Probability Issues in Without Replacement Sampling

Anwar H. Joarder and Walid S. Al-Sabah
King Fahd University of Petroleum & Minerals

Key Words: *probability; sampling without replacement; combinatorial method; tree diagram; conditional probability*

Abstract

Sampling without replacement is an important aspect in teaching conditional probabilities in elementary statistics courses. Different methods scattered in different texts for calculating probabilities of events in this context are reviewed and their relative merits and limitations in applications are pinpointed. An alternative representation of hypergeometric distribution resembling binomial distribution may often provide more insight into the problems often encountered.

1. Introduction

Suppose that we are interested in the number of a kind of items say x ash colored balls in a sample of n units drawn from a lot containing $N = a + b$ balls, of which there are a ash colored balls and b blue colored balls. Let A_i denote the event of having an ash ball at i th draw, B_i denote the event of having a blue ball at i th draw. For a sample of size $n = 3$, the sample space would be

$$S = \{A_1A_2A_3,\ A_1A_2B_3,\ A_1B_2A_3,\ A_1B_2B_3,\ B_1A_2A_3,\ B_1A_2B_3,\ B_1B_2A_3,\ B_1B_2B_3\}.$$

Let us assume that the items be drawn with replacement. Then the probability of $A_1A_2B_3$ is given by

$$\begin{aligned} P(A_1A_2B_3) &= P(A_1)P(A_2 \mid A_1)P(B_3 \mid A_1A_2) \\ &= \frac{a+0}{a+b}\ \frac{a+0}{a+b}\ \frac{0+b}{a+b} \\ &= \left(\frac{a}{a+b}\right)^2 \frac{b}{a+b}. \end{aligned} \tag{1.1}$$

Since each of the $\binom{3}{2} = 3$ sample points namely $\{A_1A_2B_3,\ A_1B_2A_3,\ B_1A_2A_3\}$ have the same probability, the probability of having 2 ash balls in a sample of size 3 drawn with replacement is given by

$$P(X=2)=\binom{3}{2}\left(\frac{a}{a+b}\right)^2\frac{b}{a+b} \tag{1.2}$$

where X is the number of ash balls in the sample. For a sample of size n, the probability that all the balls in the sample are ash is given by

$$P(X=n)=\left(\frac{a}{a+b}\right)^n$$

while the probability that all the balls in the sample are blue is given by

$$P(Y=n)=\left(\frac{b}{a+b}\right)^n$$

where Y is the number of blue balls in the sample. In general, let us have a sequence $A_1A_2\cdots A_xB_{x+1}\cdots B_n$ of x ash balls and $n-x$ blue balls. Then the probability of this sequence would be

$$P\left(\underbrace{A_1A_2\cdots A_x}_{x}\underbrace{B_{x+1}\cdots B_n}_{n-x}\right)=\frac{a+0}{a+b}\ \frac{a+0}{a+b}\cdots\frac{a+0}{a+b}\ \frac{0+b}{a+b}\ \frac{0+b}{a+b}\cdots\frac{0+b}{a+b},$$

i.e. $P(A_1A_2\cdots A_xB_{x+1}\cdots B_n)=\left(\frac{a}{a+b}\right)^x\left(\frac{b}{a+b}\right)^{n-x}$. (1.3)

Since x ash balls (from a ash balls) and $n-x$ blue balls (from b blue balls) can happen in $\binom{n}{x}$ sequences, the probability of any outcome of this type is given by

$$P(X=x)=\binom{n}{x}\left(\frac{a}{a+b}\right)^x\left(\frac{b}{a+b}\right)^{n-x},(x=0,1,\cdots,n). \tag{1.4}$$

Note that while (1.3) is the probability of a sample point, (1.4) is that of a compound event where each of the sample points in the compound event has the same probability as given by (1.3). Both the events in (1.3) and (1.4) are important in elementary courses in statistics.

When the sampling is without replacement in the above situations, a close analogue of (1.4) is not noticed in any text book. The investigation resulted in an alternative derivation of hypergeometric probability function presented in Section 3. Since there are some other ways scattered in different textbooks, first we review them. It appears that the methods discussed in Section 2 may be preferred by readers not acquainted with conditional probabilities (cf. Barnett, 1998, 182).

2. Unconditional Probability Approach

(i) Labeling Method

Let there be $a+b$ distinguishable items in the lot. They are labeled and n items are drawn without replacement. In case the items in the lot are indistinguishable, they may be labeled to make them distinguishable to use the method.

Suppose that there are three ash balls and two blue balls in an urn. If the balls are indistinguishable, we cannot use the **Labeling Method** unless we label the balls of the same kind. Let A^1, A^2 and A^3 be ash balls while B^1 and B^2 be blue balls. We want to draw 1 ball from them in succession without replacement. Then the sample space would be $S=\{A^1,A^2,A^3,\ B^1,B^2\}$. The probability that A^1 is selected in the sample would be $P(A^1)=1/5$. In fact the probability that A^i $(i=1,2,3)$ is selected in the sample is $1/5$, while the probability that B^i $(i=1,2,3)$ is selected in the sample is also $1/5$. But the probability that an ash ball is selected would be $P(A)=P(A^1)+P(A^2)+P(A^3)=3/5$, while the probability that a blue ball will be selected is $P(B)=P(B^1)+P(B^2)=2/5$.

Example 2.1 Suppose that there are three males and two females in a family. You want to invite three of them for a dinner party. What is the probability that two males are invited?

Let M^1, M^2 and M^3 be males while F^1 and F^2 be females. Then the sample space would be

$$S=\{M^1M^2M^3,\ M^1M^2F^1,\ M^1M^2F^2,\ M^1M^3F^1,\ M^1M^3F^2,$$
$$M^1F^1F^2,\ M^2M^3F^1,\ M^2M^3F^2,\ M^2F^1F^2,\ M^3F^1F^2\}.$$

Assuming that each sample point is equally likely, the probability that the male M^i $(i=1,2,3)$ is selected in the sample is $6/10$ and the probability that female F^i is selected in the sample is $6/10$. Thus, each element in the lot has the chance of being selected in the sample.

Assuming that each sample point is equally likely, the probability that two males will be invited is given by

$$P(M^1M^2F^1)+P(M^1M^2F^2)+P(M^1M^3F^1)$$
$$+P(M^1M^3F^2)+P(M^2M^3F^1)+P(M^2M^3F^2)=6/10.$$

The number 10 in the denominator of the above probability is the number of elements in the sample space. In case the sample size increases, it is difficult to label the sample points. This is why we have the following equivalent method.

(ii) Combinatorial Method

The Combinatorial Method, better known as Hypergeometric Method in this case, is a mathematical way of counting the groups or combinations where the items in the lot are distinguishable.

A general proof of the problem encountered in the introduction when the sampling is without replacement is outlined here for some avid readers. The x ash balls can be chosen (from a ash balls in the lot) in $\binom{a}{x}$ ways, and the $(n-x)$ blue balls can be chosen (from b blue balls in the lot) in $\binom{b}{n-x}$ ways. Hence the x ash balls and $(n-x)$ blue balls can be chosen in the sample in ways $\binom{a}{x}\binom{b}{n-x}$ ways. Also n balls can be chosen from $a+b$ items in $\binom{a+b}{n}$ ways. If we consider the possibilities as equally likely, the probability of having x ash balls and $(n-x)$ blue balls in a sample of size n is given by

$$P(X=x)=\binom{a}{x}\binom{b}{n-x}\div\binom{a+b}{n}, \quad x=0,1,\cdots,n$$

where $x \le a$ and $n-x \le b$, i.e., $\max\{0,\ n-b\} \le x \le \min\{a,\ n\}$ (Rohatgi, 1984, 336).

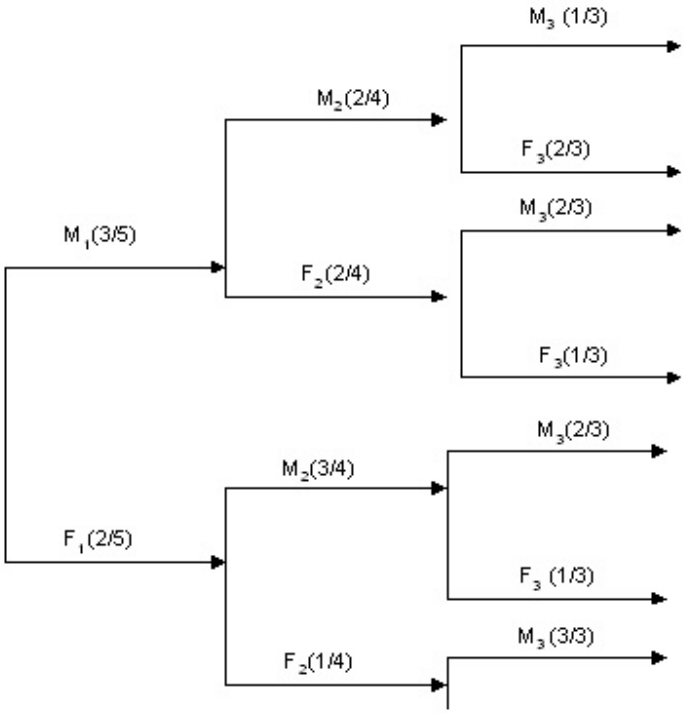
3. Conditional Probability Approach

In most elementary textbooks the Tree Diagram Method is used to calculate the probability of a simple event or any compound event. However, it is difficult to fit a tree diagram in a A-4 sized paper when the sample size increases. While calculating probability, we show a breakdown of items in the lot size and also in the sample size allowing us to solve any problem with relatively larger sample size in an efficient manner without having recourse to the Tree Diagram itself. What we have discussed in the introduction is actually an example of Tree Diagram Method though the sampling method is with replacement.

Example 3.1 The Example 2.1 is a natural example for the **Tree Diagram Method.** Let M_i be the event that in the i^{th} $(i=1,2)$ selection without replacement we have a male and F_j be the event that in the j^{th} $(j=1,2)$ th selection we have a female. Then by the Tree Diagram Method the sample space is given by

$$S=\{M_1M_2M_3,\ M_1M_2F_3,\ M_1F_2M_3,\ M_1F_2F_3,\ F_1M_2M_3,\ F_1M_2F_3,\ F_1F_2M_3\}.$$

Note that sample space does not include $F_1F_2F_3$ as we have only two females. A tree diagram is provided below:



Tree Diagram for Example 3.1

Since

$$P(M_1M_2F_3)=\frac{3+0}{3+2}\times\frac{2+0}{2+2}\times\frac{0+2}{1+2}=\frac{2}{10},$$
$$P(M_1F_2M_3)=\frac{3+0}{3+2}\times\frac{0+2}{2+2}\times\frac{2+0}{2+1}=\frac{2}{10},$$
$$P(F_1M_2M_3)=\frac{0+2}{3+2}\times\frac{3+0}{3+1}\times\frac{2+0}{2+1}=\frac{2}{10},$$

the probability that two males are invited in the sample of 3 persons is given by

$P(M_1M_2F_3)+P(M_1F_2M_3)+P(F_1M_2M_3)=6/10$.

We now generalize the Tree Diagram method. With the notations in the introduction, the probability of selecting x ash balls (from a ash balls) and $n-x$ blue balls (from b blue balls) under sampling without replacement, we proceed as follows. Let A_i denote the event of having an ash ball at i th draw, B_i denote the event of having a blue ball at i th draw. For a sample of size $n=3$ without replacement, the sample space would be

$$S=\{A_1A_2A_3,\ A_1A_2B_3,\ A_1B_2A_3,\ A_1B_2B_3,\ B_1A_2A_3,\ B_1A_2B_3,\ B_1B_2A_3,\ B_1B_2B_3\}.$$

Then the probability of having 2 ash balls in the sample is given by

$$\begin{aligned} P(A_1A_2B_3) &= P(A_1)P(A_2\mid A_1)P(B_3\mid A_1A_2) \\ &= \frac{a+0}{a+b}\,\frac{(a-1)+0}{(a-1)+b}\,\frac{0+b}{(a-2)+b} \\ &= \frac{(a+0)^{\{2\}}}{(a+b)^{\{2\}}}\,\frac{b}{(a-2)+b} \end{aligned} \tag{3.1}$$

where $a^{\{x\}}=a(a-1)\cdots(a-x+1)$. The equation (3.1) is an analogue for (1.1). Also since each of the $\binom{3}{2}=3$ sample points namely $\{A_1A_2A_3,\ A_1A_2B_3,\ B_1A_2A_3\}$ have the same probability, the probability of having two ash balls in a sample of size 3 is given by

$$P(X=2)=\binom{3}{2}\,\frac{(a+0)^{\{2\}}}{(a+b)^{\{2\}}}\,\frac{b}{(a-2)+b} \tag{3.2}$$

where X is the number of ash balls in a sample of size 3. This is an analogue of (1.2)

It is easy to prove that the probability of having an ash ball in a particular draw, say, in the 2nd draw, is given by $P(A_1A_2A_3)+P(A_1A_2B_3)+P(B_1A_2A_3)+P(B_1A_2B_3)=a/N$. For a sample of size n the probability of having all ash balls in the sample is given by

$$P(X=n)=\frac{(a+0)^{\{n\}}}{(a+b)^{\{n\}}}$$

while the probability of having all blue balls in the sample is given by

$$P(Y=n)=\frac{(0+b)^{\{n\}}}{(a+b)^{\{n\}}}$$

where Y is the number of blue balls in a sample.

In general, let us have the sequence of $A_1A_2\cdots A_xB_{x+1}\cdots B_n$ of x ash balls and $n-x$ blue balls. Then the probability of this sequence would be

$$P(A_1A_2\cdots A_x B_{x+1}\cdots B_n)$$
$$=\frac{a+0}{a+b}\ \frac{(a-1)+0}{(a-1)+b}\cdots\frac{(a-x+1)+0}{(a-x+1)+b}$$
$$\times\frac{0+b}{(a-x)+b}\ \frac{0+(b-1)}{(a-x)+(b-1)}\cdots\frac{0+\big(b-(n-x)+1\big)}{(a-x)+(b-(n-x)+1)},$$
$$\text{i.e. } P(A_1A_2\cdots A_x B_{x+1}\cdots B_n)=\frac{(a+0)^{\{x\}}}{(a+b)^{\{x\}}}\frac{(0+b)^{\{n-x\}}}{(a-x+b)^{\{n-x\}}} \tag{3.3}$$

which is analogous to (1.3). Since x ash balls from (a ash balls) and $n-x$ blue balls (from b blue balls) can happen in $\binom{n}{x}$ sequences, the probability of any outcome of this type is given by

$$P(X=x)=\binom{n}{x}\ \frac{{}_aP_x}{{}_NP_x}\ \frac{{}_bP_{n-x}}{{}_{N-x}P_{n-x}}=\binom{n}{x}\frac{(a+0)^{\{x\}}}{(a+b)^{\{x\}}}\frac{(0+b)^{\{n-x\}}}{(a-x+b)^{\{n-x\}}} \tag{3.4}$$

which can also be written as

$$P(X=x)=\binom{n}{x}\ \frac{{}_aP_x}{{}_NP_x}\ \frac{{}_bP_{n-x}}{{}_{N-x}P_{n-x}}=\binom{n}{x}\frac{(a+0)^{\{x\}}}{(a+b)^{\{x\}}}\frac{(0+b)^{\{n-x\}}}{(a-x+b)^{\{n-x\}}}$$

where X is the number of ash balls in a sample of size n selected without replacement from the lot of size $a+b=N$ and ${}_aP_x=a(a-1)\cdots(a-x+1)$. This is analogous to (1.4). In case $a\to\infty, N\to\infty$ as $a/N\to p$, $(0<p<1)$, then (3.4) can be approximated by binomial probability function (See Appendix).

Since ${}_nP_x=\frac{n!}{(n-x)!}$ and $\binom{n}{x}=\frac{n!}{x!(n-x)!}=\frac{{}_nP_x}{x!}$, the above probability in (3.4) can be represented by

$$P(X=x)=\binom{n}{x}\ \frac{{}_aP_x}{{}_NP_x}\ \frac{{}_bP_{n-x}}{{}_{N-x}P_{n-x}}$$
$$=\frac{n!}{x!(n-x)!}\ \frac{{}_aP_x\ {}_bP_{n-x}}{{}_NP_n}$$
$$=\frac{{}_aP_x}{x!}\ \frac{{}_bP_{n-x}}{(n-x)!}\div\frac{{}_{a+b}P_n}{n!},$$

$$\text{i.e. } P(X=x)=\binom{a}{x}\binom{b}{n-x}\div\binom{a+b}{n} \tag{3.5}$$

which is the well known formula for the hypergeometric distribution (Johnson, 2005, 110) . We highlight below why the use of (3.4) instead of (3.5) would provide more insight into conditional probabilities and often result in less mistakes by students.

Example 3.2 Suppose that a shipment of 9 digital voice recorders contains 4 that are defective. If n recorders are randomly chosen without replacement for inspection, what is the probability that

(i) the first two of $n=3$ checked will be defective but the third one will be non-defective?
(ii) 2 of the $n=3$ recorders will be defective?
(iii) 3 of the $n=5$ recorders will be defective?

Solution:

(i) The probability is $P(D_1D_2D_3') = \frac{4+0}{4+5}\ \frac{3+0}{3+5}\ \frac{0+5}{2+5} = \frac{5}{42}$.

Students, who memorize combinatorial rule, erroneously calculate the probability to be $P(X=2) = \binom{4}{2}\binom{5}{1} \div \binom{9}{3} = \frac{5}{14}$.

(ii) Since $4+5=a+b,\ 2+1=x+(n-x)$, the probability that 2 of the 3 voice recorders will be defective is given by

$$\begin{aligned} P(X=2) &= P(D_1D_2D_3') + P(D_1D_2'D_3) + P(D_1'D_2D_3) \\ &= \left(\frac{4+0}{4+5}\times\frac{3+0}{3+5}\times\frac{0+5}{2+5}\right) + \left(\frac{4+0}{4+5}\times\frac{0+5}{3+5}\times\frac{3+0}{3+4}\right) + \left(\frac{0+5}{4+5}\times\frac{4+0}{4+4}\times\frac{3+0}{3+4}\right) \\ &= \frac{5}{14}. \end{aligned}$$

The above insight of conditional probabilities and the number of sequences resulting 2 defectives is prominent in

$$P(X=2) = \binom{3}{2} \times \frac{(4+0)^{\{2\}}}{(4+5)^{\{2\}}} \times \frac{0+5}{(4-2)+5}$$

but not in the combinatorial method

$$P(X=2) = \binom{4}{2}\binom{5}{1} \div \binom{9}{3}.$$

(iii) Since $4+5=a+b,\ 3+2=x+(n-x)$, by the representation (3.4), the probability that 3 of the 5 voice recorders will be defective is given by

$$P(X=3)=\binom{5}{3}\times\frac{(4+0)^{\{3\}}}{(4+5)^{\{3\}}}\times\frac{(0+5)^{\{2\}}}{(4-3+5)^{\{2\}}}$$
$$=10\times\frac{4(4-1)(4-2)}{9(9-1)(9-2)}\times\frac{5(5-1)}{6(6-1)}$$

i.e. $P(X=3)=\frac{20}{63}$.

Though the above can be quickly calculated by the combinatorial method as

$$P(X=3)=\binom{4}{3}\binom{5}{2}\div\binom{9}{5}=\frac{20}{63},$$

it does not provide insights of conditional probabilities which is expected in sampling without replacement.

4. Conclusion

Students with little background in combinatorics, say in elementary schools, may use the Labeling Method provided the lot size or the sample size is small. Most readers would find the Tree Diagram Method suitable though it can be increasingly difficult to fit in a A4 sized paper as the sample size exceeds 3. The multiplication rule in the Tree Diagram Method, though seems obvious, provides good insight into conditional probability. The breakdown of the lot size and the sample size in the numerator and denominator of (3.4) is easily understood by students, and can be done with or without tree diagrams. The combinatorial method works regardless of the lot size or the sample size and whether the items are distinguishable or not, and would be certainly preferred by readers not acquainted with conditional probabilities (cf. Barnet, 1998, 182). Though the probability function in (3.4) and (3.5) are algebraically the same, the former provides insight into conditional probabilities, and much analogous to the Binomial Probability function in (1.4).

Appendix

In case $a\to\infty, N\to\infty$ as $a/N\to p$, $(0<p<1)$, then the second term in (3.4) can be written as

$$\frac{a^{\{x\}}}{N^{\{x\}}}=\frac{a(a-1)\cdots(a-x+1)}{N(N-1)\cdots(N-x+1)}$$
$$=\frac{a}{N}\frac{(a-1)/N}{(N-1)/N}\cdots\frac{(a-x+1)/N}{(N-x+1)/N}.$$

Each of the above x ratios will tend to p so that

$$\frac{a^{\{x\}}}{N^{\{x\}}} \to p^x .$$

Similarly the third term in (3.4) can be written as

$$\begin{aligned}\frac{b^{\{n-x\}}}{N-x^{\{n-x\}}} &= \frac{b(b-1)\cdots(b-(n-x)+1)}{(N-n)(N-(n-1))\cdots(N-n+1)} \\ &= \frac{b/N}{(N-x)/N}\frac{(b-1)/N}{(N-x-1)/N}\cdots\frac{(b-(n-x)+1)/N}{(N-n-1)/N}, \\ &= \frac{1-a/N}{(N-x)/N}\frac{1-(a+1)/N}{(N-x-1)/N}\cdots\frac{1-(a+n-x-1)/N}{(N-n-1)/N}\end{aligned}$$

which will tend to $(1-p)^{n-x}$. Thus (3.4) will have a binomial distribution in the limit with probability function

$$P(X=x)=\binom{n}{x}p^x(1-p)^x,\ (x=0,1,\cdots,n;0<p<1).$$

Acknowledgements

The authors are thankful to two referees and the Associate Editor for constructive suggestions that have improved the quality of the paper significantly. The authors also gratefully acknowledge the excellent research support provided by King Fahd University of Petroleum and Minerals, Saudi Arabia.

References

Barnett, S. (1998). *Discrete Mathematics: Numbers and Beyond.* Essex, England: Pearson Education Limited.

Johnson, R. (2005). Miller and Freund's Probability and Statistics for Engineers. New Jersey, USA: Pearson Educational International.

Lapin, L.L. (1997). *Modern Engineering Statistics.* New York, USA: International Thompson Publishing Co.

Rohatgi, V.K. (1984). *Statistical Inference.* New York, USA: John Wiley and Sons.

Anwar H. Joarder and Walid S. Al-Sabah

Department of Mathematical Sciences

King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia
Emails: anwarj@ kfupm.edu.sa, walid@kfupm.edu.sa

File: Pedastat\ Hypergeometricpa.doc