PROBLEM CORNER

Md S. Warasi
Department of Mathematics and Statistics
Radford University, VA 24142
email: msarker@radford.edu

Problem 1

To screen blood donors for HIV, the American Red Cross often implements pool testing, where pools are formed by compositing a set of individual donations and then the pooled samples are tested for the presence or absence of HIV; see Figure 1. A pool is positive when at least one individual in the pool has disease; however, a pool is negative when all individuals in the pool are free of disease. Unfortunately, the assay being used for diagnosis is subject to errors. When a positive pool is tested, there is a 97% probability that the test result is positive (a correct result). When a negative pool is tested, there is a 98% probability that the test result is negative (also a correct result). Assume that the individuals are independent and have an identical probability of 1% to be HIV positive. Also, assume that the test accuracy does not depend on the pool size. Suppose a pool comprised of 3 individuals is tested for HIV.

a. What is the probability that the pool tests positive?

b. Write an algorithm to approximate the probability in 1(a) by simulation.

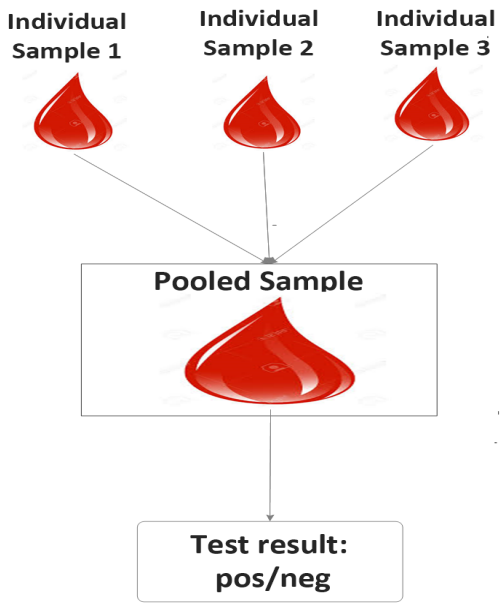

Figure 1: Pool testing to screen blood donors for HIV.

Solution:

a. Let A be the event that a pool is truly positive so that the complement of A, denoted by A^C, refers to the event that a pool is truly negative. Note that a pool is negative when all individuals in the pool are negative. Thus, for a pool of 3 independent individuals, $P(A^C) = 0.99 \times 0.99 \times 0.99 = 0.9703$, and $P(A) = 1 - P(A^C) = 0.0297$.

Suppose B is the event that a pool tests positive. We need to calculate $P(B)$. Possible partitions of the sample space are shown in Figure 2, where the event B is highlighted in pink. It is easy to observe that $P(B) = P(B \text{ and } A) + P(B \text{ and } A^C) = 0.0482$. The tree diagram in Figure 3 depicts the calculation of $P(B \text{ and } A)$ and $P(B \text{ and } A^C)$.

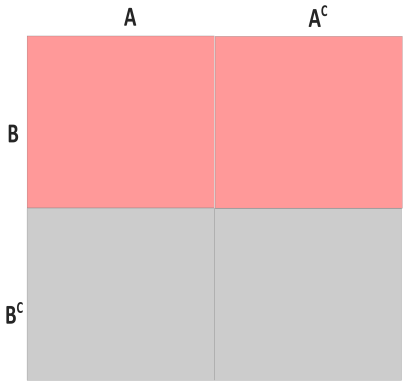

Figure 2: Partitions of the sample space.

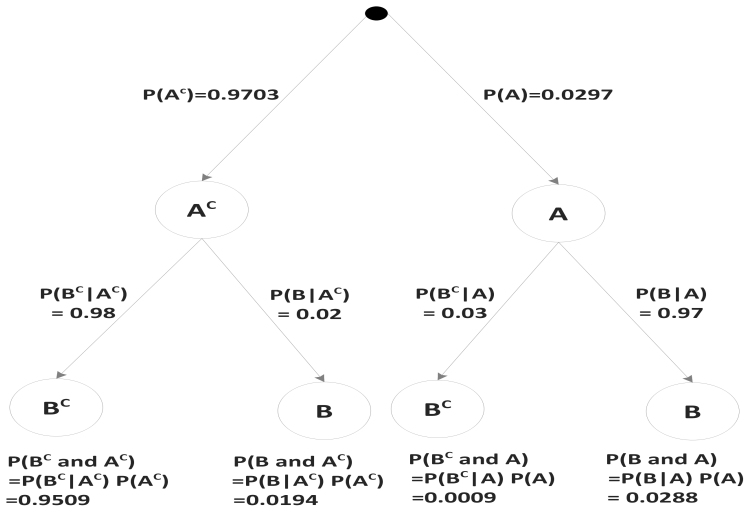

Figure 3: Tree diagram to calculate the probability in 1(a) that a pool is diagnosed as positive.

b. To approximate $P(B)$, a sample of pool testing results are simulated. When sample size is sufficiently large (e.g., 10,000), the sample mean will be a good approximate for $P(B)$. The flowchart in Figure 4 describes the steps to simulate one pool testing result. The steps can be repeated to obtain a large sample using any statistical software including `R`, `Minitab`, and `Matlab`. An example of R codes is shown below.

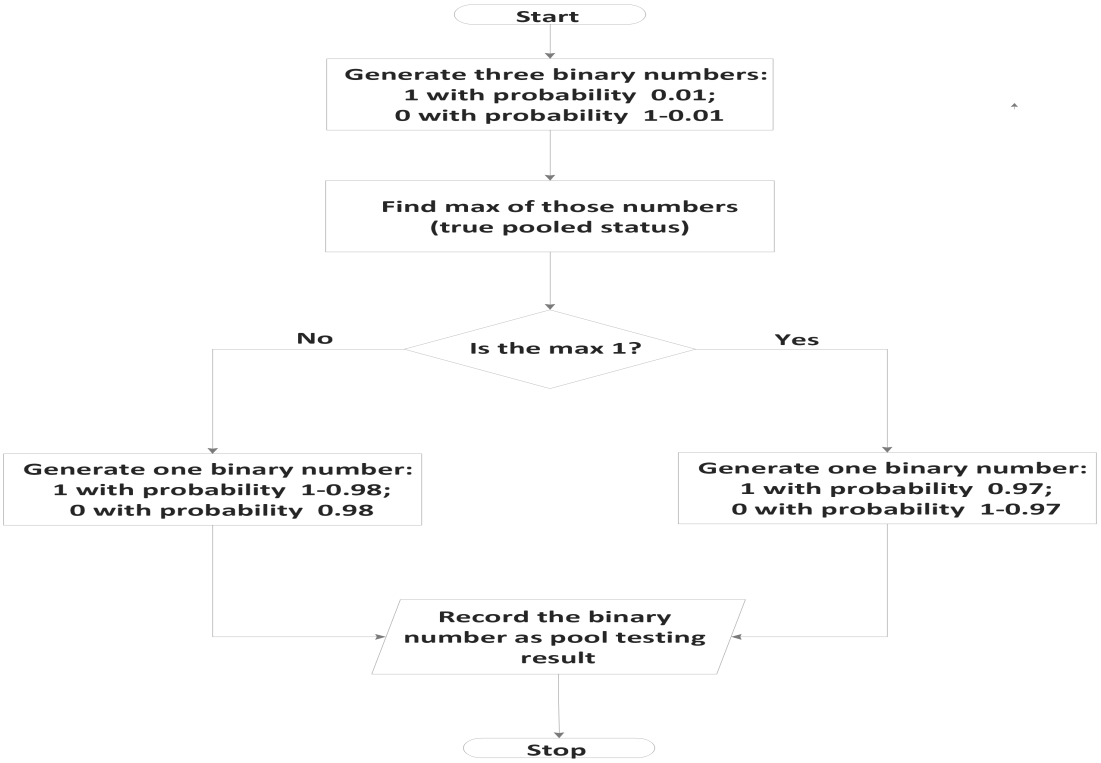

Figure 4: Flowchart to approximate the probability in 1(a) by simulation.

Note that accuracy of the approximation depends on the simulated data sample size. The law of large numbers states that sample mean converges to population mean as the sample size increases; thus, with very large sample size, the true value of $P(B)$ and its estimate will be nearly equal.

To illustrate how sample size affects the accuracy, we perform a simulation study with sample size $N = 100, 1000, 10000$. For each N, the probability $P(B)$ is estimated 2000 times. The mean and standard deviation of these estimates are shown in Table 1. The mean estimate with each sample size is close to the true value. However, the estimates with larger sample are more precise (less variable). The estimates are also presented in Figure 5, where the true value, $P(B) = 0.0482$, is shown by the horizontal line for comparison. The boxplots in Figure 5 reiterate the findings discussed above. In practice, exact calculation of some probabilities or integrations can be very hard or impossible. However, implementing the approximation technique can simplify the problem dramatically and work well for nearly all problems.

Table 1: Mean and standard deviation of 2000 estimates of $P(B)$ with sample size N.

	$N = 100$	$N = 1000$	$N = 1000$
Mean	0.0477	0.0481	0.0483
Standard deviation	0.0214	0.0068	0.0022

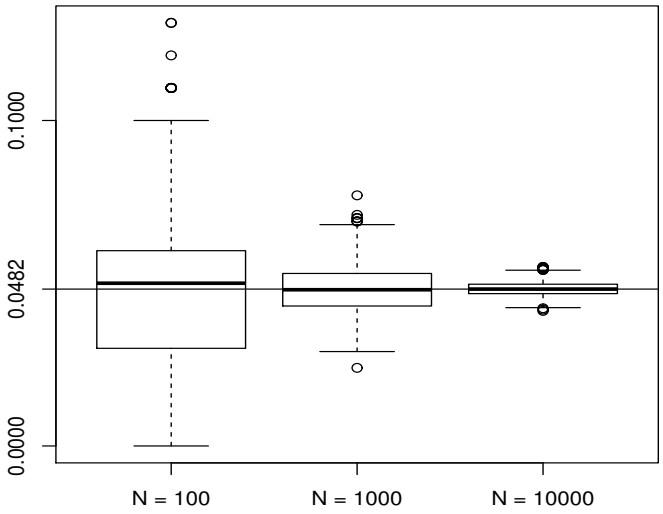

Figure 5: Boxplots of 2000 estimates of $P(B)$ with sample size N.

R codes to approximate $P(B)$:

```
Z <- NULL  # Note: Z is the pool testing result
for(j in 1:10000){
   Y <- rbinom(3, 1, .01)
   T <- ifelse(max(Y)==1, 1, 0)
   prob <- ifelse(T==1, 0.97, 1-0.98)
   Z[j] <- rbinom(1, 1, prob)
}
# An approximate of P(B):
mean(Z)
```

Note: R is a free software and widely used in research and applications. R can be downloaded from the link `https://www.r-project.org`.

Problem 2

In statistics, maximum likelihood is a procedure of estimating the parameters of a probabilistic model. In the context of pool testing, the maximum likelihood technique is used to estimate individual-level disease prevalence using data observed from pools; see Problem 1 for more details about pool testing.

Consider a pool testing application, where μ denotes the probability that an individual has HIV. Suppose J pools, each of which is comprised of n individuals, are tested for HIV. Let z_j, for $j = 1, 2, ..., J$, denote testing responses, where $z_j = 1$ if a pool tests positive and $z_j = 0$ if otherwise. Finding maximum likelihood estimate of the parameter μ involves maximizing $L(\mu) = \prod_{j=1}^{J} \theta^{z_j}(1-\theta)^{1-z_j}$ as a function of μ, where $\theta = 1-(1-\mu)^n$ and $\mu \in (0,1)$; i.e., if $\widehat{\mu}$ denotes the maximum likelihood estimate of μ, then $\widehat{\mu} = \arg\max_{\mu} L(\mu)$. Show that

$$\widehat{\mu} = 1 - \left(1 - \frac{\sum_{i=1}^{J} z_i}{J}\right)^{1/n}$$

and find $\widehat{\mu}$ for the following data, where $J = 10$ and $n = 4$.

Pool testing data										
z	1	0	1	0	0	1	1	1	1	1

Solution:

For the given data, the likelihood function $L(\mu)$ is shown in Figure 6. One easily finds that a unique maximum exists and the maximizer of $L(\mu)$ is between $\mu = 0.2$ and $\mu = 0.4$. To calculate the exact value, we will prove and use the formula for the maximum likelihood estimator $\widehat{\mu}$. In practice, it is instructive to plot a function when its optimal value is searched over a one-dimensional parameter space.

Note that $\theta = 1-(1-\mu)^n \neq 0$ and $\frac{d\theta}{d\mu} = n(1-\mu)^{n-1} \neq 0$, for $\mu \in (0,1)$, so that $\frac{1}{\theta(1-\theta)}\frac{d\theta}{d\mu} \neq 0$. For the proof, it is more convenient to use $\ln L(\mu)$ instead of $L(\mu)$. Because natural logarithm is strictly increasing, the maximum value of $L(\mu)$ and $\ln L(\mu)$ will occur at the same point. We have $\ln L(\mu) = \sum_{j=1}^{J} z_j \ln\theta + \sum_{j=1}^{J}(1-z_j)\ln(1-\theta)$. Taking derivative with respect to μ yields

$$\frac{d\ln L(\mu)}{d\mu} = \frac{d\theta}{d\mu}\left(\frac{\sum_{j=1}^{J} z_j}{\theta} - \frac{J - \sum_{j=1}^{J} z_j}{1-\theta}\right).$$

Setting $\frac{d\ln L(\mu)}{d\mu} = 0$ and solving for μ result in

$$\widehat{\mu} = 1 - \left(1 - \frac{\sum_{i=1}^{J} z_i}{J}\right)^{1/n}.$$

To emphasize that the solution is an estimator, the notation $\widehat{\mu}$ is used in place of μ in the final step. When J is sufficiently large, the global maximum of $L(\mu)$ occurs at $\widehat{\mu}$. For the given

data, we find $\widehat{\mu} = 0.26$. A close inspection reveals that the likelihood function $L(\mu)$ in Figure 6 is maximum at $\widehat{\mu} = 0.26$.

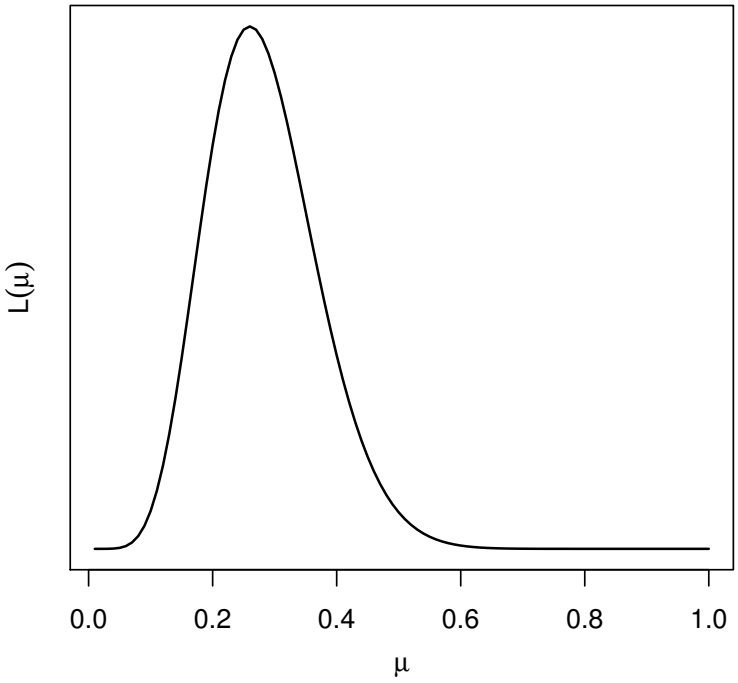

Figure 6: The likelihood function $L(\mu)$ with the given pool testing data.