References

1. O. Bretscher, *Linear Algebra with Appications* (3rd ed.), Pearson Prentice Hall, 2005.
2. H. J. Larson, *Introduction to Probability Theory* (2nd ed.), Wiley, 1974.

Exhaustive sampling and related binomial identities

Jim Ridenhour (ridenhourj@apsu.edu) and David Grimmett (grimmettd@apsu.edu), Austin Peay State University, Clarksville, TN 37044

There are many situations that involve repeated sampling from the same set of observations. For example, suppose a professor has a test bank of 100 questions for a particular course and randomly chooses 25 of these questions for the final exam each semester. A persistent but not very talented student repeats the course several times. Obviously, the student has no chance of having seen all the questions before taking the course four times. What is the probability that the student will have seen all the questions after k repetitions? That is, what is the probability that the entire test bank will have been exhausted after k repetitions?

A more practical example involves drug testing. Suppose, for example that a bicycle race has 100 contestants and consists of several stages where random samples of 20 contestants are taken at each stage and screened for banned substances. If the race has 10 stages, what is the probability that each contestant will be tested for banned substances at least once?

Probability of exhaustion. We will assume that we are selecting k samples of size n from a population containing N members. We want to find the probability of the event E that the population has been exhausted in the k samples. That is, every member of the population has been included in at least one sample. Denote the members of the population by $x_1, x_2, \ldots, x_N$. We will calculate the probability of the complementary event E^C. Let E_i be the event that x_i has not been included in any of the k samples. Then $E^C = E_1 \cup E_2 \cup \cdots \cup E_N$. By the addition law of probability and the method of inclusion and exclusion,

$$P(E^C) = \sum_{i=1}^{N} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{p<q<r} P(E_p \cap E_q \cap E_r) - \cdots, \quad (1)$$

where the last sum consists of all terms with $N - n$ intersections. That is because each sample has n distinct elements and so the greatest number of elements that cannot be included is $N - n$. For a particular x_i, let B_{ij} be the event that x_i is not included in the jth sample. Then $E_i = B_{i1} \cap B_{i2} \cap \cdots \cap B_{ik}$. Moreover,

$$P(B_{ij}) = \frac{\binom{N-1}{n}}{\binom{N}{n}}$$

since we must choose a sample from N elements without including x_i. Since the samples are chosen with replacement, the events $B_{i1}, B_{i2}, \ldots, B_{ik}$ are independent,

and so

$$P(E_i) = P(B_{i1} \cap B_{i2} \cap \cdots \cap B_{ik}) = \prod_{j=1}^{k} P(B_{ij}) = \prod_{j=1}^{k} \frac{\binom{N-1}{n}}{\binom{N}{n}} = \left[\frac{\binom{N-1}{n}}{\binom{N}{n}}\right]^k. \quad (2)$$

Consider the first term in (1). There are $\binom{N}{1}$ ways to choose i, and each $P(E_i)$ has the probability given in (2), so this first term has the value

$$\binom{N}{1}\left[\frac{\binom{N-1}{n}}{\binom{N}{n}}\right]^k.$$

Next, consider the second term in (1). There are $\binom{N}{2}$ ways to choose the integers i and j between 1 and N with $i < j$. Also $E_i \cap E_j$ means that neither x_i nor x_j is in any of the k independent samples. But the probability that neither x_i nor x_j is in any of these independent samples is $\binom{N-2}{n}/\binom{N}{n}$. Since the samples are independent, the probability that neither x_i nor x_j is in any of the k independent samples is $\left[\binom{N-2}{n}/\binom{N}{n}\right]^k$. Consequently, the second term of the sum (1) is $\binom{N}{2}\left[\binom{N-2}{n}/\binom{N}{n}\right]^k$. Each of the later terms can be analyzed in a similar manner, and therefore, (1) can be rewritten as

$$P(E^C) = \binom{N}{1}\left[\frac{\binom{N-1}{n}}{\binom{N}{n}}\right]^k - \binom{N}{2}\left[\frac{\binom{N-2}{n}}{\binom{N}{n}}\right]^k + \binom{N}{3}\left[\frac{\binom{N-3}{n}}{\binom{N}{n}}\right]^k - \cdots$$
$$+ (-1)^{N-n+1}\binom{N}{N-n}\left[\frac{\binom{N-(N-n)}{n}}{\binom{N}{n}}\right]^k.$$

Simplifying, we get

$$P(E^C) = \frac{\binom{N}{1}\binom{N-1}{n}^k - \binom{N}{2}\binom{N-2}{n}^k + \binom{N}{3}\binom{N-3}{n}^k - \cdots + (-1)^{N-n+1}\binom{N}{N-n}\binom{n}{n}^k}{\binom{N}{n}^k}. \quad (3)$$

Then the probability of exhaustion is $P(E) = 1 - P(E^C)$.

We again consider the example where a professor has a test bank of 100 questions and randomly chooses 25 for the final exam each semester and a persistent student continues to repeat the course each semester. What is the probability of the event E that the student has seen all 100 questions after taking the course k times? As we noted earlier, the student must take the course at least four times to have any chance of having seen all the questions. However, the probability of having seen them all in just four repetitions is only

$$P(E) = \frac{\binom{75}{25}\binom{50}{25}\binom{25}{25}}{\binom{100}{25}^3} = 4.66 \times 10^{-37},$$

so it is extremely unlikely that this will occur. It is not hard to write a program to calculate $P(E)$ but care must be taken to avoid overflow and underflow errors due to the nature of the numbers involved. The following table gives the results up to $k = 30$

when $N = 100$ and $n = 25$ (the numbers have been rounded off to eight decimal places). From this table, we see that students must be *very* persistent if they want at least a fifty-fifty chance of seeing all of the questions.

k	$P(E)$	k	$P(E)$
4	0.00000000	18	0.56269617
5	0.00000000	19	0.65090382
6	0.00000000	20	0.72545763
7	0.00000001	21	0.78656843
8	0.00000294	22	0.83553496
9	0.00011925	23	0.87411444
10	0.00150604	24	0.90413151
11	0.00890754	25	0.92726961
12	0.03159226	26	0.94498164
13	0.07868290	27	0.95846997
14	0.15278444	28	0.96870217
15	0.24836613	29	0.97644193
16	0.35514273	30	0.98228379
17	0.46257357		

Associated binomial identities. If $kn < N$, then it is impossible to have sampled all N members of the population with k samples of size n. Consequently, $P(E^C) = 1$ and the numerator of (3) must equal the denominator. This yields the following family of identities:

$$\binom{N}{1}\binom{N-1}{n}^k - \binom{N}{2}\binom{N-2}{n}^k + \binom{N}{3}\binom{N-3}{n}^k - \cdots + (-1)^{N-n+1}\binom{N}{N-n}\binom{n}{n}^k = \binom{N}{n}^k.$$

Rewritten with summation notation, this is

$$\sum_{j=1}^{N-n} (-1)^{j+1}\binom{N}{j}\binom{N-j}{n}^k = \binom{N}{n}^k. \tag{4}$$

These identities hold for any positive integer k for which $kn < N$. As an example, we consider the case $N = 7$ and $n = 2$. Part of Pascal's triangle is given below, and the numbers involved in the identities are in boldface type. In this case, (4) holds for $k = 1$, 2, and 3.

	n							
N	0	1	2	3	4	5	6	7
0	1							
1	1	1						
2	1	2	**1**					
3	1	3	**3**	1				
4	1	4	**6**	4	1			
5	1	5	**10**	10	5	1		
6	1	6	**15**	20	15	6	1	
7	1	**7**	**21**	**35**	**35**	**21**	7	1

Here $\binom{N}{n} = 21$. The first factors of the terms in the sum on the left side of (4) start with $\binom{N}{1} = 7$ and progress to the right along the row for $N = 7$. The second factors of these terms are raised to the kth power and begin with $\binom{N-1}{2} = 15$ and progress up the column for $n = 2$. The numerical values for the identities in this example for $k = 1, 2$, and 3 are given below.

$$7 \cdot 15 - 21 \cdot 10 + 35 \cdot 6 - 35 \cdot 3 + 21 \cdot 1 = 21$$
$$7 \cdot 15^2 - 21 \cdot 10^2 + 35 \cdot 6^2 - 35 \cdot 3^2 + 21 \cdot 1^2 = 21^2$$
$$7 \cdot 15^3 - 21 \cdot 10^3 + 35 \cdot 6^3 - 35 \cdot 3^3 + 21 \cdot 1^3 = 21^3$$

The number of identities in the family is determined by how small n is relative to N. For example, if $N = 30$ and $n = 4$, then (4) holds for $k \leq 7$. The general relationship of each identity with respect to Pascal's triangle is the same as in the example. The first and second factors for the terms on the left side of (4) are found by starting with $\binom{N}{1}$ and moving to the right for the first factor and starting with $\binom{N-1}{2}$ and moving upwards for the second factor. The extensive literature on binomial coefficients has identities similar to the case where $k = 1$:

$$\sum_{j=1}^{N-n} (-1)^{j+1} \binom{N}{j} \binom{N-j}{n} = \binom{N}{n}.$$

For example, the reader is referred to the first chapter of Riordan's classic book, *Combinatorial Identities*. However, these identities do not have terms where factors are raised to an arbitrary power k as is the case in (4). The identity (4) is interesting in that it holds for all positive integers less than N/n. This allows us to write identities that hold for any number of consecutive integers but not beyond. For example, if $N = 1000$ and $n = 10$, then (4) holds for all $k \leq 99$ but not for any values beyond 99.

References

1. J. Riordan, *Combinatorial Identities*, Wiley, 1968.

Controlling the discrepancy in marginal analysis calculations

Michael W. Ecker (DrMWEcker@aol.com), Pennsylvania State University, Wilkes-Barre Campus, Lehman, PA 18627

Despite technology, we professors still love our little tricks in designing shortcuts and problems involving "nice numbers" that lead to easily predictable outcomes. Here's one such shortcut that I discovered recently.

Consider the typical case of a quadratic cumulative-cost function, often encountered in Calculus I and "Business Calculus" as a differentiation application. Here a hypothetical business produces x "widgets" for a total cost of $C(x) = ax^2 + bx + c$. (In order for $C(x)$ to be increasing, we restrict our attention to $0 \leq x \leq -b/2a$ with