items and select a composite of size k from that pool. This is the essence of the entropy-based retesting method due to Hwang (1984).

Suppose we have a large collection of items to be classified. At each stage of the entropy method, this collection is divided into three disjoint parts: those items that have been classified as positive, those that have been classified as negative, and the remaining unclassified pool. Let k be a fixed composite sample size and write k^* for the smaller of k and the current size of the unclassified pool. The entropy-based retesting method is described below. Notice that whenever items are put back into the unclassified pool, we have no information about these items. Throughout, then, the unclassified pool can be looked upon as a collection of independent Bernoulli trials with an unchanging parameter p.

Step 1. If $k^* = 0$, then exit; otherwise, select a subset of size k^* from the unclassified pool and test the resulting composite. If the test is negative, classify all k^* items as negative and go to Step 1. Otherwise, go to Step 2 with $k' = k^*$.

Step 2. Here we have a set of k' items whose corresponding composite has tested (or would test) positive. If $k' = 1$, classify the item as positive and go to Step 1. Otherwise, split the k' items into two disjoint subsets, of sizes $k_1 \leq k_2$, with k_1 and k_2 as nearly equal as possible. Form a composite of size k_1 from the first of these subsets and test the composite.

Step 2a. If the test is negative, classify all k_1 items as negative and go to Step 2 with the remaining $k' = k_2$ items.

Step 2b. If the test is positive, return the remaining k_2 items to the unclassified pool and go to Step 2 with $k' = k_1$ items.

Notice that the algorithm differs from binary retesting only in Step 2b, where items are returned to the unclassified pool. Starting from a given composite in Step 1, processing of that composite results in at most one item being classified as positive; along the way, several items may be classified as negative.

Let the k^* items in the composite at Step 1 be arranged in sequence from left to right and let the binary splitting maintain this sequential arrangement. If J is the position of the first item in the sequence that possesses the trait, then processing of the composite results in classifying the first $J - 1$ items as negative, classifying the Jth item as positive, and returning the remaining $k^* - J$ items to the unclassified pool. Consistent with earlier sections, we write $J = 0$ when a starting composite contains no item with the trait. In this case all k^* items are classified as negative. Write T_m for the number of tests required to fully classify a pool of m items using composites of size k.

Near the end of the classification procedure, smaller composite sample sizes may have to be used. For the finite case, when the number of unclassified individual samples drops below k, the composite sample size being used, all the remaining individual samples are used to form the next composite sample. The following cases develop the asymptotic relative cost and give several finite relative costs for various composite sample sizes.

Case 1. Composites of Size $k = 2$ Are Used. Processing each composite requires one or two tests, so that

$$\begin{aligned} E[T_m] &= E[T_m|J=0]\,q^2 + E[T_m|J=1]\,p + E[T_m|J=2]\,qp \\ &= \left\{1 + E\left[T_{m-2}\right]\right\} q^2 + \left\{2 + E\left[T_{m-1}\right]\right\} p + \left\{2 + E\left[T_{m-2}\right]\right\} qp \\ &= q^2 + 2p + 2qp + pE\left[T_{m-1}\right] + qE\left[T_{m-2}\right] \\ &= 2 - q^2 + pE\left[T_{m-1}\right] + qE\left[T_{m-2}\right], \quad m = 2, 3, \ldots. \end{aligned}$$

Since $T_1 = 1$ and $T_0 = 0$, we obtain

$$E[T_2] = 2 - q^2 + p = 3 - q - q^2,$$

$$
\begin{aligned}
E[T_3] &= 2 - q^2 + p\left(3 - q - q^2\right) + q \\
&= 5 - 3q - q^2 + q^3, \\
E[T_4] &= 2 - q^2 + pE[T_3] + q[T_2] \\
&= 2 - q^2 + p\left(5 - 3q - q^2 + q^3\right) + q\left(3 - q - q^2\right) \\
&= 7 - 5q + q^3 - q^4, \\
E[T_5] &= 9 - 7q + q^2 - q^4 + q^5.
\end{aligned}
$$

We observe that $E[T_1] - E[T_0] = 1$, and

$$
\begin{aligned}
E[T_2] - E[T_1] &= 2 - q - q^2, \\
E[T_3] - E[T_2] &= 2 - 2q + q^3, \\
E[T_4] - E[T_3] &= 2 - 2q + q^2 - q^4, \\
E[T_5] - E[T_4] &= 2 - 2q + q^2 - q^3 + q^5.
\end{aligned}
$$

Recall that $E[T_m] = 2 - q^2 + pE\left[T_{m-1}\right] + qE\left[T_{m-2}\right]$, $m = 2, 3, \ldots$. So, $qE\left[T_{m-1}\right] = 2q - q^3 + pqE\left[T_{m-2}\right] + q^2E\left[T_{m-3}\right]$. Subtracting and simplifying gives

$$
E[T_m] - E\left[T_{m-1}\right] = p(2 - q^2) + q^2\left\{E\left[T_{m-2}\right] - E\left[T_{m-3}\right]\right\}, \quad m = 3, 4, \ldots.
$$

This difference is the average number of tests needed to classify one item when there are m items to be classified. For m large, this difference is essentially a constant independent of p. That is, for large m, the asymptotic relative cost (RC) can be found by solving

$$
\mathrm{RC} = p\left(2 - q^2\right) + q^2\,(\mathrm{RC}).
$$

Thus, the asymptotic relative cost is

$$
\mathrm{RC} = \frac{p\left(2 - q^2\right)}{1 - q^2} = \frac{2 - q^2}{1 + q}. \tag{2.9}
$$

Case 2. Composites of Size $k = 3$ Are Used. Processing of one composite results in one, two, or three tests. Consider

$$
\begin{aligned}
E[T_m] &= E[T_m|J = 0]\,q^3 + E[T_m|J = 1]\,p + E[T_m|J = 2]\,qp \\
&\quad + E[T_m|J = 3]\,q^2 \\
&= \left\{1 + E\left[T_{m-3}\right]\right\}q^3 + \left\{2 + E\left[T_{m-1}\right]\right\}p \\
&\quad + \left\{3 + E\left[T_{m-2}\right]\right\}qp + \left\{3 + E\left[T_{m-3-1}\right]\right\}q^2p \\
&= 2 + q - 2q^3 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2E\left[T_{m-3}\right], \quad \text{for} \quad m \geq 3.
\end{aligned}
$$

As before, $T_0 = 0$, $T_1 = 1$, and T_2 is the same as in Case 1. Thus,

$$E[T_2] = 3 - q - q^2,$$

and

$$\begin{aligned} E[T_3] &= 2 + q - 2q^3 + p\left(3 - q - q^2\right) + qp = 5 - 2q - q^2 - q^3, \\ E[T_4] &= 2 + q - 2q^3 + p\left(5 - 2q - q^2 - q^3\right) + qp\left(3 - q - q^2\right) + q^2 p \\ &= 7 - 3q - 2q^2 - 2q^3 + 2q^4, \\ E[T_5] &= 9 - 4q - 3q^2 - 2q^3 + 3q^4 - q^5. \end{aligned}$$

Recall

$$\begin{aligned} E[T_m] &= 2 + q - 2q^3 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2 E\left[T_{m-3}\right], \\ qE\left[T_{m-1}\right] &= 2q + q^2 - 2q^4 + qpE\left[T_{m-2}\right] + q^2 pE\left[T_{m-3}\right] + q^3 E\left[T_{m-4}\right]. \end{aligned}$$

Subtracting the latter from the former and simplifying gives

$$E[T_m] - E\left[T_{m-1}\right] = p\left(2 + q - 2q^3\right) + q^3\left\{E\left[T_{m-3}\right] - E\left[T_{m-4}\right]\right\}.$$

For large m, this difference is essentially a constant equal to the relative cost. Thus,

$$\mathrm{RC} = \frac{2 + q - 2q^3}{1 + q + q^2}. \tag{2.10}$$

Case 3. Composites of Size $k = 4$ Are Used. Again, processing of a composite results in one or three tests. Thus,

$$\begin{aligned} E[T_m] &= \left\{1 + E\left[T_{m-4}\right]\right\} q^4 + \left\{3 + E\left[T_{m-1}\right]\right\} p \\ &\quad + \left\{3 + E\left[T_{m-2}\right]\right\} qp + \left\{3 + E\left[T_{m-3}\right]\right\} q^2 p + \left\{3 + E\left[T_{m-4}\right]\right\} q^3 p \\ &= 3 - 2q^4 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2 pE\left[T_{m-3}\right] + q^3 E\left[T_{m-4}\right]. \end{aligned}$$

Subtracting $qE\left[T_{m-1}\right]$ and simplifying gives

$$E[T_m] - E\left[T_{m-1}\right] = p\left(3 - 2q^4\right) + q^4\left\{E\left[T_{m-4}\right] - E\left[T_{m-5}\right]\right\}.$$

For large m, the relative cost is

$$\mathrm{RC} = \frac{3 - 2q^4}{1 + q + q^2 + q^3}. \tag{2.11}$$

Case 4. Composites of Size $k = 5$ Are Used. In this case, one, three, or four tests are required for processing a composite. Thus,

$$\begin{aligned} E[T_m] &= \left\{1 + E\left[T_{m-5}\right]\right\} q^5 + \left\{3 + E\left[T_{m-1}\right]\right\} p + \left\{3 + E\left[T_{m-2}\right]\right\} qp \\ &\quad + \left\{3 + E\left[T_{m-3}\right]\right\} q^2 p + \left\{4 + E\left[T_{m-4}\right]\right\} q^3 p + \left\{4 + E\left[T_{m-5}\right]\right\} q^4 p \\ &= 3 + q^3 - 3q^5 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2 pE\left[T_{m-3}\right] \\ &\quad + q^3 pE\left[T_{m-4}\right] + q^4 E\left[T_{m-5}\right]. \end{aligned}$$

Subtracting $qE\left[T_{m-1}\right]$ and simplifying gives

$$E[T_m] - E\left[T_{m-1}\right] = p\left(3 + q^3 - 3q^5\right) + q^5\left\{E\left[T_{m-5}\right] - E\left[T_{m-6}\right]\right\}.$$

For large m this is the relative cost resulting in

$$\mathrm{RC} = \frac{3 + q^3 - 3q^5}{1 + q + q^2 + q^3 + q^4}. \tag{2.12}$$

Note that the asymptotic relative cost with a composite sample size k is the ratio of two polynomials in q, the denominator being a geometric series $1 + q + q^2 + \cdots + q^k$. The numerators of the respective relative costs corresponding to composite sample sizes of $k = 6, 7, 8, 9$, and 10 are tabulated below:

k	Numerator of the relative cost
6	$3 + q - q^3 + q^4 - 3q^6$
7	$3 + q - 3q^7$
8	$4 - 3q^8$
9	$4 + q^7 - 4q^9$
10	$4 + q^3 - q^5 + q^8 - 4q^{10}$

The optimal composite sample size depends on the number of individual samples to be classified and on the prevalence p of individual samples possessing the trait, and hence testing positive. Table 2.7 gives the optimal composite sample size as a function of p for values of $m = 2$–12. See also Fig. 2.14 for the relative costs corresponding to these composite sample sizes.

Table 2.7 can be used interactively by choosing or estimating the optimal composite sample size to use initially. After the first composite sample has been processed, the optimal composite sample size is determined by entering the table with the number of remaining individual samples to be classified. Also, it is possible to update the estimate of q after some initial samples have been classified.

In general, one would like to optimize the entropy-based procedure by permitting k in Step 1 and (k_1, k_2) in Step 2 to vary depending upon the current size of the unclassified pool. On the basis of exhaustive computer searches, Snyder and

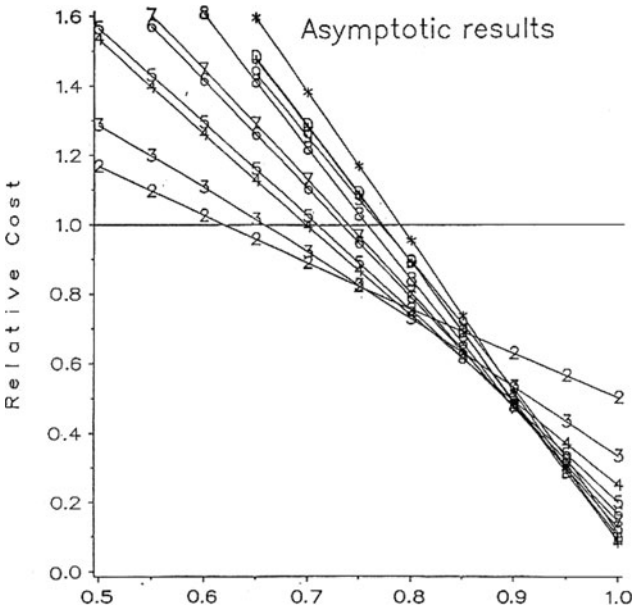

Fig. 2.14 Relative cost for composite sample sizes of $k = 2$–12

Larson (1969) have published such optimized algorithms for $p = 0.01(0.01)0.1$ and where the initial size of the unclassified pool can be as large as $m = 50$.

2.2.8 Exhaustive Retesting in the Presence of Classification Errors

The problem is to classify every individual sample as polluted or not polluted using presence/absence measurements. Suppose that there is a positive probability of misclassifying any sample, either individual or composite, and assume that the probability of misclassification depends only on whether or not the sample is polluted. In particular, composite samples and individual samples have the same misclassification rates. Let r_n be the probability of a false-negative classification. That is,

$$r_n = \Pr[\text{negative test result} \mid \text{sample is polluted}].$$

Similarly, let r_p be the probability of a false-positive classification. That is,

$$r_p = \Pr\left[\text{positive test result} \mid \text{sample is not polluted}\right].$$

Now consider using the exhaustive retesting procedure with composites of size k. Let

$$d_n = \Pr\left[\text{negative classification} \mid \text{individual sample is polluted}\right]$$

Table 2.7 Optimal composite sample size k over the tabulated ranges of p for $m = 1$–16, 25, 50, and ∞

k	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
1	0.0–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.0– 0.62
2	–	0.0–0.38	0.29–0.38	0.24–0.38	0.27–0.38	0.25–0.38	0.25–0.38	0.25– 0.38
3	–	–	0.0–0.29	–	0.19–0.27	0.17–0.25	0.21–0.25	0.19–0.25
4	–	–	–	0.0–0.24	–	0.16–0.17	0.14–0.21	0.12–0.19
5	–	–	–	–	0.0–0.19	–	–	–
6	–	–	–	–	–	0.0–0.16	–	–
7	–	–	–	–	–	–	0.0–0.14	–
8	–	–	–	–	–	–	–	0.0–0.12
k	$m=9$	$m=10$	$m=11$	$m=12$	$m=13$	$m=14$	$m=15$	$m=16$
1	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0
2	0.25–0.38	0.22–0.38	0.25–0.38	0.25–0.38	0.25–0.38	0.25–0.38	0.25–0.38	0.24–0.38
3	0.18–0.25	0.19–0.22	0.19–0.25	0.19–0.25	0.18–0.25	0.18–0.38	0.18–0.25	0.19–0.24
4	–	0.17–0.19	0.13–0.19	0.14–0.19	0.16–0.18	0.16–0.18	0.15–0.18	0.13–0.19
5	0.11–0.18	0.10–0.17	–	0.13–0.14	0.12–0.16	0.12–0.16	0.11–0.15	–
6	–	–	0.12–0.13	–	–	–	–	–
7	–	–	0.09–0.12	0.08–0.13	0.08–0.12	0.07–0.12	–	0.11–0.13
8	–	–	–	–	–	–	–	0.10–0.11
9	0.89–1.0	–	–	–	–	–	0.07–0.11	–
10	–	0.0–0.10	–	–	–	–	–	–
11	–	–	0.0–0.09	–	–	–	–	0.07–0.10
12	–	–	–	0.0–0.08	–	–	–	–
13	–	–	–	–	0.0–0.08	–	–	–
14	–	–	–	–	–	0.0– 0.07	0.0–0.07	–
15	–	–	–	–	–	–	–	–
16	–	–	–	–	–	–	–	0.0–0.07

$m = 25$

$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0.38–1.0	0.25–0.38	0.18–0.25	0.15–0.18	0.11–0.15
$k=7$	$k=9$	$k=10$	$k=16$	$k=25$
0.09–0.11	0.08–0.09	0.06–0.08	0.03–0.06	0.0–0.03

$m = 50$

$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=7$
0.38–1.0	0.25–0.38	0.18–0.25	0.15–0.18	0.11–0.15	0.09–0.11
$k=8$	$k=9$	$k=15$	$k=17$	$k=24$	$k=50$
0.08–0.09	0.06–0.08	0.05–0.06	0.03–0.05	?–0.03	0.0–?

∞ (asymptotic case)

$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0.38–1.0	0.25–0.38	0.18–0.25	0.14–0.18	0.11–0.14
$k=7$	$k=8$	$k=9$	$k=16$	$k=32$
0.09–0.11	0.08–0.09	0.06–0.08	0.03–0.06	?–0.03

The question mark (?) indicates the detection limit (which is a small positive number).

and

$$d_p = \Pr\left[\text{positive classification} \mid \text{individual sample is not polluted}\right].$$

Consider the computation of d_n. There are two ways in which an individual sample can be misclassified as negative. First, the composite is misclassified as negative, so that every individual sample is automatically (but incorrectly) classified as negative. The probability of this happening is r_n. Second, the composite is correctly

classified as positive (which occurs with probability $1 - r_n$), but the individual sample incorrectly tests negative (the probability of this happening is r_n). Thus

$$d_n = r_n + (1 - r_n)r_n = 2r_n - r_n^2.$$

If the misclassification rate r_n is small, then the negative misclassification rate with exhaustive retesting is approximately twice that for individual testing.

Similarly, the probability of a false-positive classification can be shown to be

$$d_p = r_p \left\{ r_p + \left(1 - q^{k-1}\right)\left(1 - r_n - r_p\right) \right\}.$$

For example, suppose $r_n = r_p = 0.2$ and suppose 10% of the individual samples are polluted. Using a composite sample size of $k = 4$, we have $d_n = 0.36$ and $d_p = 0.07$. Notice that d_n is approximately twice r_n while the false-positive rate under compositing, d_p, is substantially less than for individual testing.

Misclassification also affects the relative cost of a retesting procedure. The exhaustive retesting procedure results in either one test or $k + 1$ tests to process k individual samples. If the tests were free of error, then a single test is required when all k items are unpolluted. In the presence of testing error, a single test can also occur when the composite incorrectly tests as negative. Examining the two cases, we see that

$$\begin{aligned} \Pr[\text{one test}] &= r_n \left(1 - q^k\right) + \left(1 - r_p\right) q^k \\ &= r_n + q^k \left(1 - r_n - r_p\right). \end{aligned}$$

Also

$$\begin{aligned} \Pr[k + 1 \text{ tests}] &= 1 - \Pr[\text{one test}] \\ &= 1 - r_n - q^k \left(1 - r_n - r_p\right). \end{aligned}$$

The relative cost of classification becomes

$$\begin{aligned} \text{RC} &= \frac{1}{k} + \Pr[k + 1 \text{ tests}] \\ &= 1 + \frac{1}{k} - r_n - q^k \left(1 - r_n - r_p\right). \end{aligned}$$

Note that this expression reduces to the relative cost given in (2.2) when $r_n = r_p = 0$.

2.2.9 Other Costs

Laboratory procedures become more costly and error-prone when numerous steps must be performed in sequence. From this point of view, the exhaustive retesting

method of Dorfman is certainly the easiest composite strategy to implement since no more that two steps are required. By contrast, the sequential retesting method may require $k + 1$ steps.

But what, precisely, is meant by the number of steps? Conceptually, let us suppose that each test requires one unit of time and, further, that all tests are performed as soon as possible. That is, a test is deferred only if its execution requires the results of other, earlier, tests. Then a composite classification design is said to be *R-step* if its performance to completion could require as many as R units of time. We also refer to R as the *maximum duration* of the design. Clearly, $R = 2$ for exhaustive retesting.

Retesting strategies require that aliquots or duplicates be maintained for each separate item. The maintenance of these duplicates can be a significant portion of total cost and a relevant consideration in deciding which design to select. In addition, another source of error is introduced if true duplicates are difficult to achieve – and the impact of this error source will grow as the needed number of duplicates grows. We define the *maximum aliquot count* (MAC) to be the number of duplicates of each item that must be available to complete the procedure under all possible circumstances. For simple hierarchical designs the MAC is the same as the maximum duration of the design. For feedback designs like the entropy-based procedure, the MAC and the maximum duration can be different – and each can be unreasonably large. These issues are explored further in the Exercises.

2.3 Continuous Response Variables

The previous sections have examined retesting strategies for presence/absence measurements, i.e., where the response variable is binary. The rationale for the different strategies is found in the following two properties.

Property *N*. If the composite is negative, then *every* item in the composite is negative.

Property *P*. If the composite is positive, then *at least one* item in the composite is positive.

Property N is the fundamental premise of compositing for classification and accounts for the method's efficiency since it allows an entire group of items to be classified on the basis of a single measurement. Property P is the justification for curtailment and can lead to improved efficiencies, but is not fundamental to compositing for classification.

We now want to turn our attention to response variables X that are nonnegative but continuously distributed. An individual item is classified as positive if $X \geq c$ for that item, where c is a specified criterion level. The proportion of positive items, i.e., the *prevalence*, is given as $p = \Pr[X \geq c]$. Consider a composite whose k items have individual values $X_1, X_2, \ldots, X_k$. Barring measurement error, the measured

response Y on the composite is the average of $X_1, X_2, \ldots, X_k$ so that

$$kY = X_1 + X_2 + \cdots + X_k.$$

We cannot classify the composite as negative whenever $Y < c$, for then Property N would fail. But, since X is nonnegative, it is certainly the case that $X_i < c$ for all i whenever $X_1 + X_2 + \cdots + X_k < c$. Thus, Property N will be true provided a composite is classified as negative whenever $kY < c$. The probability of a negative composite is then

$$\begin{aligned} q_k &= \Pr[\text{composite is negative}] \\ &= \Pr[X_1 + X_2 + \cdots + X_k < c]. \end{aligned} \tag{2.13}$$

When the individual items can be regarded as statistically independent, then the distribution of $X_1 + X_2 + \cdots + X_k$ in (2.13) is that of the k-fold convolution of X. Throughout we suppose that the individual items can in fact be treated as independent.

The exhaustive retesting procedure is exactly the same as in the case of a binary response: Measurement is made on a composite of size k. All individual items are classified as negative, if the composite is negative; otherwise, measurement is made individually on each item. As in the case of a binary response, the relative saving is

$$\begin{aligned} \text{RS} &= \Pr[\text{composite is negative}] - 1/k \\ &= q_k - 1/k. \end{aligned} \tag{2.14}$$

Recall that $q_k = q^k = (1 - p)^k$ for a binary response variable. Notice that RS depends upon k and also upon p through the criterion level c. But RS also depends upon the distribution of X because of the k-fold convolution occurring in (2.13). This latter dependence is the principal distinction between the continuous and binary response scenarios. Unfortunately, the relative savings is very sensitive to the underlying distribution of X, as we shall see below.

We have computed the relative savings as a function of k and p for the gamma distribution with index parameter a and for the lognormal distribution with logarithmic variance σ^2. The results, for $k = 2, 3, 4$, are shown in Fig. 2.15. Since the relative savings, considered as a function of p, does not depend upon the scaling of X, it was convenient to index the distributions in Fig. 2.15 by their coefficients of variation CV. The corresponding parameter values are

CV	a	σ^2
1.5	0.44	1.09
1.0	1.00	0.83
0.5	4.00	0.47

For comparison purposes, Fig. 2.15 also shows the relative savings for a binary response variable (Bernoulli distribution).

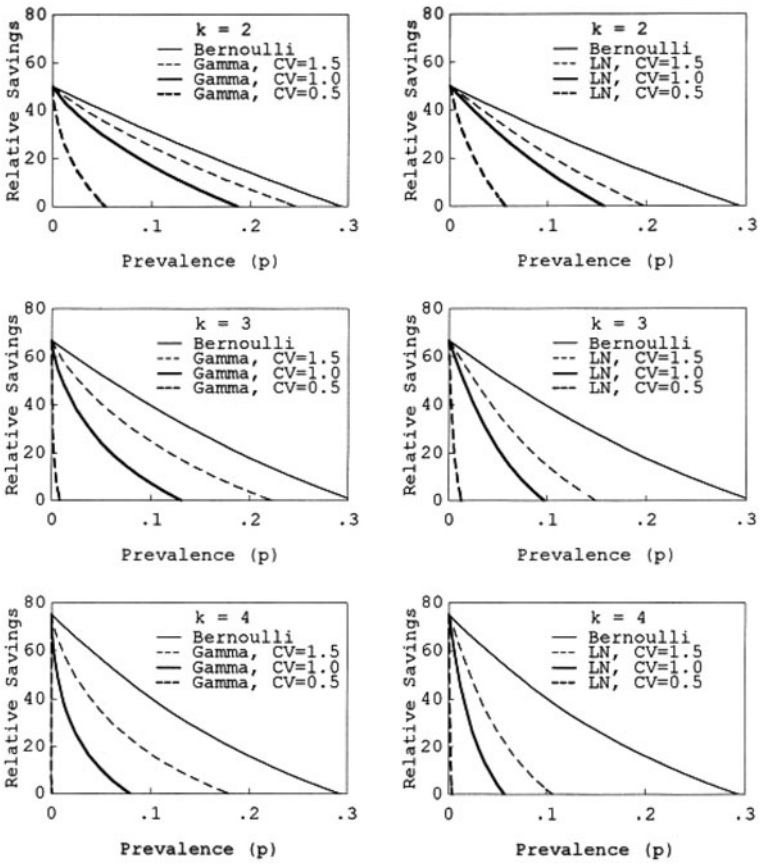

Fig. 2.15 Relative savings of exhaustive retesting when the response variable is binary (Bernoulli) or continuously distributed with a gamma distribution or a lognormal distribution (LN). Relative savings is expressed as a percent and the coefficient of variation is denoted by CV

Several conclusions emerge from an examination of Fig. 2.15:

1. For fixed k and p, the relative savings is greater for the binary response variable than for any continuous response variable.
2. Within each family of distributions (gamma and lognormal) the relative savings *increases* with the skewness of the distribution and, conversely, in the other direction the relative savings decreases as the distribution approaches normality.
3. For fixed p, the performance is quite sensitive to k and to the underlying distribution of X. The relative savings decreases rapidly with increasing k and with decreasing variability of X.

In view of the sensitivity described in item 3, there would be little purpose in attempting to optimize with respect to k. In fact, it appears unlikely that one would want to consider composite sizes much larger than $k = 2$ or $k = 3$.

We now examine items 1 and 2 to determine the extent to which they do or do not hold generally. Let p be fixed and determine the criterion level c by the requirement that $\Pr[X \geq c] = p$. We define a random variable $X(p)$ and its corresponding distribution by truncating X to the interval $[0, c)$ and then rescaling to the interval $[0, 1)$ so that

$$X(p) = \frac{1}{c} X|_{x<c}. \tag{2.15}$$

Now, the event $\{X_1 + \cdots + X_k < c\}$ is contained in the event $\{X_1 < c, \ldots, X_k < c\}$ which implies that q_k is given by

$$\begin{aligned} \Pr[X_1 + \cdots + X_k < c] &= \Pr[X_1 + \cdots + X_k < c | X_1 < c, \ldots, X_k < c] \\ &\quad \times \Pr[X_1 < c_1, \ldots, X_k < c] \\ &= \Pr[X_1(p) + \cdots + X_k(p) < 1] q^k, \end{aligned} \tag{2.16}$$

where $q = 1 - p$ and $X_1(p), \ldots, X_k(p)$ are independent realizations of $X(p)$. Equation (2.16) implies that

$$q_k \leq q^k, \tag{2.17}$$

so that the relative savings is always greater for a binary response variable than for a continuous response variable.

A distribution function F is stochastically smaller than a distribution function G provided $F(x) \geq G(x)$ uniformly in x. Given two response variables X and $\tilde{X}$, we shall say that X is more *zero-concentrated* than $\tilde{X}$ if the distribution of $X(p)$ is stochastically smaller than that of $\tilde{X}(p)$ for all p. In this case,

$$\Pr[X_1(p) + \cdots + X_k(p) < 1] \geq \Pr[\tilde{X}_1(p) + \cdots + \tilde{X}_k(p) < 1].$$

Comparing with (2.16), we see for all k and p that the relative savings of exhaustive retesting is greater for the more zero-concentrated response variable. This result is related to item 2 above. We generally associate skewness of a nonnegative random variable with a heavy right-hand tail. But for typical families of distributions, the right tail and the left tail are linked in a way that makes the left tail more zero-concentrated as the right tail becomes more elongated. Thus, typically but not always, a large skewness goes hand in hand with a large relative savings.

2.3.1 Quantitatively Curtailed Exhaustive Retesting

It would be possible to curtail the exhaustive retesting procedure along the lines used for a binary response variable. But the quantitative nature of the measurements allows for a more sophisticated form of curtailment. Suppose the measurement Y on a composite indicates that retesting is needed, i.e., $kY \geq c$. After the first j items have been measured, the individual values $X_1, \ldots, X_j$ are available and, consequently, the total for the remaining items can be calculated as

$$X_{j+1} + \cdots + X_k = kY - (X_1 + \cdots + X_j).$$

If this residual total is less than c, then these $k - j$ remaining items can be classified as negative without further testing. If the residual total exceeds c *and* if $j < k - 1$ then measurement is made on item $j + 1$ and the procedure iterates. If $j = k - 1$, then the value X_k is known so that item k can be classified without measurement on that item.

The number of retests R_k can take the values $0, 1, 2, \ldots, k - 1$ and the total number of measurements, $T_k = 1 + R_k$, ranges from 1 to k. Now $R_k \geq j$ means that item j must be tested, which is equivalent to saying that $X_j + \cdots + X_k \geq c$. Thus,

$$\begin{aligned}
&\Pr[R_k \geq 0] = 1 \\
&\Pr[R_k \geq 1] = \Pr[X_1 + \cdots + X_k \geq c] = 1 - q_k \\
&\Pr[R_k \geq 2] = \Pr[X_2 + \cdots + X_k \geq c] = 1 - q_{k-1}.
\end{aligned}$$

This pattern continues until

$$\Pr[R_k \geq k - 1] = \Pr[X_{k-1} + X_k \geq c] = 1 - q_2.$$

Now

$$\begin{aligned}
E[T_k] &= \Pr[T_k > 0] + \Pr[T_k > 1] + \cdots + \Pr[T_k > k - 1] \\
&= \Pr[R_k \geq 0] + \Pr[R_k \geq 1] + \cdots + \Pr[R_k \geq k - 1] \\
&= k - (q_2 + q_3 + \cdots + q_k).
\end{aligned}$$

From this, we obtain that the relative cost is

$$\mathrm{RC} = (q_2 + q_3 + \cdots + q_k)/k,$$

and the relative savings becomes

$$\mathrm{RS} = 1 - (q_2 + q_3 + \cdots + q_k)/k.$$

As the exercises indicate, quantitative curtailment improves the performance of exhaustive retesting rather markedly. But keep in mind that the derivations of this

section do not account for the effects of measurement error, imperfect mixing, or imperfect duplicates.

2.3.2 Binary Split Retesting

As in the case of presence/absence measurement, any positively testing composite sample is divided into two groups as nearly equal in size as possible. In the presence/absence case, both the composite samples are tested. For continuous measurements, only one composite sample need be formed and tested. The value of the second composite sample can be calculated. Let T_k be the number of tests necessary to classify all k samples starting with a composite sample of size k. Let $U_j = X_1 + \cdots + X_j$, and let $V_j = X_{j+1} + \cdots + X_k = U_k - U_j$. Here we assume that the X_i's are independent and identically distributed Bernoulli random variables. Consider the five mutually exclusive cases defined in terms of values of J_k given below.

Let $\left[\frac{k}{2}\right]$ be the largest integer less than or equal to $k/2$. Then $V_{\left[\frac{k}{2}\right]}$ has the same distribution as $U_{k-\left[\frac{k}{2}\right]}$. Let

$$J_k = \begin{cases} 0 & \text{if} \quad U_k < c \\ 1 & \text{if} \quad U_k \geq c, U_{\left[\frac{k}{2}\right]} < c, V_{\left[\frac{k}{2}\right]} < c \\ 2 & \text{if} \quad U_{\left[\frac{k}{2}\right]} \geq c, V_{\left[\frac{k}{2}\right]} < c \\ 3 & \text{if} \quad U_{\left[\frac{k}{2}\right]} < c, V_{\left[\frac{k}{2}\right]} \geq c \\ 4 & \text{if} \quad U_{\left[\frac{k}{2}\right]} \geq c, V_{\left[\frac{k}{2}\right]} \geq c. \end{cases}$$

Then

$$\begin{aligned} \Pr[J_k = 0] &= q_k, \\ \Pr[J_k = 2] &= \left(1 - q_{\left[\frac{k}{2}\right]}\right) q_{k-\left[\frac{k}{2}\right]}, \\ \Pr[J_k = 3] &= q_{\left[\frac{k}{2}\right]} \left(1 - q_{k-\left[\frac{k}{2}\right]}\right), \\ \Pr[J_k = 4] &= \left(1 - q_{\left[\frac{k}{2}\right]}\right) \left(1 - q_{k-\left[\frac{k}{2}\right]}\right). \end{aligned}$$

Thus

$$\begin{aligned} \Pr[J_k = 1] &= 1 - \Pr[J_k = 0] - \Pr[J_k = 2] - \Pr[J_k = 3] - \Pr\left[j_k = 4\right] \\ &= q_{\left[\frac{k}{2}\right]} q_{k-\left[\frac{k}{2}\right]} - q_k. \end{aligned}$$

So

$$
\begin{aligned}
E[T_k] &= E[E(T_k|J_k)] = \sum_{j=0}^{4} E(T_k|J_k = j)\Pr\left[J_k = j\right] \\
&= 1 \cdot \Pr[J_k = 0] + 2\Pr[J_k = 1] + \left(1 + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\right)\Pr[J_k = 2] \\
&\quad + \left(1 + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\right)\Pr[J_k = 3] \\
&\quad + \left(2 + \left\{E\left(T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right) - 1\right\} + \left\{E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil}\right] - 1\right\}\right)\Pr[J_k = 4] \\
&= \Pr[J_k = 0] + 2\Pr[J_k = 1] + \Pr[J_k = 2] + \Pr[J_k = 3] \\
&\quad + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\{\Pr[J_k = 2] + \Pr[J_k = 4]\} \\
&\quad + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\{\Pr[J_k = 3] + \Pr[J_k = 4]\}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
E[T_k] &= 1 + \Pr[J_k = 1] - \Pr[J_k = 4] + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{\lceil \frac{k}{2} \rceil}\right) \\
&\quad + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{k-\lceil \frac{k}{2} \rceil}\right) \\
&= q_{\lceil \frac{k}{2} \rceil} + q_{k-\lceil \frac{k}{2} \rceil} - q_k \\
&\quad + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{\lceil \frac{k}{2} \rceil}\right) \\
&\quad + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{k-\lceil \frac{k}{2} \rceil}\right).
\end{aligned}
$$

Now

$$
\begin{aligned}
E\left[T_j\right] &= E\left[T_j|U_j < c\right]\Pr\left[U_j < c\right] + E\left[T_j|U_j \geq c\right]\Pr\left[U_j \geq c\right] \\
&= q_j + E\left[T_j|U_j \geq c\right](1 - q_j).
\end{aligned}
$$

So

$$
E\left[T_j|U_j \geq c\right] = \frac{E\left[T_j\right] - q_j}{1 - q_j}.
$$

Therefore,

$$E[T_k] = q_{\lceil \frac{k}{2} \rceil} + q_{k-\lceil \frac{k}{2} \rceil} - q_k + E\left[T_{\lceil \frac{k}{2} \rceil}\right] - q_{\lceil \frac{k}{2} \rceil} + E\left[T_{k-\lceil \frac{k}{2} \rceil}\right] - q_{k-\lceil \frac{k}{2} \rceil}$$
$$= E\left[T_{\lceil \frac{k}{2} \rceil}\right] + E\left[T_{k-\lceil \frac{k}{2} \rceil}\right] - q_k, \quad k = 2, 3, \ldots.$$

Now $E[T_1] = 1$ so, $E[T_2] = 2 - q_2$, $E[T_3] = 3 - q_2 - q_3$, $E[T_4] = 2(2 - q_2) - q_4$, etc.

This recurrence (difference) equation can be solved iteratively for composite samples of arbitrary size k. If the composite sample size is a power of 2, the recurrence formula can be solved, giving

$$E(T_{2^r}) = 2^r - q_{2^r} - 2q_{2^{r-1}} - \cdots - 2^{r-1}q_2, \quad r = 2, 3, \ldots. \tag{2.18}$$

The following examples may be useful:

$$\begin{aligned}
E[T_5] &= 5 - 2q_2 - q_3 - q_5, \\
E[T_6] &= 6 - 2q_2 - 2q_3 - q_6, \\
E[T_7] &= 7 - 3q_2 - q_3 - q_4 - q_7, \\
E[T_9] &= 9 - 4q_2 - q_3 - q_4 - q_5 - q_9, \\
E[T_{10}] &= 10 - 4q_2 - 2q_3 - 2q_5 - q_{10}, \\
E[T_{11}] &= 11 - 4q_2 - 3q_3 - q_5 - q_6 - q_{11}, \\
E[T_{12}] &= 12 - 4q_2 - 4q_3 - 2q_6 - q_{12}, \\
E[T_{13}] &= 13 - 5q_2 - 3q_3 - q_4 - q_6 - q_{13}, \\
E[T_{14}] &= 14 - 6q_2 - 2q_3 - 2q_4 - 2q_7 - q_{14}, \\
E[T_{15}] &= 15 - 7q_2 - q_3 - 3q_4 - q_7 - q_8 - q_{15}.
\end{aligned}$$

It is interesting to note that the coefficients in the expressions for $E(T_k)$ add up to 1. The relative cost can be found by dividing $E[T_k]$ by the composite sample size, i.e., $\mathrm{RC}_k = E[T_k]/k$.

The optimal binary split retesting procedure starts with the largest possible composite sample size. To understand why this is so, consider combining two composite samples to form a larger composite sample. If the larger composite sample tests negative, then one test is saved. If the larger composite sample tests positive, then one of the original (smaller) composite samples is tested and the value of the other composite sample is calculated. This results in a total of two tests, namely one on the larger composite sample and the other on one of the smaller composite samples. That is to say, it takes the same number of tests to reach this point if the two composite samples were tested separately. In practice, the composite sample size will be limited by the number of subsamples that can be mixed or ratio of the detection limit to the action level.

2.3.3 Entropy-Based Retesting

This method appears to be inappropriate for continuous response variables. Recall that a positively testing composite results in the formation of a subcomposite of size about half of that of the original composite. If this subcomposite tests positive, then the remaining items in the original composite are returned to the pool of unclassified items. However, the total for these remaining items can be calculated, and it seems unreasonable not to use this information. If these items are not returned to the unclassified pool, then the resulting procedure reduces to the curtailed binary split retesting procedure.

2.4 Cost Analysis of Composite Sampling for Classification

2.4.1 Introduction

Sampling plans for environmental and public health monitoring often involve expensive laboratory methods for quantifying observations on individual sample units. United States Environmental Protection Agency (US EPA) and the regulated community spend an estimated $5 billion every year on collecting data for research, regulatory decision making, and regulatory compliance (US EPA, 1994). Johnson and Patil (2001) have carried out a cost analysis of composite sampling for classification. The cases of presence/absence and continuous measurements are considered. The general cost expression is using probability theory and the relative cost of composite sampling is derived in comparison with the conventional method of using individual sample measurements.

Cost analysis of composite sampling is not as easy as determining the expected number of measurements to be made. Composite sampling involves costs that do not arise in the conventional method of making measurements on individual sample units. For instance, forming a composite sample of several soil samples requires careful laboratory procedure of cleaning the containers and rinsing the solvent before forming every composite. Similarly, when composite sampling is used for sampling of fluids or gases, appropriate procedures have to be carefully implemented while forming composite samples. Also, archiving aliquots of individual samples for possible retesting involves storage and retrieval costs. Finally, any extra labor costs must be taken into account before concluding as to whether or not composite sampling will be truly cost-effective.

2.4.2 General Cost Expression

A general cost expression is derived by taking into consideration various cost components and different retesting strategies. The following notation is used in the derivation of the cost expression and relative cost of composite sampling:

m = number of individual sample units to be classified,
n = number of composite sample units,
k = number of individual sample units contributing to a single composite sample,
C_s = cost of acquisition of an individual sample unit,
C_a = cost of archiving an individual sample unit,
C_c = cost of forming a composite sample,
C_t = cost of testing a sample unit, either individual or composite,
Y_k = number of composites to be formed, each of size k,
T_k = number of tests for classifying k individual sample units in a composite.

Then the relative cost of composite sampling is given by

$$\mathrm{RC} = \{(C_s + C_a) + C_c E[Y_k] + C_t E[T_k]\}/(C_s + C_t).$$

The procedure to be followed for classifying m individual sample units by composite sampling with selective retesting is as follows:

1. Obtain m individual sample units.
2. Obtain an aliquot from every sample unit and archive the remaining material for possible retesting.
3. Form n composite samples, each consisting of aliquots from l individual sample units.
4. Analyze the n composite samples.
5. Classify a composite sample as "clean" or "contaminated." If a composite sample is "clean," then all individual samples contributing aliquots to that composite are classified as "clean." Otherwise, retesting must be undertaken to classify individual samples as "clean" or "contaminated."
6. Retest the archived sample units based on the result of Step 5. All the m individual sample units are classified.

2.4.3 *Effect of False Positives and False Negatives on Composite Sample Classification*

Testing mechanisms are subject to some degree of error. In the case of binary classification, an error is either a false-positive or a false-negative test. Let r_{n} and r_{p} denote the rates of these two errors, respectively. These rates are defined by the following probability statements:

$r_{\mathrm{p}} = \Pr(\text{positive test result} \mid \text{clean sample})$,
$r_{\mathrm{n}} = \Pr(\text{negative test result} \mid \text{contaminated sample})$.

Composite sampling with retesting reduces the overall false-positive error rate with a trade-off that the false-negative error rate can be magnified due to

compositing. The false-negative error rate may be controlled by retesting some of the negative testing composites, though this would increase the expected number of tests.

The interest is in $E[Y_k]$ and $E[T_k]$ for the specified values of C_s, C_a, C_c, C_t, and k. In situations where r_p and r_n are not negligible, the overall design false-positive error rate, d_p, *and* false-negative error rate, d_n, are also derived.

2.4.4 Presence/Absence Measurements

The measurement of a sample returns a binary response indicating presence or absence of the trait of interest in the tested sample. The Bernoulli model with parameter p is appropriate for this situation under the assumption that all m individual measurements are independent and identically distributed with probability p of testing positive.

2.4.4.1 Exhaustive Retesting

Exhaustive retesting, as proposed by Dorfman (1943), does not result in re-formation of composites and hence $E[Y_k] = 1$. The number of tests will be 1 (when the composite tests negative) or $k + 1$ (when the composite tests positive). Writing $q = 1 - p$, where p denotes the probability that the trait is present in an individual sample, the expected number of tests is given by

$$\begin{aligned} E[T_k] &= 1 \cdot q^k + (k+1) \cdot (1 - q^k) \\ &= 1 + k(1 - q^k). \end{aligned}$$

The overall design false-negative error rate is given by

$$\begin{aligned} d_\mathrm{n} &= r_\mathrm{n} + (1 - r_\mathrm{n}) r_\mathrm{n} \\ &= 2r_\mathrm{n} - r_\mathrm{n}^2. \end{aligned}$$

The overall design false-positive error rate is given by

$$\begin{aligned} d_\mathrm{p} &= r_\mathrm{p} \cdot [\Pr(\text{retest} | \text{at least one of } k-1 \text{ is positive}) \cdot (1 - q^k) \\ &\qquad + \Pr(\text{retest} | \text{all } k-1 \text{ are negative})\, q^{k-1}] \\ &= r_\mathrm{p} \cdot [(1 - r_\mathrm{n})(1 - q^{k-1} + r_\mathrm{p} \cdot q^{k-1}] \\ &= r_\mathrm{p} \cdot [1 - r_\mathrm{n} - q^{k-1}(1 - r_\mathrm{n} - r_\mathrm{p})]. \end{aligned}$$

$$\begin{aligned} \Pr(T_k = 1) &= r_\mathrm{n}(1 - q^k) + (1 - r_\mathrm{p})\, q^k \\ &= r_\mathrm{n} + q^k(1 - r_\mathrm{n} - r_\mathrm{p}), \end{aligned}$$

$$\begin{aligned}\Pr(T_k = k+1) &= 1 - \Pr(T_k = 1)\\ &= 1 - r_{\mathrm{n}} - q^k(1 - r_{\mathrm{n}} - r_{\mathrm{p}}).\end{aligned}$$

Therefore,

$$E[T_k] = k + 1 - k[r_{\mathrm{n}} + q^k(1 - r_{\mathrm{n}} - r_{\mathrm{p}})].$$

2.4.4.2 Sequential Retesting

Sterret (1957) proposed sequential retesting method as an improvement in the exhaustive retesting method. In this method,

$$\begin{aligned}E[Y_k] &= 1 + (k-2)p,\\ E[T_k] &= 2k - (k-3)q - q^2 - (1 - q^{k+1})/(1-q), \quad k = 2, 3, \ldots.\end{aligned}$$

If r_{n} and r_{p} are not negligible, then

$$\begin{aligned}E[Y_k] &= 1 + (k-2)[r_{\mathrm{p}}q + (1 - r_{\mathrm{n}})p],\\ E[T_k] &= 2k - (k-3)[r_{\mathrm{n}}p + (1 - r_{\mathrm{p}})q] - [r_{\mathrm{n}} + (1 - r_{\mathrm{p}})q]^2\\ &\quad - \{1 - [r_{\mathrm{n}}p + (1 - r_{\mathrm{p}})q]^{k+1}\}/\{r_{\mathrm{p}}\, q + (1 - r_{\mathrm{n}})p\}, \quad k = 3, 4, \ldots.\end{aligned}$$

2.4.4.3 Binary Split Retesting

The binary split retesting method of Gill and Gottlieb (1974) entails a recurrence relation for the expected number of composites to be formed and for expected number of tests to be carried out.

$$\begin{aligned}E[Y_k] &= E[Y_{k1}] + E[Y_{k2}] + 1 - 2q^k, \quad k = 4, 5, \ldots,\\ E[T_k] &= E[T_{k1}] + E[T_{k2}] + 1 - 2q^k, \quad k = 2, 3, \ldots,\end{aligned}$$

where $k_1 = k_2 = k/2$ if k is even and $k_1 = (k-1)/2$ and $k_2 = (k+1)/2$ if k is odd.

The design false-positive rate is given by

$$\begin{aligned}d_{\mathrm{p}} =&[(1 - r_{\mathrm{n}})(1 - q^k) + r_{\mathrm{p}}q^k] \times [(1 - r_{\mathrm{n}})(1 - q^{k1}) + r_{\mathrm{p}}q^{k1}]\\ &\times [(1 - r_{\mathrm{n}})(1 - q^{k11}) + r_{\mathrm{p}}q^{k11}] \times \cdots \times r_{\mathrm{p}}.\end{aligned}$$

Here, k_1 and k_2 denote sizes of subcomposites of the composite of size k. These are in turn split into subcomposites of sizes k_{11} and k_{12}, k_{21} and k_{22}, respectively, and so on. The design false-negative error rate d_{n} for an initial composite of size k is given by

$$\begin{aligned} d_{\mathrm{n}}(k) &= r_{\mathrm{n}} - r_{\mathrm{n}}^2 + d_{\mathrm{n}}(k/2) \quad \text{if k is even,} \\ &= r_{\mathrm{n}} - r_{\mathrm{n}}^2 + d_{\mathrm{n}}((k+1)/2) \quad \text{if k is odd.} \end{aligned}$$

Here, the argument of d_{n} indicates the composite sample size.

The expected number of composites and tests is, respectively, given by

$$\begin{aligned} E[Y_k] &= E[Y_{k1}] + E[Y_{k2}] + 1 - 2[(1 - r_{\mathrm{p}})q + r_{\mathrm{n}}p]^k, \quad k = 4, 5, \ldots, \\ E[T_k] &= E[T_{k1}] + E[T_{k2}] + 1 - 2[(1 - r_{\mathrm{p}})q + r_{\mathrm{n}}p]^k, \quad k = 2, 3, \ldots. \end{aligned}$$

These equations can be solved recursively for the appropriate value of k.

2.4.5 Continuous Measurements

Measurement on a continuous random variable results in classifying a sample unit as negative if its measured value is less than some numerical criterion. However, a size k composite cannot be classified as negative using the same criterion because it would not imply that every individual sample unit contributing to the composite is negative. For this purpose, if c denotes the criterion for an individual sample, the numerical criterion for a composite sample of size k is c/k.

If $X_1, X_2, \ldots, X_k$ denote individual sample values and Y denotes the composite value, then the probability that a composite sample tests negative is

$$\begin{aligned} q_k &= \Pr[X_1 + X_2 + \cdots + X_k < c] \\ &= \Pr[X_1 + \cdots + X_k < c | X_1 < c, \ldots, X_{\mathrm{k}} < \mathrm{c}] \times \Pr[X_1 < c, \ldots, X_k < c] \\ &= \Pr[X_1 + \cdots + X_k < c | X_1 < c, \ldots, X_k < c] \times q^k, \end{aligned}$$

where $q = \Pr[X_1 < c]$ and $X_1, \ldots, X_k$ are independent and identically distributed.

Since $q_k < q^k$, the relative cost for measuring a continuous variable has a lower bound representing the relative cost for measuring a presence/absence variable. The expected number of composites and tests heavily depends on the probability distribution of the individual sample values. Approximations can be obtained through expressions in case of presence/absence measurement. However, such an approximation can have a dual impact on the result.

On the one hand, this leads to over-optimistic results due to an upper bound on the probability of a composite sample testing negative. On the other hand, the presence/absence expressions are based on independence among individual sample measurements achieved through random formation of composite samples. When individual sample values are autocorrelated, better strategies of forming composites can be developed to improve the performance of composite sampling by reducing the relative cost. In this case, there is no comparison with the case of presence/absence measurements.

http://www.springer.com/978-1-4419-7627-7

Composite Sampling
A Novel Method to Accomplish Observational Economy in Environmental Studies

Patil, G.P.; Gore, S.D.; Taillie, C.

2011, XIII, 275 p., Hardcover
ISBN: 978-1-4419-7627-7