Computational Pool-Testing with Retesting Strategy

Cox Lwaka Tamba *clwaka@yahoo.com*
Faculty of Science/Department of Mathematics/Division of Statistics
Egerton University
P.o. Box 536-20115, Egerton, Kenya

Jesse Wachira Mwangi *jdmwangi@egerton.ac.ke*
Faculty of Science/Department of Mathematics/Division of Statistics
Egerton University
P.o. Box 536-20115, Egerton, Kenya

.

Abstract

Pool testing is a cost effective procedure for identifying defective items in a large population. It also improves the efficiency of the testing procedure when imperfect tests are employed. This study develops computational pool-testing strategy based on a proposed pool testing with re-testing strategy. Statistical moments based on this applied design have been generated. With advent of digital computers in 1980's, pool-testing with re-testing strategy under discussion is handled in the context of computational statistics. From our study, it has been established that re-testing reduces misclassifications significantly as compared to Dorfman procedure although re-testing comes with a cost i.e. increase in the number of tests. Re-testing considered improves the sensitivity and specificity of the testing scheme.

Keywords: Pool, Pooling, Re-Specificity, Sensitivity, Tests, Misclassifications.

1. INTRODUCTION

The idea of pool testing was initiated by Dorfman (1943) during World War II as an economical method of testing blood samples of army inductees in order to detect the presence of infection. Pooling procedures involve putting together individuals to form a group/pool and then testing the group rather than testing each individual for evidence of a characteristic of interest. A negative reading indicates that the group contains no defective items and a positive reading indicates the presence of at least one defective individual in the group. There are two objectives of pool testing: classification of the units of a population as either defective or non-defective (Dorfman, 1943) and estimation of the prevalence of a disease in a population (Sobel and Elashoff, 1975). Pooling procedures have proved to reduce the cost of testing when the prevalence rate is low. This is because if a pool tests negative, it implies all its constituent members are non-defective and hence it is not necessary to test each member of the pool. A procedure of classifying the population into defective and non-defective when each unit i of the population has a different probability p_i of being defective (which is called a generalized binomial group test) problem has been studied (Hwang, 1975). The generalized binomial group test problem reduces to a binary pool testing problem which is the Dorfman, (1943) procedure when all the units have the same probability p of being defective. Hwang (1976) has also considered a pool testing model in the presence of dilution effect. Dilution effect in this case refers to a situation where a pool which contains a few defective items may be misidentified as containing no such items, especially when the size of the pool is large.

Johnson *et al.* (1992) has studied the cost effectiveness of pooling algorithm for the objective of identifying individuals with a characteristic of interest using hierarchical procedures. In this procedure, each pool that test positive is divided into two equal groups, which are tested, groups that tested positive are further subdivided and tested and so on. This work has been extended by considering pooling algorithms when there are errors and showed that some of these algorithms can reduce the error rates of the screening procedures (the false positives and false negatives) compared to individual testing (Litvak *et al.,* 1994). Computational statistics has been used in pool testing to compute the statistical measures when perfect and imperfect tests are used has been considered (Nyongesa and Syaywa, 2011; Nyongesa and Syaywa, 2010; Tamba et al., 2012).

Pool testing has vast applications (Sobel and Groll, 1966). It has been applied industries, and recently it has been applied in screening the population for the presence of HIV antibody (Kline *et al*., 1989 and Manzon *et al*., 1992). Pool testing has been used in screening HIV antibody to help curb the further spread of the virus (Litvak *et al*., 1994).

In this study, we consider the computation of statistical measures based on a pool testing with re-testing strategy via computer package MATLAB. The tests used in this procedure are assumed imperfect i.e. the specificity and sensitivity are less than 100%. The rest of the paper is arranged as follows: Section 2 formulates the problem while the pool-testing with re-testing strategy is discussed in Section 3. Section 4 provides the number of tests and moments while the results in this design are provided in Section 5. Misclassifications in the proposed testing design are discussed in Section 6. Section 7 provides the discussion and conclusion to this study.

2. PROBLEM FORMULATION

Consider a population of size *N* pooled into *n* pools each of size *k*. Each pool is subjected to an initial test. We re-test each pool irrespective of whether it tests positive or negative on the initial test. Pools that test positive on re-testing of pools that initially tested positive and negative, their constituent members are tested individually. We wish to establish the number of defective individuals in the population *N*. For efficiency and cost effectiveness, pool the population *N* into *n* independent pools each of equal sizes say *k*. The procedure is described diagrammatically below.

Pools

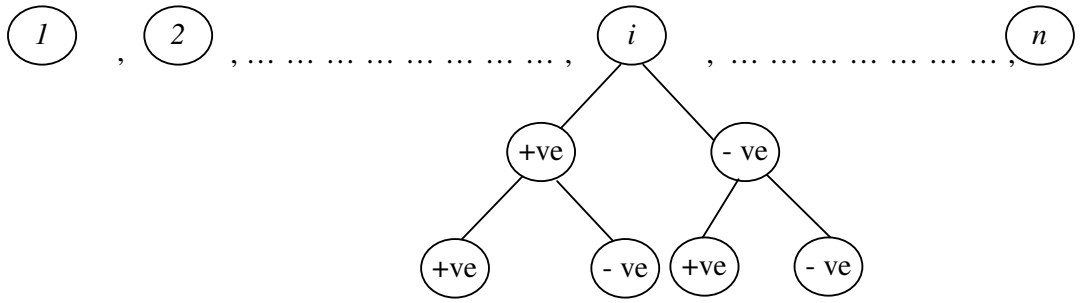

Figure 1: Diagrammatic description of the Pool Testing with Re-testing Strategy

The figure shows the *n* constructed pools and the test result on the i^{th} pool, for i=1, 2, …, *n*. The result on the test is binary: positive or negative. The analysis in this study will require the following indicator functions:

Let

$$T_i = \begin{cases} 1; & if\ the\ i^{th}\ pool\ tests\ positive\ on\ the\ test\ kit \\ 0; & otherwise \end{cases}$$

$$T'_i = \begin{cases} 1; & if\ the\ i^{th}\ pool\ test\ is\ positive\ on\ the\ re-test\ on\ the\ test-kit \\ 0; & otherwise \end{cases}$$

$$D_i = \begin{cases} 1; & if\ the\ i^{th}\ pool\ is\ positive \\ 0; & otherwise \end{cases}$$

$$T_{ij} = \begin{cases} 1; & if\ the\ j^{th}\ individual\ in\ an\ i^{th}\ pool\ tests\ positive\ on\ the\ test\ kit \\ 0; & otherwise \end{cases}$$

and

$$\zeta_{ij} = \begin{cases} 1; & if\ the\ j^{th}\ individual\ in\ the\ i^{th}\ pool\ is\ positive\ with\ probability\ p \\ 0; & otherwise \end{cases}$$

The indicator functions provided above are essential in the subsequent developments. The constituent members of the i^{th} pool will be represented by $(\zeta_{i1}, \zeta_{i2}, ..., \zeta_{ij}, ..., \zeta_{ik})$ or simply $\{\zeta_{ij}\}_{j=1}^{k}$. Clearly,

$$\Pr(D_i = 0) = \Pr(\zeta_{i1} = 0, \zeta_{i2} = 0, ..., \zeta_{ij} = 0, ..., \zeta_{ik} = 0) \quad (1)$$

by definition. For analysis purposes, we shall assume that the constituent member of a pool act independently of each other, hence

$$\Pr(D_i = 0) = (1-p)^k \quad (2)$$

where p is the prevalence rate.

3. POOL TESTING WITH RE-TESTING STRATEGY

Let *N* be a universal set and ξ be a $\delta-$ field on N. Let X_{11}, X_{12}, X_{21} and X_{22} be random variables defined on N. Now subdivide N into n partitions representing pools each of size k. We perform a test on each pool as discussed above. Basically, let X_{11} be the number of pools that test positive on re-testing initial declared positive pools, X_{12} be the number of pools that test negative on re-testing initial declared positive pools, X_{21} be the number of pools that test positive on re-testing of pools that initially tested negative test and X_{22} be the number of pools that test negative on retesting initially declared negative pools. Let p be the probability measure on ξ such that an individual is positive (prevalence rate). We derive new set functions π_1, π_2, π_3 and π_4 on ξ, where $\pi_1 = \Pr(T_i = 1, T'_i = 1)$, $\pi_2 = \Pr(T_i = 1, T'_i = 0)$, $\pi_3 = \Pr(T_i = 0, T'_i = 1)$ and $\pi_4 = \Pr(T_i = 0, T'_i = 0)$. To obtain these probabilities, we require the application of the law of total probability (Ross, 1997). First,

$$\pi_1 = (1-p)^k(1-\phi) + [1-(1-p)^k]\eta^2. \quad (3)$$

where η is the sensitivity of the test kit and ϕ the specificity of the test kit. By sensitivity, we mean the probability of correctly classifying a defective pool or defective individual while ϕ is specificity of the test kits and by specificity here means the probability of correctly classifying a non-defective pool or non-defective individual. Ideally, we have introduced the error element in our model. The error component will be assumed to be based on the manufacturers' specifications and will remain constant in the entire experiment. That is, sensitivity and specificity will remain constant at group level and individual level. Clearly $p \in [0,1]$ and so $1-\phi \le \pi_1 \le \eta^2$ which implies that π_1 is a continuous function bounded below by $1-\phi$ and above by η^2. Next,

$$\pi_2 = \phi(1-\phi)(1-p)^k + \eta(1-\eta)[1-(1-p)^k]. \quad (4)$$

The probability $\Pr(T_i = 0, T'_i = 1)$ is,

$$\pi_3 = \phi(1-\phi)(1-p)^k + \eta(1-\eta)\left[1-(1-p)^k\right] \quad (5)$$

and the probability $\Pr(T_i = 0, T'_i = 0)$ has been derived as,

$$\pi_4 = \phi^2(1-p)^k + (1-\beta)^2\left[1-(1-p)^k\right] \quad (6)$$

The probabilities π_1, π_2, π_3 and π_4 will enable us to compute the joint probability distribution of X_{11}, X_{12}, X_{21} and X_{22}. The joint probability density function of X_{11}, X_{12}, X_{21} and X_{22} is a multinomial probability density i.e.,

$$f_{X_{11},X_{12},X_{21},X_{22}}(x_2, x_{11}, x_{12}) = \binom{n}{x_{11}\ x_{12}\ x_{12}\ x_{22}} \pi_1^{x_{11}} \pi_2^{x_{12}} \pi_3^{x_{21}} \left(1-\pi_1-\pi_2-\pi_3\right)^{n-x_{11}-x_{12}-x_{12}}. \quad (7)$$

Utilizing the probabilities π_1, π_2, π_3, π_4 and Model (7) above, we propose a computer intensive pool-testing with re-testing strategy. With the advent of the digital computers in early 1980's,

computational statistics has evolved (Martinez and Martinez, 2002; L'Ecuyer, 2004). In a similar format we wish to develop a computational pool-testing with re-testing model. The next section discusses the moments of the number of tests in the testing scheme.

4. THE NUMBER OF TESTS AND MOMENTS

Model (7) is of major interest in this study. The overall number of tests in this design is

$$Z = 1 + 2n + kX_{11} + kX_{21}, \tag{8}$$

where *n* is the number of pools and *k* is the pool size. It then follows that,

$$E[Z] = 1 + 2n + kn[\pi_1 + \pi_3]. \tag{9}$$

In the field experiments, the sensitivity and specificity of the test kits are normally provided by the manufacturers' specifications. Therefore, for given η and ϕ, we can compute (9) and similarly, the variance of the number of test is

$$Var(Z) = k^2\left[n\pi_1(1-\pi_1) + n\pi_3(1-\pi_3) - 2n\pi_1\pi_3\right] \tag{10}$$

from which, the standard deviation is given by $\sqrt{k^2\left[n\pi_1(1-\pi_1) + n\pi_3(1-\pi_3) - 2n\pi_1\pi_3\right]}$.

We shall utilize Equations (9) and (10) to generate the mean, standard deviation in the proposed pool-testing with re-testing strategy. Next, we consider misclassification arising from this testing design.

5. MISCLASSIFICATIONS

Note that since we allowed testing with errors in our design as is the case real life problem i.e. the test kits in use are not 100% perfect, two possible misclassifications can arise in practice; false- positive and false- negatives. A false- positive refers to a non- defective item being classified as defective whereas a false- negative means that a defective item is classified as non-defective. The probability of correctly classifying a defective individual is referred to as sensitivity. The sensitivity of the testing procedure is derived as,

$$\begin{aligned}\text{Sensitivity} &= \Pr(T_i = 1, T'_i = 1, T_{ij} = 1 \mid \delta_{ij} = 1) + \Pr(T_i = 0, T'_i = 1, T_{ij} = 1 \mid \delta_{ij} = 1) \\ &= \eta^2. \end{aligned} \tag{12}$$

The probability of false positives arising from this model is

$$f_p = 1 - \eta^2. \tag{13}$$

We know that η<1 in practice, this implies that $\eta^2 < \eta$, hence pool-testing scheme lowers the sensitivity in general. The sensitivity of this procedure is the same as that of the pool testing without re-testing. (c.f Tamba et al., 2012). We derive the probability of correctly classifying a non-defective individual herein referred as specificity of the testing procedure.

$$\begin{aligned} Specificity &= \Pr(T_i = 1, T'_i = 0 \mid \delta_{ij} = 0) + \Pr(T_i = 0, T'_i = 0 \mid \delta_{ij} = 0) + \\ &\quad \Pr(T_i = 1, T'_i = 1, T_{ij} = 0 \mid \delta_{ij} = 0) + \Pr(T_i = 0, T'_i = 1, T_{ij} = 0 \mid \delta_{ij} = 0) \\ &= \left[2\phi - \phi^2\right]\left[(1-p)^{k-1}\right] + \left[(1 - \eta + \eta\phi) + \right]\left[1 - (1-p)^{k-1}\right] \end{aligned} \tag{14}$$

One minus the specificity of the testing scheme yields the probability of false negative as

$$f_n = 1 - \left[2\phi - \phi^2\right]\left[(1-p)^{k-1}\right] + \left[(1 - \eta + \eta\phi) + \right]\left[1 - (1-p)^{k-1}\right]. \tag{15}$$

To investigate the performance of this design we shall utilize Equation (13) and (15), in computing the false-positive and false-negative. The pool size in (14) appears to be $k-1$ since the j^{th} individual is known to be non- negative and we have employed sampling without replacement.

6. RESULTS

To this end, we have presented formulas that can be used to compute the central moments of the number of tests in group testing with re-testing scheme. We illustrate the procedure by computing the central moment measures for various sensitivity and specificity. In the tables we have used the word total testing cost to imply the overall cost of performing the procedure.

Characteristics	p=0.01		p=0.05		p=0.1	
	μ	σ	μ	σ	μ	σ
Number of defectives	2.0010	0.9232	6.0020	2.2183	10.8000	3.0205
Number of defective groups	1.0060	0.9172	4.0700	1.4931	6.5840	1.4860
Number of group tests	21.0000	-	21.0000	-	21.0000	-
Number of individual tests	10.0600	9.1720	40.7000	14.9310	65.8400	14.8600
Total number of tests	31.6600	9.1720	61.7000	14.9310	86.8400	14.8600
Total testing cost	31.6600	9.1720	61.7000	14.9310	86.8400	14.8600
Percentage savings	68.3400	9.1720	38.3000	14.9310	13.1600	14.8600

Table 1: Various characteristics for pool testing with re-testing strategy with 1000 runs, *N* =100, k=10, $\eta = \phi = 99$%

Characteristics	p=0.01		p=0.05		p=0.1	
	μ	σ	μ	σ	μ	σ
Number of defectives	9.7590	2.1833	29.5240	4.6742	54.0840	6.2065
Number of defective groups	4.7710	1.8682	15.9250	2.4625	21.7910	1.4881
Number of group tests	51.0000	-	51.0000	-	51.0000	-
Number of individual tests	95.5420	37.3640	318.5000	49.2500	435.8200	29.7620
Total number of tests	146.4200	37.3640	369.5000	49.2500	486.8200	29.7620
Total testing cost	29.2840	37.3640	73.9000	49.2500	97.3640	29.7620
Percentage savings	70.7160	37.3640	26.1000	49.2500	2.6360	29.7620

Table 2: Various characteristics for pool testing with re-testing strategy with 1000 runs, *N* =500, k=20, $\eta = \phi = 99$%

Characteristics	p=0.01		p=0.05		p=0.1	
	μ	σ	μ	σ	μ	σ
Number of defectives	6.0190	1.0691	9.4710	2.0904	13.8480	2.7201
Number of defective groups	1.3370	0.7994	4.0610	1.4019	6.3180	1.3284
Number of group tests	21.0000	-	21.0000	-	21.0000	-
Number of individual tests	13.3700	7.9944	40.6100	14.0190	63.1800	13.2840
Total number of tests	34.3700	7.9944	61.6100	14.0190	84.1800	13.2840
Total testing cost	34.3700	7.9944	61.6100	14.0190	84.1800	13.2840
Percentage savings	65.6300	7.9944	38.3900	14.0190	15.8200	13.2840

Table 3: Various characteristics for pool testing with re-testing strategy with 1000 runs, *N* =100, k=10, $\eta = \phi = 95$%

Characteristics	p=0.01		p=0.05		p=0.1	
	μ	σ	μ	σ	μ	σ
Number of defectives	29.4820	2.2064	47.3720	4.27323	70.2500	6.1414
Number of defective groups	5.2730	1.8166	15.5170	2.2012	21.0340	1.4323
Number of group tests	51.0000	-	51.0000	-	51.0000	-
Number of individual tests	105.4600	36.3320	310.3400	44.0240	420.6800	28.6460
Total number of tests	156.4600	36.3320	361.3400	44.0240	471.6800	28.6460
Total testing cost	31.2920	36.3320	72.2680	44.0240	94.3360	28.6460
Percentage savings	68.7080	36.3320	27.7320	44.0240	5.6640	28.6460

Table 4: Various characteristics for pool testing with re-testing strategy with 1000 runs, *N* =500, k=20, $\eta = \phi = 95\%$

Probability, p	*N* =100, k=10		*N* =500, k=20		*N* =1000, k=20	
	μ	σ	μ	σ	μ	σ
0.01	0.5652	0.7142	2.8714	1.6098	5.7724	2.2825
0.02	0.6572	0.7701	3.3189	1.7307	6.6502	2.4499
0.03	0.7399	0.8172	3.7829	1.8477	7.4711	2.5967
0.04	0.8429	0.8722	4.2054	1.9482	8.4221	2.7570
0.05	0.9141	0.9083	4.6254	2.0431	9.2276	2.8838
0.1	1.3625	1.1089	6.8062	2.4784	13.6571	3.5105
0.15	1.7953	1.2729	9.0324	2.8551	18.0663	4.0379

Table 5: Number of false positives in the pool testing strategy for different pool sizes

$\eta = \phi = 99\%$

	N =100, k=10		*N* =500, k=20		*N* =1000, k=20	
	μ	σ	μ	σ	μ	σ
0.01	0.0388	0.1949	0.1964	0.4387	0.3941	0.6215
0.02	0.0593	0.2410	0.2940	0.5368	0.5912	0.7612
0.03	0.0767	0.2741	0.3915	0.6194	0.7825	0.8757
0.04	0.0987	0.3110	0.4902	0.6932	0.9809	0.9805
0.05	0.1171	0.3387	0.5805	0.7543	1.1705	1.0711
0.1	0.2147	0.4587	1.0749	1.0264	2.1483	1.4510
0.15	0.3143	0.5552	1.5672	1.2393	3.1239	1.7498

Table 6: Number of false positives in the pool testing strategy for different pool sizes

$\eta = \phi = 95\%$

Probability, p	N=100, k=10		N=500, k=20		N=1000, k=20	
	μ	σ	μ	σ	μ	σ
0.01	0.0929	0.3046	0.8839	0.9393	1.7676	1.3283
0.02	0.1679	0.4093	1.5640	1.2486	3.1289	1.7660
0.03	0.2356	0.4848	2.1158	1.4514	4.2329	2.0529
0.04	0.2957	0.5430	2.5609	1.5960	5.1248	2.2577
0.05	0.3507	0.5911	2.9186	1.7031	5.8332	2.4077
0.1	0.5438	0.7352	3.8264	1.9477	7.6527	2.7545
0.15	0.6447	0.7999	3.9813	1.9859	7.9683	2.8094

Table 7: Number of false negatives in the pool testing strategy for different pool sizes

$$\eta = \phi = 99\%$$

Probability, p	N=100, k=10		N=500, k=20		N=1000, k=20	
	μ	σ	μ	σ	μ	σ
0.01	0.6007	0.7726	4.8543	2.1918	9.7127	3.1004
0.02	0.9303	0.9597	7.8444	2.7771	15.6995	3.9287
0.03	1.2270	1.1003	10.2817	3.1709	20.5605	4.4836
0.04	1.4941	1.2123	12.2386	3.4512	24.4819	4.8812
0.05	1.7317	1.3033	13.8131	3.6594	27.6312	5.1757
0.1	2.5874	1.5842	17.8083	4.1317	35.6026	5.8419
0.15	3.0186	1.7049	18.5098	4.2034	39.0662	5.9485

Table 8: Number of false negatives in the pool testing strategy for different pool sizes

$$\eta = \phi = 95\%$$

Remark : In all the above tables we have; $\mu = mean, \sigma = s\tan dard\ deviation$

In the next section, we provide the discussion of our findings and the conclusion to the study.

7. DISCUSSION AND CONCLUSION

This study has presented a computational pool testing strategy with re-testing. It has been shown from the results; Tables 1, 2, 3 and 4 that when the pool size and prevalence rate are small, significant savings are realized. This is an empirical result since pool testing is only feasible when the prevalence rate is small otherwise individual testing is preferred. Similarly large pools are prone to increase the dilution effect and hence increase the misclassifications. It has been established that re-testing pools increases the cost of testing however, the misclassifications significantly reduce as compared to the Dorfman procedure when imperfect tests are used (Tamba *et al.*, 2011). The results in Tables 5, 6, 7 and 8 show that the higher the efficiency of the tests, the lower the misclassifications. This implies that pool testing should be carried out when specificity and sensitivity of the testing procedure are high. It has also been noted that this re-testing strategy improves the specificity and sensitivity of the testing procedure. Misclassifications are high when the prevalence rate is high and the efficiency of the test kits is low.

8. REFERENCES

1. C. L Tamba, K. L. Nyongesa, J. W. Mwangi, (2012). “Computational Pool-Testing Strategy”. Egerton University Journal, 11:51-56.

2. E. Litvak, X. M. Tu, and M. Pagano, (1994). “Screening for the presence of a disease by pooling sera samples.” Journal of the America statistical Association, **89**, 424-434.

3. F.K. Hwang, (1975). “A Generalized Binomial Group Testing Problem”. Journal of the American Statistical Association, **70**,923- 926.

4. F.K. Hwang, (1976). “Group testing with a dilution effect”. Biometrika **63**, 611-613.

5. L. K. Nyongesa and J. P. Syaywa, (2010). “Group Testing with Test Errors Made Easier.” International Journal of Computational Statistics. Volume (1): Issue (1).

6. L. K. Nyongesa and J. P. Syaywa, (2011). “Block Testing Strategy with Imperfect Tests and its Improved Efficient Testing Model for Donor Blood.” Communication in Statistics-Computational Statistics. (Submitted).

7. L. K. Nyongesa, (2005). “Hierarchical Screening with Retesting in a low Prevalence Population.” The Indian Journal of Statistics.**66,** 779-790.

8. L. K. Nyongesa, (2010). “Dual Estimation of Prevalence and Disease Incidence in Pool-Testing Strategy.” Communication in Statistics Theory and Method. Vol. (1): Issue (1).

9. M. Sobel, and P.A. Groll, (1966). “Binomial Group-Testing with an Unknown Proportion of Defectives.” American Statistical Association and American Society for Quality, **8**,631-656.

10. M. Sobel, and, R.M. Elashoff, (1975). “Group-testing with a new goal, Estimation.” Biometrika**, 62,** 181-193.

11. N. L. Johnson, S. Kotz, and, X Wu, (1992). “Inspection errors for attributes in quality control.” London; Chapman and Hall.

12. O.T. Monzon, F.J.E Palalin, E. Dimaal, , A.M. Balis, C. Samson, and S. Mitchel, (1992). “Relevance of antibody content and test format in HIV testing of pooled sera.” AIDS, **6**, 43-48.

13. P. L’Ecuyer,, (2004). Uniform Random Number Generation. Handbook of Computational Statistics. Springer-Verlag, Berlin, pp. 35–70.

14. R. B. Hunt, L. L. Ronald, M. R. Jonathan (2004). A Guide to MATLAB for Beginners and Experienced Users. Cambridge University Press, pp.101-119.

15. R. Dorfman, (1943). “The detection of defective members of large population”. Annals of Mathematical Statistics **14**, 436-440.

16. R.L. Kline, T. Bothus, R. Brookmeyer, S. Zeyer,, and T. Quinn, (1989). “Evaluation of Human Immunodeficiency Virus seroprevalence in population surveys using pooled sera.” Journal of clinical microbiology, **27**, 1449-1452.

17. S.M. Ross, (1997). Introduction to Probability Models. 6th Edition, Academic Press, pp 4-7.

18. W.L. Martinez,and A.L. Martinez,, (2002). Computational Statistics Handbook with MATLAB. Chapman & Hall/CRC, pp. 96-115.