Sampling, Amplifying, and Resampling

Tinajiao Chu

April 19, 2002

Technical Report No. CMU-PHIL-133

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

$$G^2 = -2\sum_{i=1}^{k} \frac{Z_i}{c_i} \log \frac{Z_i}{n_i \hat{p}_i}$$

where $\hat{p} = \dfrac{\sum_{i=1}^{s} Z_i/c_i}{\sum_{j=1}^{k} n_j/c_j}$

The above tests could be easily extended to tests for whether a set of categories all have constant relative frequencies in a set of populations. In the cases where $Z_1, \cdots, Z_k$ are small, it is preferred to use bootstrap method to get the distribution of the test statistics, rather than relying on the asymptotic distribution.

Our asymptotic results also show the possible consequences of treating SAR samples as multinomial samples. Accordion to Theorems 3 and 4, asymptotically the covariance matrix of a SAR sample is exactly c times the covariance matrix of a multinomial sample of the same size from the same population, where c is the normalizing factor of the SAR sample. When c is close to 1 (c is always greater than 1), it might be OK to treat the SAR sample as if it were multinomial. However, when c is large, the multinomial model will significantly underestimate the variance of the SAR sample. For example, suppose we are given k SAR samples from k populations, and asked to test the null hypothesis that a certain category has the same relative frequency in all the k populations. If we treated the k SAR samples as multinomial samples, we would use the traditional χ^2 or G^2 tests, and thought that they were χ^2_{k-1} distributed under the null hypothesis. However, the values of the traditional χ^2 and G^2 statistics are much higher than that of the test statistics modified for the SAR samples, which are indeed χ^2_{k-1} distributed under the null hypothesis. Needless to say, in such a situation, a traditional level α test based on the multinomial model will actually have a type I error rate much higher than α when applied to the SAR samples.

ACKNOWLEDGMENTS

I would like to thank Professors L. Wasserman, C. Glymour and P. Spirtes for the helpful discussions related to this work.

REFERENCE

Ash, R., Doleans-Dade, C. (2000), *Probability and Measure Theory,* 2ed., San Diego CA: Academic Press.

Bishop, Y., Fienberg, S., & Holland, P. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: The MIT Press.

Hajek, J. (1960), "Limit Distributions in Simple Random Sampling from a Finite Population", *Publ. Math. Inst. Hungar. Acad. Sci. 5*: 361-374.

Sun, F. (1999), "The Polymerase Chain Reaction and Branching Processes".

Velculescu, V., Zhang, L., Vogelstein, B., & Kinzler, K. (1995), "Serial Analysis of Gene Expression", *Science 270*: pp.484–487

Velculescu, V., Zhang, L., Zhou, W., Traverso, G., St. Croix, B., Vogelstein, B., & Kinzler, K. (2000), "Serial Analysis of Gene Expression Detailed Protocol", version 1.0e, John Hopkins Oncology Center and Howard Hughes Medical Center.

- Given the above result, conditional on $U_m = \sqrt{m}(\mathbf{X}_m - \mathbf{p}) = \mathbf{u}$, it can be shown that:

$$\sqrt{n_m}\left(\mathbf{Z}_m - \mathbf{p}\right) = \sqrt{n_m}\left(\mathbf{Z}_m - \mathbf{X}_m\right) + \sqrt{\frac{n_m}{m}}\mathbf{u} \implies N\left(\sqrt{\gamma_m}\mathbf{u}, \left(1 + \gamma_m \frac{\sigma^2}{\mu^2}\right)\Sigma_p\right)$$

  Then it is easy to show that, for any $\mathbf{z}$, $P(\mathbf{Z}_m \leq z|\mathbf{X}_m = \mathbf{x})$ is decreasing in $\mathbf{x}$, hence there is a dual distribution function $H'_{m,z}$ of $\mathbf{x}$ such that $H'_{m,z} = P(\mathbf{Z}_m \leq z|\mathbf{X}_m = \mathbf{x})$ almost surely with respect to the measure induced by $\mathbf{X}_m$. Thus, by the lemmas of conditional convergence,

$$\sqrt{n_m}\left(\mathbf{Z}_m - \mathbf{p}\right) \implies N\left(\mathbf{0}, \left(1 + \gamma_m \frac{\sigma^2}{\mu^2} + \gamma_m\right)\Sigma_p\right)$$

□

Like Theorem 2, Theorem 4 can be generalized to allow different distributions of the amplification factors for elements belonging to different categories. Let μ_i be the mean of the amplification factor for the i^{th} category, X_i be the relative frequency of the i^{th} category in the original sample. Using the delta method, we can get the asymptotic distribution of $\left(\frac{\mu_1 X_1}{\sum_{i=1}^{k+1}\mu_i X_i}, \cdots, \frac{\mu_k X_k}{\sum_{i=1}^{k+1}\mu_i X_i}\right)^T$. The generalized version of Theorem 2 is also needed. Otherwise, the proof of Theorem 4 remains largely unchanged. Of course, in this case, the covariance matrix for the asymptotic distribution of the relative frequencies will be extremely complicated.

The asymptotic results of Theorems 3 and 4 imply that the relative frequencies in a SAR sample are consistent estimators of the relative frequencies in the population. In particular, if the amplification factor is the simple type branch process described in Lemma 1, the relative frequencies in a SAR sample are indeed unbiased estimators of the relative frequencies in the population.

5 Discussion

In this paper we showed the asymptotic normality of the relative frequencies in the final sample. It is interesting to note that there is a striking similarity between the asymptotic distribution of the relative frequencies of the categories in the final sample of an SAR sample, and the asymptotic distribution of the relative frequencies in a multinomial sample. Let Z be the relative frequencies of the categories in an SAR sample, according to Theorem 4, $(\mathbf{Z}_m - \mathbf{p}) \implies N\left(\mathbf{0}, \frac{c_m}{n_m}\Sigma_p\right)$, where n_m is the size of the final sample, c_m the normalizing factor, and $\mathbf{p}$ the relative frequencies of the categories in the population. Thus, asymptotically Z_m behaves like the relative frequencies in a multinomial sample of size $\frac{n_m}{c_m}$ drawn from the same population. This observation leads immediately to tests of whether a category has the same relative frequency in two or more populations. In particular, we can extend the traditional χ^2 test or the G test for association in contingency tables in the following way:

Suppose we have k populations, and k SAR samples obtained from them respectively. Let $n_1, \cdots, n_k$ be the sizes of the final samples of the k SAR samples, and $c_1, \cdots, c_k$ be the normalizing factors for the k final samples. Suppose $Z_1, \cdots, Z_k$ are the counts of elements belonging to a specific category in the k final samples respectively. Under the null hypothesis that this category has the same relative frequency in the k populations, the following two test statistics both have asymptotically a χ^2_{k-1} distribution:

$$X^2 = \sum_{i=1}^{k} \frac{(Z_i - n_i\hat{p})^2}{c_i n_i \hat{p}}$$

Theorem 4 *Consider a multinomial sample of size m drawn from a population of $k+1$ categories of elements with relative frequencies $p_1, \cdots, p_k$, and $1 - \sum_{i=1}^{k} p_i$ respectively. Suppose each element of the multinomial sample is subject to i.i.d. amplifying processes such that the mean and variance of the amplification factor are μ and σ^2 respectively. A sample of size n_m is then drawn with replacement from the intermediate sample. Suppose $\mathbf{Z}_m = (Z_{m1}, \cdots, Z_{mk})^T$ is the relative frequencies of the first k categories in the final sample. As $m, n_m \to \infty$, we have:*

$$\mathbf{Z}_m \sim N\left(\mathbf{p}, \frac{1}{n_m} c_m \Sigma_p\right)$$

where $\mathbf{p} = (p_1, \cdots, p_k)^T$, $c_m = 1 + \frac{n_m}{m}\left(1 + \frac{\sigma^2}{\mu^2}\right)$, and Σ_p is a matrix with $\sigma_{p,ii} = p_i(1 - p_i)$ and $\sigma_{p,ij} = -p_i p_j$ for $i \neq j$.

Proof:

The general idea is similar to the proof of Theorem 3. Below is a sketch of the proof.

Let $\mathbf{X}_m = (X_{m1}, \cdots, X_{mk})^T$ and $\mathbf{Y}_m = (Y_{m1}, \cdots, Y_{mk})^T$ be the relative frequencies of the first k categories in the original sample and the intermediate sample respectively. Let $\frac{n_m}{m} = \gamma_m$. By the central limit theorem for the multinomial data, and Theorem 2 of the asymptotic distribution of the ratio in amplification, as $m, n_m \to \infty$:

$$\sqrt{m}\,(\mathbf{X}_m - \mathbf{p}) \implies N(\mathbf{0}, \Sigma_p)$$

Conditional on $X_m = \mathbf{x}$:

$$\sqrt{m}(\mathbf{Y}_m - \mathbf{x}) \implies N\left(\mathbf{0}, \frac{\sigma^2}{\mu^2}\Sigma_x\right)$$

where Σ_x is a matrix with $\sigma_{x,ii} = x_i(1 - x_i)$, and $\sigma_{x,ij} = -x_i x_j$ for $i \neq j$.

Conditional on $\mathbf{Y}_m = \mathbf{y}$:

$$\sqrt{n_m}\,(\mathbf{Z}_m - \mathbf{y}) \implies N(\mathbf{0}, \Sigma_y)$$

where Σ_y is a matrix with $\sigma_{y,ii} = y_i(1 - y_i)$, and $\sigma_{y,ij} = -y_i y_j$ for $i \neq j$.

Then we can show that:

- Conditional on $X_m = \mathbf{x}$ and $V_m = \sqrt{m}(\mathbf{Y}_m - \mathbf{x}) = \mathbf{v}$, we have $\mathbf{Y}_m \to \mathbf{x}$, hence:

$$\sqrt{n_m}\,(\mathbf{Z}_m - \mathbf{x}) = \sqrt{n_m}\,(\mathbf{Z}_m - \mathbf{Y}_m) + \sqrt{\frac{n_m}{m}}\mathbf{v} \implies N(\sqrt{\gamma_m}\mathbf{v}, \Sigma_x)$$

- Let $G_{m,z} = P(\mathbf{Z}_m \leq z | \mathbf{Y}_m = \mathbf{y})$. From the probability mass function of the multinomial distribution, it is easy to check that $G_{m,z}$ is continuous and decreasing in $\mathbf{Y}_m$ for $0 \leq y_i \leq 1$. Extend $G_{m,z}$ to a function $G'_{m,z}$ defined on $\mathbb{R}^k$ such that when $y_i \to \infty$ for some $1 \leq i \leq k$, $G'_{m,z} \to 0$ decreasingly and continuously, while when $y_i \to \infty$ for all $1 \leq i \leq k$, $G'_{m,z} \to 1$ increasingly and continuously. Because $G'_{m,z} = P(\mathbf{Z}_m \leq z | \mathbf{Y}_m = \mathbf{y})$ a.s. with respect to the measure induced by $\mathbf{Y}_m$, by the lemmas of conditional convergence,

$$\sqrt{n_m}\,(\mathbf{Z}_m - \mathbf{x}) \implies N\left(\mathbf{0}, \left(1 + \gamma_m \frac{\sigma^2}{\mu^2}\right)\Sigma_x\right)$$

number of labels $l_1, \cdots, l_{x_m}$ is less than or equal to z, and $C_{x_m+1,z}$ the set of possible final samples where the total number of labels $l_1, \cdots, l_{x_m+1}$ is less than or equal to z. Now $P(Z_m \leq z|X_m = x_m)$ is simply the probability of getting a possible final sample s such that $s \in C_{x_m,z}$, and $P(Z_m \leq z|X_m = x_m + 1)$ is the probability of getting a possible final sample s' such that $s' \in C_{x_m,z}$. Clearly $C_{x_m+1,z} \subset C_{x_m,z}$, which implies that $P(Z_m \leq z|X_m = x_m + 1) \leq P(Z_m \leq z|X_m = x_m)$

- As a function of x_m, $P(Z_m \leq z|X_m = x_m)$ is defined at finite points. We can interpolate linearly between these points, and extend smoothly and increasingly below the smallest point, which is 0, so that the function converges to 1 as $x_m \to -\infty$, and extend smoothly and increasingly above the largest point, which is m, so that the function converges to 0 as $x_m \to \infty$. Call this extended function H_m. Clearly, $H_m = P(Z_m \leq z|X_m = x_m)$ a.s. with respect to the measure introduced by X_m.

Now we have shown that the conditional distribution function of Z_m given $X_m = x_m$ is a dual distribution function in x_m, hence the conditional distribution function of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma m p)$ given $U_m = u$ is a dual distribution function in u. Given that a normal distribution function is a dual distribution function in its mean, and the fact that $U_m \implies N(0, p(1-p))$, by the lemmas of conditional independence,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma m p) \implies N\left(0, \left(1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2} + \gamma\right) p(1-p)\right)$$

Let $c_m = 1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2} + \gamma = 1 - \frac{N_m}{m\mu} + \frac{N_m}{m}\left(1 + \frac{\sigma^2}{\mu^2}\right)$, given that $\frac{1}{\sqrt{\gamma m}}(N_m - \gamma m) \to 0$ w.p.1., as $m \to \infty$

$$Z_m \implies N\left(N_m p, N_m c_m p(1-p)\right)$$

□

We note that the assumption that the original sample is binomial is not essential to our proof. If it is hypergeometric with the ratio of the sample to population being δ_m, the above result still holds, with the exception that the normalizing factor now is changed to $c'_m = 1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2} + \gamma(1 - \delta_m)$.

It seems reasonable to conjecture that Theorem 3 could be extended to the multivariate SAR samples, where the original samples are multinomial or multivariate hypergeometric, and the final samples are multivariate hypergeometric conditional on the intermediate samples. Let $\mathbf{X}_m$ and $\mathbf{Z}_m$ be the counts of first k categories of elements in the original sample and the final sample respectively. To show the asymptotic normality of $\mathbf{Z}_m$, we would only need to show the asymptotic normality of $\mathbf{Z}_m$ given $\mathbf{X}_m$. One approach would be to prove directly the asymptotic normality by the lemmas of conditional convergence. Another approach would be using the Cramer-Wold's theorem, i.e., showing the appropriate asymptotic normality of $\mathbf{u}^T\mathbf{Z}$ for an arbitrary $\mathbf{u} = (u_1, \cdots, u_k)^T$ conditional on $\mathbf{X}_m$. We tried both approaches, but were unable to get the desired result. Here we shall leave the multivariate version of Theorem 3 as a conjecture.

Although it is difficult to extend Theorem 3 to the multivariate SAR sample with multivariate hypergeometric final sample, we could show the asymptotic normality of the multivariate SAR sample if the final sample is drawn *with* replacement from the intermediate sample:

by defining $G'_{m,z}(y_m, -r_m) = G_{m,z}(\max(0, y_m), \min(0, -r_m))$. Clearly $G'_{m,z}$ is still decreasing in $(y_m, -r_m)$. Now extend $G'_{m,z}$ to a continuous decreasing function $G''_{m,z}$. It is easy to check that $G''_{m,z}$ has the desired limit behavior. That is, it goes decreasingly to 0 when either $y_m \to \infty$ or $-r_m \to \infty$, and goes increasingly to 1 when both $y_m \to -\infty$ and $-r_m \to -\infty$.

$G''_{m,z}$ is a dual distribution function in $(y_m, -r_m)$ such that $G''_{m,z} = P(Z_m \leq z | Y_m = y_m, -R_m = -r_m)$ a.s. with respect to the measure induced by $(Y_m, -R_m)$. Thus, $P\left(\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \leq z \middle| V_m, W_m\right)$ can also be extended to a function $H_{m,z}$ such that $H_{m,z} = P\left(\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \leq z \middle| V_m, W_m\right)$ a.s. with respect to the measure introduced by $(V_m, -W_m)$, and $H_{m,z}$ is a dual distribution function in $(v, -w)$. It is also easy to check that the distribution function for $N(\mu, \sigma^2)$ is a dual distribution function in μ. Moreover, if $\mu = x - y$, this distribution function is a dual distribution function in $(x, -y)$.

Given that the joint distribution of V_m and W_m converges to a bivariate normal, [15] by the lemmas of conditional convergence, the asymptotic normality of the conditional distribution of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m)$ given $\frac{X_m}{m} = x$ follows immediately. [16]

3. Next, we show that, conditional on $U_m = \frac{1}{\sqrt{m}}(X_m - mp) = u$,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp) \implies N\left(\sqrt{\gamma}u, \left(1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2}\right) p(1-p)\right)$$

   The argument is similar to the one used in step 1. Basically, conditional on $U_m = u$, as $m \to \infty$, $\frac{X_m}{m}\left(1 - \frac{X_m}{m}\right) = \left(p + \frac{u}{\sqrt{m}}\right)\left(1 - p + \frac{u}{\sqrt{m}}\right) \to p(1-p)$. Then we notice that:

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp) = \frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) + \frac{1}{\sqrt{\gamma m}}(\gamma X_m - \gamma mp) = \frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) + \sqrt{\gamma}u$$

4. Finally, given the asymptotic conditional normality of Z_m given X_m, and the asymptotic normality of X_m, to derive the asymptotic distribution of Z_m using our lemmas of conditional convergence, we only need to show that the conditional distribution of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma mp)$ given $U_m = u$ is a dual distribution function in u. This is equivalent to showing that, conditional on $X_m = x_m$, the distribution function of Z_m is a dual distribution function in x_m.

   - First, we show that the conditional distribution function of Z_m given $X_m = x_m$ is decreasing in x_m.

     Imagine that we mark each of the m elements in the original sample with a unique label, say, l_i for the i^{th} element. Then after the amplification and the resampling steps, a final sample of size N_m of the labels is obtained. The probability of getting a specific sample is uniquely determined by the size m of the original sample. Let $C_{x_m,z}$ be the set of possible final samples where the total

[15] The mean is $\mathbf{0}$, and the covariance matrix is:

$$\begin{bmatrix} \sigma^2 x & 0 \\ 0 & \sigma^2(1-x) \end{bmatrix}$$

[16] If $M \sim N(0, \sigma_1^2)$, and conditional on $M = \mu$, $Z \sim N(\mu, \sigma^2)$, then $Z \sim N(0, \sigma_1^2 + \sigma^2)$.

We observe that conditional on $Y_m = y_m$ and $R_m = r_m$, Z_m has a hypergeometric distribution with parameters $(y_m + r_m, y_m, N_m)$, where $y_m + r_m$ is the population size, y_m the number of elements in population belonging to the first category, and N_m the sample size. Let $h_{N,M,n}$ be the distribution function for a hypergeometric function with parameters (N, M, n). Obviously, we need to show that, for any (N, M, n), and any z:

$$h_{N,M,n}(z) \geq h_{N+1,M+1,n}$$

$$h_{N,M,n}(z) \geq h_{N-1,M,n}$$

- Let $F_1 = h_{N,M,n}(z)$, $F_2 = h_{N+1,M+1,n}$, and f_1 and f_2 be the probability mass functions corresponding to F_1 and F_2 respectively:

  $$F_1(z) = \sum_{x=0}^{z} f_1(z) = \sum_{x=0}^{z} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

  $$F_2(z) = \sum_{x=0}^{z} f_2(z) = \sum_{x=0}^{z} \frac{\binom{M+1}{x}\binom{N-M}{n-x}}{\binom{N+1}{n}}$$

  Solve the inequality $f_1(x) > f_2(x)$ for $0 \leq x \leq \min(M, n)$, [13] we get: $f_1(x) > f_2(x)$ if and only if $x < \frac{n}{N+1}(M+1)$. Thus, $F_1(z) \geq F_2(z)$ when $z \leq \frac{n}{N+1}(M+1)$.
  Also note that if $n \leq M$, then $F_1(n) = 1 = F_2(n)$, and if $n > M$, then

  $$F_1(M) = 1 > F_2(M) = 1 - \frac{\binom{N-M}{n-M-1}}{\binom{N+1}{n}}$$

  Thus, $F_1(z) \geq F_2(z)$ for $z \geq \min(M, n)$. For $\frac{n}{N+1}(M+1) < z < \min(n, M)$,

  $$F_1(z) - F_2(z) > F_1(z+1) - F_2(z+1) > \cdots > F_1(\min(n, M)) - F_2(\min(n, M)) \geq 0$$

  because $f_1(z) > f_2(z)$ for $\frac{n}{N+1}(M+1) < z < \min(n, M)$.
- Let $F_3 = h_{N-1,M,n}$, and f_3 be the corresponding probability mass function. The proof for $F_1(z) \geq F_3(z)$ is similar to the above one. Basically, we solve the inequality $f_1(x) > f_3(x)$ for $0 \leq x \leq \min(M, n)$, and get: $f_1(x) > f_3(x)$ if and only if $x < \frac{nM}{N}$.
  Thus, $F_2(z) \geq F_3(z)$ for $z \leq \frac{nM}{N}$. It is also easy to see that $F_1(\min(n, M)) = 1 = F_3(\min(n, M))$. Therefore, by a similar argument as above, we can show that $F_1(z) \geq F_3(z)$ for any z.

- The continuity and the proper convergence to 0 and 1 are easy to satisfy. Fix a z such that $0 \leq z < N_m$, [14] define $G_{m,z}(y_m, -r_m) = P(Z_m \leq z | Y_m = y_m, -R_m = -r_m)$. Clearly $G_{m,z}$ is defined for only a countable number of nonnegative/nonpositive integer pairs of $(y_m, -r_m)$. First extend $G_{m,z}$ to allow negative integer values for Y_m and/or positive integer values for $-R_m$

[13] $f_1(x) = 0$ for $x > \min(M, n)$, and $f_2(x) = 0$ for $x > \min(M+1, n)$.

[14] There is no need to worry about $z < 0$, where $P(Z_m \leq z | Y_m = y_m, -R_m = -r_m) = 0$, and $z \geq N_m$, where $P(Z_m \leq z | Y_m = y_m, -R_m = -r_m) = 1$.

$$= \frac{1}{\sqrt{N_m}}\left(Z_m - N_m\frac{Y_m}{Y_m+R_m}\right) + \frac{\sqrt{N_m m}}{Y_m+R_m}[(1-x)v - xw]$$

Conditional on $\frac{X_m}{m} = x$, $V_m = \frac{1}{\sqrt{m}}(Y_m - m\mu x) = v$, and $W_m = \frac{1}{\sqrt{m}}(R_m - m\mu(1-x)) = w$, [11] as $m \to \infty$,

$$N_m \to \gamma m$$

$$\frac{1}{\sqrt{\gamma m}}(N_m - \gamma m) \to 0$$

$$\frac{\sqrt{N_m m}}{Y_m+R_m} = \frac{\sqrt{N_m m}}{m\mu + \sqrt{m}(v+w)} \to \frac{\sqrt{\gamma}}{\mu}$$

$$\frac{N_m}{Y_m+R_m} = \frac{N_m}{m\mu + \sqrt{m}(v+w)} \to \frac{\gamma}{\mu}$$

$$\frac{Y_m R_m}{(Y_m+R_m)^2} = \frac{(m\mu x + \sqrt{m}v)[m\mu(1-x) + \sqrt{m}w]}{m\mu + \sqrt{m}(v+w)} \to x(1-x)$$

Thus, conditional on $\frac{X_m}{m} = x$, $V_m = v$, and $W_m = w$,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \implies N\left(\frac{\sqrt{\gamma}}{\mu}[(1-x)v - xw], \left(1 - \frac{\gamma}{\mu}\right)x(1-x)\right)$$

2. Next we show that conditional on $\frac{X_m}{m} = x$,

$$\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m) \implies N\left(0, \left(1 - \frac{\gamma}{\mu} + \gamma\frac{\sigma^2}{\mu^2}\right)x(1-x)\right)$$

First we show that given $\frac{X_m}{m} = x$, the conditional distribution of $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m)$ given $V_m = v$ and $W_m = w$ is a dual distribution function in $(v, -w)$. This is equivalent to showing that the conditional distribution of Z_m given Y_m and $-R_m$ is a dual distribution function in Y_m and $-R_m$. [12]

- First we show that $P(Z_m \leq z | Y_m = y_m, -R_m = -r_m)$ is decreasing in $(y_m, -r_m)$. Because the minimal change in y_m or r_m is 1, it suffices to show that, for any z,

$$P(Z_m \leq z | Y_m = y_m, -R_m = -r_m) \geq P(Z_m \leq z | Y_m = y_m + 1, -R_m = -r_m)$$

$$P(Z_m \leq z | Y_m = y_m, -R_m = -r_m) \geq P(Z_m \leq z | Y_m = y_m, -R_m = -r_m + 1)$$

[11] Note that here x, v, and w are constant, hence independent of m.

[12] The equivalence relation holds because $\frac{1}{\sqrt{\gamma m}}(Z_m - \gamma X_m)$, V_m, and W_m are linear in Z_m, Y_m, and R_m respectively, and the slopes of the linear transformations are all positive.

Theorem 3 *Consider the following SAR scheme: The original sample is a binomial sample with parameters (m, p), where m is the sample size, and p the relative frequency of the elements belonging to the first category in the population. The mean and the variance of the amplification factor for the amplifying process are μ and σ^2 respectively. Let M_m be the intermediate sample size. The final sample of size N_m is drawn without replacement from the intermediate sample, where N_m is a random variable such that, for some $0 < \gamma < \mu$, $N_m = M_m$ if $M \leq \gamma m$ and $N_m = \gamma m$ otherwise.* [10] *Suppose Z_m is the count of elements belonging to the first category in the final sample. Then as $m \to \infty$,*

$$Z_m \sim N\left(N_m p, N_m c_m p(1-p)\right)$$

where $c_m = 1 - \dfrac{N_m}{m\mu} + \dfrac{N_m}{m}\left(1 + \dfrac{\sigma^2}{\mu^2}\right)$ *is called the normalizing factor of the SAR sample.*

Proof: Let X_m be the count of elements belonging to the first category in the original sample, and Y_m and R_m be the counts of elements belonging to the first and the second categories in the intermediate sample respectively. By the central limit theorem, as $m \to \infty$:

$$U_m = \frac{1}{\sqrt{m}}\left(X_m - mp\right) \implies N\left(0, p(1-p)\right)$$

Conditional on $\dfrac{X_m}{m} = x$:

$$V_m = \frac{1}{\sqrt{m}}(Y_m - m\mu x) \implies N\left(0, \sigma^2 x\right)$$

$$W_m = \frac{1}{\sqrt{m}}(R_m - m\mu(1-x)) \implies N\left(0, \sigma^2(1-x)\right)$$

Note that Y_m and R_m are independent conditional on X_m, hence conditional on $\dfrac{X_m}{m} = x$:

$$\frac{1}{\sqrt{m}}[(1-x)Y_m - xR_m)] = \frac{1}{\sqrt{m}}[(1-x)(Y_m - \mu m x) - x(R_m - \mu m(1-x))] \implies N\left(0, \sigma^2(1-x)x\right)$$

Conditional on $\dfrac{Y_m}{m} = y$ and $\dfrac{R_m}{m} = r$:

$$\frac{1}{\sqrt{N_m}}\left(Z_m - N_m\frac{y}{y+r}\right) \implies N\left(0, \left(1 - \frac{N_m}{m(y+r)}\right)\frac{yr}{(y+r)^2}\right)$$

Now we are going to prove the asymptotic normality of Z_m in four steps.

1. Conditional on $\dfrac{X_m}{m} = x$:

$$\frac{1}{\sqrt{N_m}}(Z_m - \frac{N_m}{m}X_m) = \frac{1}{\sqrt{N_m}}\left(Z_m - N_m\frac{Y_m}{Y_m + R_m}\right) + \frac{1}{\sqrt{N_m}}\left(N_m\frac{Y_m}{Y_m + R_m} - \frac{N_m}{m}X_m\right)$$

$$= \frac{1}{\sqrt{N_m}}\left(Z_m - N_m\frac{Y_m}{Y_m + R_m}\right) + \frac{\sqrt{N_m}}{Y_m + R_m}\left[(1-x)(Y_m - xm\mu) - x(R_m - (1-x)m\mu)\right]$$

[10] In this SAR scheme, if the intermediate sample size M_m is less than or equal to γm, then the whole intermediate sample is taken as the final sample. Otherwise, a final sample of size γm will be drawn without replacement from the intermediate sample.

Proof:

Similar as above. Note that $G_n \to G$ pointwisely implies that $\mu_n \implies \mu$, where μ_n and μ are the measures determined by G_n and G.

□

Corollary 7 and the following lemma will be called the lemmas of conditional convergence. Together they give a sufficient condition for conditional convergence, but they can also be used independently.

Lemma 3 *Consider random variables $\{\mathbf{X}_n\}$, $\mathbf{X}$, $\{\mathbf{Y}_n\}$ and $\mathbf{Y}$. Let μ_n and μ be the measures induced by $\mathbf{X}_n$ and $\mathbf{X}$ respectively. Suppose the following conditions are satisfied:*

1. *$\mathbf{X}_n \implies \mathbf{X}$, and the distribution function of X is continuous.*
2. *For any fixed $\mathbf{y}$ and for all n, there is a μ_n measurable function $G_{n,\mathbf{y}} = P(\mathbf{Y}_n \leq \mathbf{y}|\mathbf{X}_n = \mathbf{x})$ a.s.$[\mu_n]$ such that $G_{n,\mathbf{y}} \to P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})$ uniformly, and $P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})$ is continuous in $\mathbf{x}$.*

Then $\mathbf{Y}_n \implies \mathbf{Y}$.

Proof:

It suffices to show that for all $\mathbf{y}$,

$$\int P(\mathbf{Y}_n \leq \mathbf{y}|\mathbf{X}_n = \mathbf{x})\, d\mu_n(\mathbf{x}) \to \int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu(\mathbf{x})$$

or equivalently,

$$\int G_{n,\mathbf{y}}\, d\mu_n(\mathbf{x}) \to \int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu(\mathbf{x})$$

Fixed $\mathbf{y}$, because $G_{n,\mathbf{y}}$ converges to $P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})$ uniformly, for any $\epsilon > 0$, there is an $N(\epsilon)$ such that for any $n \geq N(\epsilon)$,

$$\sup_{\mathbf{x}} |G_{n,\mathbf{y}} - P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})| \leq \epsilon$$

Because $\mu_n \to \mu$ and $P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})$ is bounded and continuous, we can choose M_ϵ such that for all $n \geq M_\epsilon$,

$$\left|\int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu_n(\mathbf{x}) - \int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu(\mathbf{x})\right| < \epsilon$$

Therefore, for all $n > \max(N_\epsilon, M_\epsilon)$,

$$\begin{aligned}
&\left|\int G_{n,\mathbf{y}}\, d\mu_n(\mathbf{x}) - \int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu(\mathbf{x})\right| \\
&\leq \left|\int G_{n,\mathbf{y}}\, d\mu_n(\mathbf{x}) - \int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu_n(\mathbf{x})\right| \\
&\qquad + \left|\int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu_n(\mathbf{x}) - \int P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})\, d\mu(\mathbf{x})\right| \\
&\leq \int |G_{n,\mathbf{y}} - P(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x})|\, d\mu_n(\mathbf{x}) + \epsilon \leq 2\epsilon
\end{aligned}$$

□

Now we can derive the asymptotic distribution for the relative frequency of a category in the final sample of an SAR scheme.

Proposition 1 *A dual distribution function G on $\mathbb{R}^k$ uniquely determines a probability measure μ such that $\mu(\{\mathbf{x} : x_1 \geq y_1, \cdots, x_k \geq y_k\}) = G(\mathbf{y})$ for any $\mathbf{y} = (y_1, \cdots, y_k)^T \in \mathbb{R}^k$.*

Note that if F is the distribution function corresponding to a measure μ, then the dual distribution function G for μ in general is not equal to $1 - F$. More precisely, we have:

Proposition 2 *$G = 1 - F$ if and only if F is a continuous distribution function on $\mathbb{R}$.*

The following lemma is an extension of the well known theorem of the uniform convergence of the distribution functions on $\mathbb{R}$.[9]

Lemma 2 *Consider a continuous distribution function F defined on $\mathbb{R}^k$. If there is a sequence of distribution functions $\{F_n\}$ converge weakly to F, then F_n converges to F uniformly.*

Proof:

Let the measures corresponding to F and F_n be μ and μ_n respectively. Define a compact set C_a as $C_a = \{(x_1, \cdots, x_k)^T : |x_1| \leq a, \cdots, |x_k| \leq a\}$. Note that F is uniformly continuous on C_a. For any $\epsilon > 0$, choose an a such that $\mu(C_a^c) < \epsilon$. Then we can find a finite number of compact sets $B_1, \cdots, B_m$ such that $\bigcup_{i=1}^m B_i = C_a$ and that $\max_{\mathbf{x},\mathbf{y} \in \mathbf{B_i}}(|F(\mathbf{x}) - F(\mathbf{y})|) \leq \epsilon$ for all $1 \leq i \leq m$.

Let $\mathbf{x}_{i,max}$ and $\mathbf{x}_{i,min}$ be the maximum and the minimum points in B_i. Because $F_n \Longrightarrow F$, we can find an $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, and for all $1 \leq i \leq m$, $|F_n(\mathbf{x}_{i,max}) - F(\mathbf{x}_{i,max})|) \leq \epsilon$, $|F_n(\mathbf{x}_{i,min}) - F(\mathbf{x}_{i,min})|) \leq \epsilon$, and $|\mu_n(C_a) - \mu(C_a)| \leq \epsilon$. It then follows that, for all $n \geq N(\epsilon)$, first, $|F_n(\mathbf{x}) - F(\mathbf{x})| \leq 3\epsilon$ for any $\mathbf{x} \in C_a$; second, $\mu_n(C_a^c) \leq 2\epsilon$.

For any $\mathbf{x} = (x_1, \cdots, x_k)^T \in \mathbb{R}^k$, define a set $L_\mathbf{x} = \{\mathbf{y} = (y_1, \cdots, y_k) : y_1 \leq x_1, \cdots, y_k \leq x_k\}$. Note that for any $\mathbf{x}$, $\mu_n(L_\mathbf{x}) = F_n(\mathbf{x})$, and $\mu(L_\mathbf{x}) = F(\mathbf{x})$. Let $\mathbf{a} = (a, \cdots, a)^T$. Now let us consider the following two situations:

- Suppose $C_a \cap L_\mathbf{x} = \emptyset$. Obviously, in this case, $|F_n(\mathbf{x}) - F(\mathbf{x})| = |\mu_n(L_\mathbf{x}) - \mu(L_\mathbf{x})| \leq 2\epsilon$.
- Suppose $C_a \cap L_\mathbf{x} = C_{\mathbf{a},\mathbf{x}} \neq \emptyset$. Clearly, $C_{\mathbf{a},\mathbf{x}}$ is compact, hence has a maximum point $\mathbf{x}_{a,max}$. It is easy to see that: $L_{\mathbf{x}_{\mathbf{a},\mathbf{max}}} \subset L_\mathbf{x}$, and $(L_\mathbf{x} \setminus L_{\mathbf{x}_{\mathbf{a},\mathbf{max}}}) \cap C_a = \emptyset$. Now we have:

$$|F_n(\mathbf{x}) - F(\mathbf{x})| = |[\mu_n(L_\mathbf{x} \setminus L_{\mathbf{x}_{a,max}}) + F_n(\mathbf{x}_{a,max})] - [\mu(L_\mathbf{x} \setminus L_{\mathbf{x}_{a,max}}) + F(\mathbf{x}_{a,max})]|$$
$$\leq |\mu_n(L_\mathbf{x} \setminus L_{\mathbf{x}_{a,max}}) - \mu(L_\mathbf{x} \setminus L_{\mathbf{x}_{a,max}})| + |F_n(\mathbf{x}_{a,max}) - F(\mathbf{x}_{a,max})|$$
$$\leq 2\epsilon + 3\epsilon = 5\epsilon$$

Therefore, we have shown that for any $n > N(\epsilon)$, and any $\mathbf{x} \in \mathbb{R}^k$, $|F_n(\mathbf{x}) - F(\mathbf{x})| \leq 5\epsilon$, hence F_n converges to F uniformly.

□

The following corollary shows the uniform convergence the dual distribution functions on $\mathbb{R}^k$.

Corollary 7 *If G is a continuous dual distribution function, and a sequence of dual distribution functions G_n converge to G pointwisely. Then G_n converges to G uniformly.*

[9] For example, see Theorem 7.6.2 of Ash and Doleans-Dade (2000).

Corollary 5 *If in Theorem 2 and Corollary 4, instead of requiring $n = N_{n,1} \leq N_{n,i} \leq Mn$, we require that $Ln \leq N_{n,i} \leq Mn$, where L is some positive real number, the conclusions still hold.*

Proof: The proofs in Theorem 2 and Corollary 4 depend only on the assumption that there is a fixed number M such that:

$$M \min(N_{n,1}, \cdots, N_{n,k+1}) \geq \min(N_{n,1}, \cdots, N_{n,k+1})$$

□

Corollary 6 *If in Theorem 2, we assume that $\Sigma_n \to \Sigma$, then:*

$$(Y_{n,1}, \cdots, Y_{n,k})^T \implies N(\mathbf{0}, \Sigma)$$

4 Asymptotic Distribution of the Final Sample

Theorems 1 and 2 give the asymptotic distribution of the relative frequencies in the intermediate sample conditional on the original sample. The asymptotic distributions for the relative frequencies in the original sample, and the relative frequencies in the final sample conditional on the intermediate sample, are straightforward: Both the relative frequencies in a multinomial sample and the relative frequencies in a multivariate hypergeometric sample converge weakly to multivariate normal. [8] We need to put these pieces together to get the marginal asymptotic distribution of the relative frequencies in the final sample. The basic idea is to show, under certain conditions, that conditional convergence implies marginal convergence. More precisely, consider two sequences of random variables X_i and Y_i, as well as two random variables X and Y. We say Y_i converges to Y conditional on X_i if 1), $X_i \implies X$, and 2), there are versions of $P(Y_i \leq y | X_i = x)$ and a Borel set A such that $\mu_X(A) = 1$ and for each fixed $x \in A$, $P(Y_i \leq y | X_i = x) \to P(Y \leq y | X = x)$, where μ_X is the measure induced by X. The goal is to find a sufficient condition to guarantee $Y_i \implies Y$.

To do so, we first introduce a new concept called the *dual distribution function* (ddf). The dual distribution functions are defined in a similar way as the distribution functions so that the dual distribution functions could share some properties, such as the uniform convergence, of the distribution functions.

Definition 1 *A nonnegative function G on $\mathbb{R}^k$ is called a dual distribution function if it satisfies the following conditions:*

- *G is continuous from below.*
- *G is decreasing.*
- *Let $\mathbf{x} = (x_1, \cdots, x_k)^k$, and $i \in \{1, \cdots, k\}$. If for some i, $x_i \to \infty$, then $G(\mathbf{x}) \to 0$. If $x_i \to -\infty$ for all i, then $G(\mathbf{x}) \to 1$.*

It is easy to check the following properties of a dual distribution function:

[8] Let $\mathbf{X}$ be a k dimensional random vector following a multivariate hypergeometric distribution with parameters $(N; N_1, \cdots, N_k; n)$, where N is the population size, and n is the sample size. Let $\frac{n}{N} = \beta$, $\frac{N_i}{N} = p_i$ and $\mathbf{p} = (p_1, \cdots, p_k)^T$. Fixing $\mathbf{p}$ and β, as $n \to \infty$, $(\mathbf{X} - n\mathbf{p}) \implies N(\mathbf{0}, (1-\beta)n\Sigma_p)$, where Σ_p is the covariance matrix of a multinomial distribution with parameters $(1; p_1, \cdots, p_k)$. For a general proof, see Hajek (1960).

It then follows:

$$\frac{\sqrt{N_n}\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n}{s_n} \implies N(0,1)$$

Now because the covariance matrix for $\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n$ is the identity matrix $\mathbf{I}_k$,

$$s_n^2 = \text{Var}\left(\sqrt{N_n}\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n\right) = N_n\sum_{i=1}^{k} u_i^2$$

Therefore,

$$\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n \implies N\left(0,\sum_{i=1}^{k} u_i^2\right)$$

which is the same as the distribution of $\mathbf{u}^T\mathbf{Z}$, where $\mathbf{Z} \sim N(\mathbf{0},\mathbf{I}_k)$.
Thus,

$$\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n \implies N(\mathbf{0},\mathbf{I}_k)$$

Because $\frac{\sum_{j=1}^{N_n} X_j}{N_n\mu} \rightarrow 1$ w.p.1., it then follows that, for any k-dimensional vector $\mathbf{u} = (u_1,\cdots,u_k)^T$,

$$\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Y}_n = \mathbf{u}^T\frac{\sum_{j=1}^{N_n} X_j}{N_n\mu}\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n \implies \mathbf{u}^T N\left(\mathbf{0},\mathbf{I}_k\right)$$

□

In Theorems 1 and 2, we assume that the amplification factors of all elements are identically distributed. It is possible to generalize the two theorems to allow the amplification factors for elements belonging to different categories to have different distributions. Let μ_i and σ_i^2 be the mean and the variance of the amplification factor for the i^{th} category respectively. Under the new condition, the relative frequency of the i^{th} category converges to $q_{n,i} = \frac{N_{n,i}\mu_i}{\sum_{j=1}^{k+1}(N_{n,j}\mu_j)}$. Define $Y'_{n,i} = \sqrt{N_n}\left(\frac{\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j}{\sum_{j=1}^{N_n} X_j} - q_{n,i}\right)$, then there is a matrix Σ'_n such that $\Sigma'^{-\frac{1}{2}}_n(Y_{n,1},\cdots,Y_{n,k})^T \implies N\left(\mathbf{0},\mathbf{I}_k\right)$.

Unfortunately, the matrix Σ'_n is much more complicated than Σ_n. Below are the first two elements of the first column of Σ'_n:

$$\Sigma'_n = \begin{bmatrix} (1-q_{n,1})^2 p_{n,1}\sigma_1^2 + q_{n,1}^2\sum_{i=2}^{k+1} p_{n,i}\sigma_i^2 & \cdots \\ -(1-q_{n,1})q_{n,2}p_{n,1}\sigma_1^2 - (1-q_{n,2})q_{n,1}p_{n,2}\sigma_2^2 + q_{n,1}q_{n,2}\sum_{i=3}^{k+1} p_{n,i}\sigma_i^2 & \cdots \\ \vdots & \vdots \end{bmatrix}$$

The following corollaries are generalizations of corollaries 1, 2 and 3 respectively.

Corollary 4 *In Theorem 2, if $E[X_i^4] < \infty$ and $X_i \geq c > 0$, then the covariance matrix of $\Sigma_n^{-\frac{1}{2}}\mathbf{Y}_n^T$ converges to $\mathbf{I}_k$, where $\mathbf{Y}_n^T = (Y_{n,1},\cdots,Y_{n,k})^T$.*

Proof: From Corollary 1, $\text{E}[Y_{n,i}^4] < \infty$ for $i = 1,\cdots,k+1$. Therefore, for any $\mathbf{u} = (u_1,\cdots,u_k)^T$, $\text{E}[(\mathbf{u}^T\mathbf{Y}_n)^4] < \infty$. Thus the variance of $\mathbf{u}^T\mathbf{Y}_n$ converges to the variance of the distribution μ if $\mathbf{u}^T\mathbf{Y}_n \implies \mu$.

□

$$\Sigma_n = \begin{bmatrix} p_{n,1}(1-p_{n,1}) & -p_{n,1}p_{n,2} & \cdots & -p_{n,1}p_{n,k} \\ -p_{n,2}p_{n,1} & p_{n,2}(1-p_{n,2}) & \cdots & -p_{n,2}p_{n,k} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{n,k}p_{n,1} & -p_{n,k}p_{n,2} & \cdots & p_{n,k}(1-p_{n,k}) \end{bmatrix}$$

With the convention that $N_{n,0} = 0$, define, for $i = 1, \cdots, k+1$:

$$Y_{n,i} = \frac{\sqrt{N_n}\mu}{\sigma}\left(\frac{\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j}{\sum_{j=1}^{N_n} X_j} - p_{n,i}\right)$$

Then:

$$\Sigma_n^{-\frac{1}{2}}\mathbf{Y}_n \implies N(\mathbf{0}, \mathbf{I}_k)$$

where $\mathbf{Y}_n = (Y_{n,1}, \cdots, Y_{n,k})^T$, and $\mathbf{I}_k$ is the $k \times k$ identity matrix.

Proof:

Σ_n is the covariance matrix for a k-dimensional random vector $(V_1, \cdots, V_k)$ where $(V_1, \cdots, V_k, 1-\sum_{i=1}^k V_i)$ has a multinomial distribution with parameters $\left(1; \frac{N_{n,1}}{N_n}, \cdots, \frac{N_{n,k+1}}{N_n}\right)$. Therefore, Σ_n is positive definite, and $\Sigma_n^{-\frac{1}{2}}$ exists.

Now define a new set of random vectors $\mathbf{Z}_n = (Z_{n,1}, \cdots Z_{n,k})$:

$$Z_{n,i} = Y_{n,i}\frac{\sum_{j=1}^{N_n} X_j}{N_n\mu} = \frac{1}{\sigma\sqrt{N_n}}\left(\sum_{j=N_{n,0}+\cdots+N_{N,i-1}+1}^{N_{n,1}+\cdots+N_{n,i}} X_j - p_{n,i}\sum_{j=1}^{N_n} X_j\right)$$

It is easy to check that Σ_n is the covariance matrix of $(Z_{n,1}, \cdots, Z_{n,k})^T$. Now let $\mathbf{u} = (u_1, \cdots, u_k)^T$ be any k-dimensional vector. Using the similar method used in the proof of Theorem 1, we can decompose $\sqrt{N_n}\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n$ into the sum of n independent random variables $U_1, \cdots, U_n$ with zero mean such that the Lindeberg's condition is satisfied. This is possible because the absolute value of each entry of $\Sigma_n^{-\frac{1}{2}}$ is bounded from above by 1. The basic idea is:

First, write $\sqrt{N_n}\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n$ as:

$$\sqrt{N_n}\mathbf{u}^T\Sigma_n^{-\frac{1}{2}}\mathbf{Z}_n = \sum_{j=1}^{N_n} c_j X_j$$

Given that entries of $\Sigma_n^{-\frac{1}{2}}$ are bounded between -1 and 1, it can be shown that $\sigma|c_j| \leq 2\sum_{i=1}^k |u_i|$. Suppose $r_n n \leq N_n \leq (r_n+1)n$. Clearly, $r_n \leq M$. Now we can find a sequence of $n+1$ integers $0 = q_{n,0} < q_{n,1} < \cdots < q_{n,n} = N_n$ such that $r_n \leq q_{n,i+1} - q_{n,i} \leq r_n + 1$. Define U_i as:

$$U_i = \sum_{j=q_{n,i-1}+1}^{q_{n,i}} c_j X_j - \sum_{j=q_{n,i-1}+1}^{q_{n,i}} c_j$$

Then it is easy to check that the Lindeberg's condition is satisfied. That is, let $s_n^2 = \mathrm{Var}(\sum_{j=1}^n U_j)$, as $n \to \infty$,

$$\sum_{i=1}^n \frac{1}{s_n^2}\int_{|U_i|>s_n\epsilon} U_i^2 dP \to 0$$

$$\sup_n \mathrm{E}\left[\left(\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left|\frac{S_n}{\sum_{i=1}^{n+r_n} X_i}\right|\right)^4\right] \le \frac{4(M+1)^2}{c^4}\sup_n \mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^4\right] < \infty$$

Therefore $\frac{(n+r_n)^3}{nr_n}\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2$ is uniformly integrable, hence its mean converges to the mean of the distributions it converges to, i.e., $\frac{\sigma^2}{\mu^2}$.

□

In the proof of Theorem 1, for convenience, we assume that $n \le r_n \le Mn$ for some M. This assumption is dropped in the following corollary.

Corollary 2 *If in Theorem 1 and Corollary 1, instead of requiring $n \le r_n \le Mn$, we require that $Ln \le r_n \le Mn$, where L is some positive real number, the conclusions still hold.*

Proof: First we note that if $r_n < n$, then:

$$\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right) = -\frac{(r_n+n)^{\frac{3}{2}}}{\sqrt{r_n n}}\left(\frac{\sum_{i=n+1}^{n+r_n} X_i}{\sum_{j=1}^{n+r_n} X_j} - \frac{r_n}{r_n+n}\right)$$

while $\frac{(r_n+n)^{\frac{3}{2}}}{\sqrt{r_n n}}\left(\frac{\sum_{i=n+1}^{n+r_n+1} X_i}{\sum_{j=1}^{n+r_n} X_j} - \frac{r_n}{r_n+n}\right)$ and $\frac{(r_n+n)^{\frac{3}{2}}}{\sqrt{r_n n}}\left(\frac{\sum_{i=1}^{r_n} X_i}{\sum_{j=1}^{n+r_n} X_j} - \frac{r_n}{r_n+n}\right)$ have the same distribution.

We also note that if $X \sim N(0,1)$, then $-X \sim N(0,1)$ too.

□

The following corollary is obvious.

Corollary 3 *In Corollary 1, if we further assume that $p_n = \frac{n}{n+r_n} \to p$, where $0 < p < 1$, then:*

$$E\left[\sqrt{n+r_n}\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)\right] \to 0$$

$$E\left[(n+r_n)\left(\frac{\sum_{i=1}^n X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2\right] \to \frac{p(1-p)\sigma^2}{\mu^2}$$

Now we can give the asymptotic distribution of the relative frequencies of multiple categories in the intermediate sample, conditional on the original sample.

Theorem 2 *Theorem 1 can be generalized in the following way:*

Given a sequence of independent nonnegative random variables $X_1, \cdots$, such that $E[X_i] = \mu > 0$, and $Var(X_i) = \sigma^2$. For $n = 1, \cdots$, let $N_{n,1}, \cdots, N_{n,k+1}$ be positive integers such that $n = N_{n,1} \le N_{n,i} \le Mn$, $i = 1, \cdots, k+1$, for some fixed M. Let $N_n = \sum_{i=1}^{k+1} N_{n,i}$, and $p_{n,i} = \frac{N_{n,i}}{N_n}$ for $i = 1, \cdots, k+1$. Define Σ_n as:

Corollary 1 *Given the same condition as in Theorem 1, if* $E[X_i^4] < \infty$, *then:*

$$E\left[\frac{(n+r_n)^3}{nr_n}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2\right] \to \frac{\sigma^2}{\mu^2}$$

Proof: From Theorem 1,

$$\frac{\mu(n+r_n)^{\frac{3}{2}}}{\sigma\sqrt{nr_n}}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right) = \frac{S_n}{s_n}\frac{\mu}{\dfrac{1}{n+r_n}\displaystyle\sum_{i=1}^{n+r_n} X_i} \Longrightarrow N(0,1)$$

hence,

$$\frac{(n+r_n)^3}{nr_n}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right)^2 \Longrightarrow \frac{\sigma^2}{\mu^2}\chi_1^2$$

Therefore it suffices to show that $\sup_n \mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^{2+\epsilon}\right] < \infty$ for some $\epsilon > 0$. Actually, we shall prove the case for $\epsilon = 2$.

$$\mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^4\right] = \frac{1}{[n+r_n]^2}\mathrm{E}\left[\left((1-p_n)\sum_{i=1}^{n} X_i - p_n\sum_{j=n+1}^{n+r_n} X_j\right)^4\right]$$

$$= \frac{1}{[n+r_n]^2}\mathrm{E}\left[\left((1-p_n)\sum_{i=1}^{n}(X_i-\mu) - p_n\sum_{j=n+1}^{n+r_n}(X_j-\mu)\right)^4\right]$$

$$= \frac{1}{[n+r_n]^2}\left\{\sum_{i=1}^{n}(1-p_n)^4\mathrm{E}\left[(X_i-\mu)^4\right] + \sum_{j=n+1}^{n+r_n} p_n^4\mathrm{E}\left[(X_j-\mu)^4\right]\right.$$

$$\left. +2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1-p_n)^4\mathrm{E}\left[(X_i-\mu)^2\right]\mathrm{E}\left[(X_j-\mu)^2\right]\right\}$$

$$\left. +2\sum_{i=n+1}^{n+r_n-1}\sum_{j=i+1}^{n+r_n} p_n^4\mathrm{E}\left[(X_i-\mu)^2\right]\mathrm{E}\left[(X_j-\mu)^2\right]\right\}$$

$$\left. +\sum_{i=1}^{n}\sum_{j=n+1}^{n+r_n} p_n^2(1-p_n)^2\mathrm{E}\left[(X_i-\mu)^2\right]\mathrm{E}\left[(X_j-\mu)^2\right]\right\}$$ [7]

$$= \frac{1}{[n+r_n]^2}\left\{n(1-p_n)^4\mathrm{E}\left[(X_1-\mu)^4\right] + r_np_n^4\mathrm{E}\left[(X_1-\mu)^4\right] + n(n-1)(1-p_n)^4\sigma^4\right.$$
$$\left. +[n+r_n][n+r_n-1]p_n^4\sigma^4 + n[n+r_n]p_n^2(1-p_n)^2\sigma^4\right\}$$
$$< \mathrm{E}\left[(X_1-\mu_1)^4\right] + \sigma^4$$

Given that $\mathrm{E}[(X_1)^4] < \infty$,

$$\sup_n \mathrm{E}\left[\left(\frac{|S_n|}{\sqrt{n+r_n}}\right)^4\right] \le \mathrm{E}\left[(X_1-\mu_1)^4\right] + \sigma^4 < \infty$$

Note that $X_i \ge c > 0$, hence $\frac{\sum_{i=1}^{n+r_n} X_i}{n+r_n} \ge c$. Also, $(n+r_n)^2 \le 2(M+1)nr_n$. It then follows that:

[7] All the other terms are 0 because $\{X_1, \cdots\}$ are independent.

$$S_n = \sum_{i=1}^{n} Y_{n,i} = \sum_{i=1}^{n} X_i - p_n \left(\sum_{j=1}^{n+r_n} X_j \right)$$

$$s_n^2 = \sum_{i=1}^{n} \mathrm{Var}(Y_{n,i}) = \mathrm{Var}\left(\sum_{i=1}^{n} Y_{n,i} \right) = [n(1-p_n)^2 + r_n p_n^2]\sigma^2 = \frac{nr_n}{n+r_n}\sigma^2 = n(1-p_n)\sigma^2$$

Let $Z_{n,i} = X_i + \sum_{j=q_{n,i-1}+n+1}^{q_{n,i-1}+n+M+1} X_j + 2\mu$. It is easy to check that $|Y_{n,i}| \leq Z_{n,i} = |Z_{n,i}|$. [5] Because of the i.i.d. property of $\{X_1, \cdots\}$, the distribution of $Z_{n,i}$ is independent of n. Indeed, the distribution of $Z_{n,i}$ is the same for $i = 1, \cdots, n$, and the same as that of $\sum_{j=1}^{M+2} X_j + 2\mu$, which is independent of n. Therefore, for any $\epsilon > 0$, as $n \to \infty$, $\int_{|Z_{n,i}|>\epsilon\sigma\sqrt{n(1-p_n)}} Z_{n,i}^2 dP \to 0$, because $Z_{n,i}^2$ is integrable, and $p_n \leq 0.5$. It then follows that:

$$\sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Y_{n,i}|>s_n\epsilon} Y_{n,i}^2 dP \leq \sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Z_{n,i}|>s_n\epsilon} Y_{n,i}^2 dP$$

$$\leq \sum_{i=1}^{n} \frac{1}{s_n^2} \int_{|Z_{n,i}|>s_n\epsilon} Z_{n,i}^2 dP = \frac{n}{n(1-p_n)\sigma^2} \int_{|Z_{n,1}|>s_n\epsilon} Z_{n,1}^2 dP$$

$$\leq \frac{2}{\sigma^2} \int_{|Z_{n,1}|>s_n\epsilon} Z_{n,1}^2 dP$$

where $\int_{|Z_{n,1}|>s_n\epsilon} Z_{n,1}^2 dP \to 0$ as $n \to \infty$

By the Central Limit Theorem,

$$\frac{S_n}{s_n} \Longrightarrow N(0,1)$$

On the other hand, by the Strong Law of the Large Number,

$$\frac{\sum_{i=1}^{n+r_n} X_i}{n+r_n} \to \mu \text{ w.p.1}$$

Consequently, [6]

$$\frac{\mu(n+r_n)^{\frac{3}{2}}}{\sigma\sqrt{nr_n}} \left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n \right) = \frac{S_n}{s_n} \frac{\mu}{\frac{1}{n+r_n} \sum_{i=1}^{n+r_n} X_i} \Longrightarrow N(0,1)$$

hence,

$$\frac{(n+r_n)^3}{nr_n} \left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n \right)^2 \Longrightarrow \frac{\sigma^2}{\mu^2}\chi_1^2$$

□

If the amplification procedure is nondecreasing and bounded, i.e., the amplification factor is bounded from below by a positive value, and also bounded from above, then it can be shown that the variance of the relative frequency of the first category also converges.

[5] Note that $X_i \geq 0$, and that $p_n[q_{n,i} - q_{n,i-1}] \leq 1$.

[6] Recall that $S_n = \sum_{i=1}^{n} X_i - p_n \sum_{j=1}^{n+r_n} X_j$.

Finally, by the identity $\binom{n}{x} = \sum_{i=0}^{x} \binom{m}{i}\binom{n-m}{x-i}$, [3] and the fact that $\sum_{i=0}^{x} i \frac{\binom{m}{i}\binom{n-m}{x-i}}{\binom{n}{x}} = \frac{m}{n}x$, [4] we have:

$$\mathrm{E}\left[P_1|X_0,Y_0\right] = \sum_{c=0}^{X_0+Y_0} \frac{1}{c+X_0+Y_0}\left(X_0 + c\frac{X_0}{X_0+Y_0}\right)\lambda^c(1-\lambda)^{X_0+Y_0-c} = \frac{X_0}{X_0+Y_0} = P_0$$

Given that $\sigma(P_0) \subset \sigma(X_0, Y_0)$, it then follows that:

$$\mathrm{E}\left[P_1|P_0\right] = \mathrm{E}\left[\mathrm{E}\left[P_1|X_0,Y_0\right]|P_0\right] = P_0$$

It is easy to see that $\{P_t\}$ for $t = 0, 1, \cdots$ is a martingale, with respect to $\{\sigma(P_0), \sigma(P_0, P_1), \cdots\}$, hence for any $r > 0$, we have:

$$\mathrm{E}[P_r|P_0] = P_0$$

□

From now on we shall make no specific assumptions about the distribution of the amplification factor. In most cases, we only assume that the amplification factor has positive mean and finite variance, as required by the definition of SAR.

To get the asymptotic distribution of the relative frequencies in the intermediate sample, we begin with a simpler case, where the original sample has two categories. Let the mean and the variance of the amplification factor be μ and σ^2 respectively, and the absolute frequencies of the first and the second categories in the original sample be n and r_n respectively. Then the following theorem gives the asymptotic distribution of the relative frequency of the first category in the intermediate sample.

Theorem 1 *Given a sequence of i.i.d. nonnegative random variables $X_1, \cdots$, such that $E[X_i] = \mu > 0$, and $Var(X_i) = \sigma^2$. Let r_n be a sequence of positive integers such that $n \leq r_n \leq Mn$ for some fixed M. Then:*

$$\frac{(n+r_n)^{\frac{3}{2}}}{\sqrt{nr_n}}\left(\frac{\sum_{i=1}^{n} X_i}{\sum_{j=1}^{n+r_n} X_j} - p_n\right) \Longrightarrow N\left(0, \frac{\sigma^2}{\mu^2}\right)$$

where $p_n = \dfrac{n}{n+r_n}$

Proof: Because $n \leq r_n \leq Mn$, for any n, there is a positive integer m_n such that $m_n n \leq r_n \leq [m_n + 1]n$, where $1 \leq m_n < M$. Moreover, we can find $n+1$ integers $0 = q_{n,0} < q_{n,1} < q_{n,2} \cdots < q_{n,n} = r_n$ such that $q_{n,i+1} - q_{n,i}$ is either m_n or $m_n + 1$. Create a triangular array of random variables $Y_{i,j}$ in the following way:

The n^{th} row of the array has n elements $Y_{n,1}, \cdots, Y_{n,n}$, where:

$$Y_{n,i} = (1-p_n)X_i - p_n\left(\sum_{j=q_{n,i-1}+n+1}^{q_{n,i}+n} X_j\right) - \left[(1-p_n) - p_n(q_{n,i} - q_{n,i-1})\right]\mu$$

It follows immediately that for each n, $Y_{n,1}, \cdots, Y_{n,n}$ are independent, $\mathrm{E}[Y_{n,i}] = 0$, and:

[3] Note that this identity holds even if $x > m$ or $n < x + m$, insofar as we observe the convention that $\binom{k_1}{k_2} = 0$ if $k_1 < k_2$.

[4] This is the mean for a drawing without replacement sample.

of a category in the final sample, based on which we can estimate the relative frequency in the population. In the last section, we give two tests for whether the same category has constant relative frequency over different populations, and argue that we could get a much higher than expected type I error rate if using the traditional tests.

3 Asymptotic Distribution of the Ratio in Amplification

If the intermediate sample in the SAR scheme were obtained by multiplying the original sample by a factor k, then the relative frequencies of each category in the intermediate sample will be the same as the relative frequencies in the original sample. However, if the original sample is amplified by a noisy procedure, say, a branch process, conditional on the original sample, the relative frequencies of each category in the intermediate sample will be nondegenerate random variables. In this section we shall present the asymptotic distribution of the relative frequencies in the intermediate sample conditional on the original sample. But, first, we would like to show that for a specific type of amplification processes, the mean of the relative frequency of any category in the intermediate sample, conditional on the original sample, is exactly the same as the relative frequency in the original sample. This specific process is often used to model the PCR procedure.

Lemma 1 *Let $\{X_t\}$ and $\{Y_t\}$ be two independent branch processes with the following properties:*

1. *$X_{t+1} = X_t + U_t$, where U_t follows a binomial distribution with parameters (X_t, λ), for $0 < \lambda < 1$.*
2. *$Y_{t+1} = Y_t + V_t$, where V_t follows a binomial distribution with parameters (Y_t, λ).*

then the conditional mean of $P_{t+1} = \dfrac{X_{t+1}}{X_{t+1} + Y_{t+1}}$ given $P_t = \dfrac{X_t}{X_t + Y_t}$ is P_t.

Proof:

Without loss of generality, let $t = 0$. The joint distribution of (X_1, Y_1), given X_0 and Y_0, is then:

$$P(X_1 = x, Y_1 = y|X_0, Y_0) = \binom{X_0}{x - X_0}\lambda^{x-X_0}(1-\lambda)^{2X_0-x}\binom{Y_0}{y - Y_0}\lambda^{y-Y_0}(1-\lambda)^{2Y_0-y}$$

where $X_0 \le x \le 2X_0$, and $Y_0 \le y \le 2Y_0$.

Let $u = x - X_0$, $v = y - Y_0$, the conditional mean of $P_1 = \dfrac{X_1}{X_1 + Y_1}$ given X_0 and Y_0 is:

$$\mathrm{E}\left[P_1|X_0, Y_0\right] = \sum_{u=0}^{X_0}\sum_{v=0}^{Y_0} \frac{u + X_0}{u + v + X_0 + Y_0}\binom{X_0}{u}\lambda^u(1-\lambda)^{X_0-u}\binom{Y_0}{v}\lambda^v(1-\lambda)^{Y_0-v}$$

Let $c = u + v$, with the convention that $\binom{k_1}{k_2} = 0$ if $k_1 < k_2$, the above formula can be written as:

$$\mathrm{E}\left[P_1|X_0, Y_0\right] = \sum_{c=0}^{X_0+Y_0}\sum_{u=0}^{c} \frac{u + X_0}{c + X_0 + Y_0}\binom{X_0}{u}\binom{Y_0}{c-u}\lambda^c(1-\lambda)^{X_0+Y_0-c}$$

$$= \sum_{c=0}^{X_0+Y_0} \frac{1}{c + X_0 + Y_0}\lambda^c(1-\lambda)^{X_0+Y_0-c}\sum_{u=0}^{c}(u + X_0)\binom{X_0}{u}\binom{Y_0}{c-u}$$

2 Sampling, Amplifying, and Resampling

The basic steps of the sampling, amplifying, and resampling (SAR) are as the following:

1. Draw the original sample, which has either the multinomial, or the multivariate hypergeometric distribution.

2. Amplify the original sample. Each element in the original sample is amplified independently such that the integer valued amplification factors for each element are nonnegative and identically distributed with positive mean and finite variance. [1] The amplified sample is called the intermediate sample.

3. Generate the final sample from the intermediate sample by drawing randomly *with* or *without* replacement. The final sample is also called the SAR sample.

Note that the generation of the final sample by sampling without replacement from the intermediate sample is a little bit tricky. The problem is that the size of the intermediate sample is a random variable, hence the size of the final sample in general will also be a random variable. For example, suppose the initial plan is to draw a sample of size n, but the size of the intermediate sample is $n' < n$, then the final sample size will be n', instead of n. However, this is less an issue in asymptotic study if n is so selected that it is less than the size of the intermediate sample with probability one.

One place where we might meet the SAR sampling scheme is in a Serial Analysis of the Gene Expression (SAGE) experiment (Velculescu, Zhang, Vogelstein, & Kinzler (1995); Velculescu, Zhang, Zhou, Traverso, St. Croix, Vogelstein, & Kinzler (2000)). In a SAGE experiment, a sample of mRNA transcripts is extracted from a tissue, transcribed into cDNA clones. Then, from a specific site of each cDNA clone, a short 10 base long sequence (tag) is cut. This sample of tags is the original sample in the SAR scheme. It could be treated either as a random sample drawing without replacement from the tissue (a finite population), hence has a multivariate hypergeometric distribution. Or, approximately, we can treat it as a multinomial sample, when the sample size is small compared to the size of the tissue.

A certain number of cycles of PCR then are performed to amplify the original sample. The PCR procedure could be modeled as a super critical simple type branch process. More precisely, suppose the count of tags at cycle i is X_i, then the count of tags at cycle $i+1$ is $X_{i+1} = X_i + Y_{i+1}$, where Y_{i+1} is a bounded nonnegative (integer valued) random variable that depends on X_i. Usually, Y_{i+1} is thought to be a binomial variable with parameters (X_i, p), where p is called the efficiency of the PCR. [2] The sample we get after the PCR is the intermediate sample.

Finally, the tags are linked together to form longer sequences. Among these longer sequences, those of certain length that are suitable for sequencing are chosen (without replacement) and get sequenced. The tags contained in the sequenced sequences are the final sample, and their counts are reported as the experimental result, called the SAGE library.

In a SAGE experiment, probably also in other experiments where the SAR scheme is used, people are mostly interested in the estimation of the relative frequencies of each category in the population (from which the original sample was drawn), and whether the relative frequency of a category is constant in two or more populations. In the next section, we shall first study the asymptotic behavior of the amplifying step. Then in section 4, we present the main result of this paper, the asymptotic distribution of the count/relative frequency

[1] Starting with a single element, let X be the total number of elements obtained after the amplification, then X is the amplification factor.

[2] For simplification, the mutation during PCR is ignored in this model. For a more complicated model, see Sun (1999).

Sampling, Amplifying, and Resampling

Tinajiao Chu

Abstract

We discuss a new sampling method, sampling, amplifying, and resampling (SAR), for generating discrete data, and derive the asymptotic distribution of the SAR sample. A new statistic for the test of association is given, and its asymptotic distributions is derived. We compare the new model with the traditional multinomial model, and show that the new model predicts a significantly larger variance for the SAR sample. This implies that, when applied to the SAR sample, the tests based on the traditional model will have a much higher type I error than expected. This new model can be applied to biological experiments involving Polymerase Chain Reaction (PCR).

KEY WORDS: Sampling, Amplification, PCR, Asymptotic, Association

1 Introduction

In their classic work on multivariate discrete analysis, Bishop, Fienberg, and Holland (1975) discuss several popular sampling methods that generate multivariate discrete data (contingency tables). Basically, there are two types of sampling methods: the multinomial type, and the hypergeometric type. The multinomial type methods again include sampling methods that could generate the following three families of distributions: the multinomial sampling, which generates data of multinomial distributions; the Poisson sampling, which generates data of Poisson distributions; and the negative multinomial sampling, which generate data of negative multinomial distributions. These sampling methods are closely related to each other. For example, among the three multinomial type methods, the joint distribution of k independent Poisson random variables, conditional on their sum, is a k dimensional multinomial distribution. The multinomial distribution, on the other hand, could be seen as the limit of the multivariate hypergeometric distribution, and is often a good approximation for the latter when the population size is large compared to the sample size.

Of course, these are not the only distributions a multivariate discrete sample could have. But because of the popularity of the above models, people may tend to treat any contingency table as being generated by one of these methods, which could be problematic when the true distribution is quite different. In this paper, we introduce a new sampling scheme, i.e., the sampling, amplifying, and resampling scheme (SAR). SAR scheme could be found in current genetic study, where researchers use PCR (Polymerase Chain Reaction) to amplify the original sample from a tissue, and then perform some experiment. In the following sections, we shall illustrate the basic idea of this sampling scheme, find out the asymptotic distribution of the data generated by this sampling scheme, present test statistics for some frequently used tests, and give the asymptotic distributions for these statistics.