Chapter 4

Hypergeometric Distribution

The hypergeometric distribution plays a key role in statistical auditing. This chapter describes some important properties of the hypergeometric distribution we use in subsequent chapters. Section 4.1 will give some elementary properties of the hypergeometric probability. This section also gives some properties of the hypergeometric distribution function and quotients of hypergeometric distribution functions. These properties will be very helpful in Chapter 5. Section 4.2 gives exact and approximate confidence intervals for the probability that a certain characteristic is present within a population. Finally Section 4.3 shows how we can calculate hypergeometric probabilities in an efficient and accurate way. This section is essential to Chapter 6.

4.1 Properties of the hypergeometric distribution

Consider a population of N elements. A number of these N elements may have a certain characteristic that we are interested in, e.g. the number of travel declarations in a yearly population which were processed incorrectly. We will denote this number by M. In auditing applications this characteristic is often unwanted, and therefore the value of M is relatively small. This number is not known to us in advance. To get more information about M, a random sample of size n is taken. The sample contains K elements that have the characteristic of interest. The number K in the sample follows a hypergeometric distribution with parameters n, M, and N. We write $K \sim \mathcal{H}(n, M, N)$.

We use a well-known extended definition of the binomial coefficients, that will be very convenient in our algebraic manipulations with the hypergeometric distribution. Recall that $\binom{p}{q} = \frac{p!}{q!\,(p-q)!}$ for $q = 0, 1, \ldots, p;\ p = 0, 1, 2, \ldots$, where $0! = 1$ by definition. For other values of $p, q \in \mathbb{Z}$ it is defined $\binom{p}{q} = 0$. Using these notations we do not have to incorporate the usual domain for K, namely $K = 0, \ldots, n$ with the restriction that $K \geq n - (N - M)$ and $K \leq M$. Thus for non-negative integers k we have

$$\mathrm{P}\{K = k|n, M, N\} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}. \tag{4.1.1}$$

We know that

$$\mathrm{E}(K) = n \cdot \frac{M}{N}$$

and

$$\mathrm{Var}(K) = \frac{n \cdot M \cdot (N - M) \cdot (N - n)}{N^2 \cdot (N - 1)}.$$

The following properties for the hypergeometric distribution hold. We refer to Lieberman and Owen (1961).

Property 4.1.1. *The hypergeometric distribution has the following elementary properties:*

$$\mathrm{P}\{K = k + 1|n, M, N\} = \frac{(M - k) \cdot (n - k)}{(k + 1) \cdot (N - M - n + k + 1)} \times \\ \times \mathrm{P}\{K = k|n, M, N\}$$

$$\mathrm{P}\{K = k|n + 1, M, N\} = \frac{(n + 1) \cdot (N - M - n + k)}{(n + 1 - k) \cdot (N - n)} \times \\ \times \mathrm{P}\{K = k|n, M, N\}$$

$$\mathrm{P}\{K = k|n, M + 1, N\} = \frac{(M + 1) \cdot (N - M - n + k)}{(M + 1 - k) \cdot (N - M)} \times \\ \times \mathrm{P}\{K = k|n, M, N\}$$

$$\mathrm{P}\{K = k|n, M, N + 1\} = \frac{(N - n + 1) \cdot (N - M + 1)}{(N - M - n + k + 1) \cdot (N + 1)} \times \\ \times \mathrm{P}\{K = k|n, M, N\}$$

$$\begin{aligned}
\mathrm{P}\{K = k|n, M, N\} &= \mathrm{P}\{K = n - k|n, N - M, N\} \\
&= \mathrm{P}\{K = M - k|N - n, M, N\} \\
&= \mathrm{P}\{K = N - M - n + k|N - n, N - M, N\}
\end{aligned}$$

$$\begin{aligned}
\mathrm{P}\{K \le k|n, M, N\} &= 1 - \mathrm{P}\{K \le n - k - 1|n, N - M, N\} \\
&= 1 - \mathrm{P}\{K \le M - k - 1|N - n, M, N\} \\
&= \mathrm{P}\{K \le N - M - n + k|N - n, N - M, N\}
\end{aligned}$$

The following property is a very helpful tool that shows that we are allowed to interchange M and n without affecting the hypergeometric probabilities. This property will frequently be used in this thesis.

Property 4.1.2. *If the roles of M and n are interchanged, this does not affect the hypergeometric probabilities; i.e.*

$$\mathrm{P}\{K = k|n, M, N\} = \mathrm{P}\{K = k|M, n, N\}\,.$$

The proof of this property is simple. A probabilistic explanation for Property 4.1.2 is given in Davidson and Johnson (1993).

Notice that $\mathrm{P}\{K = k|n, M, N\}$ is a unimodal function of k, see Johnson, Kotz and Kemp (1992). It takes on its maximum for the largest integer that does not exceed $\frac{(M+1)(n+1)}{N+2}$. If $\frac{(M+1)(n+1)}{N+2}$ is an integer, say c then it takes on its maximum for this integer c , but also for $c - 1$.

4.1.1 Properties of $\Lambda(n, M, N)$

We introduce the following notation,

$$\Lambda(n, M, N) = \mathrm{P}\{K \le k_0|n, M, N\} = \sum_{k=0}^{k_0} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}. \tag{4.1.2}$$

This notation suppresses the dependence on k_0, because in most of our applications we will not allow the value of k_0 to vary. Unless stated otherwise k_0 will be considered fixed in the sequel. In fact we will be more interested in the behaviour of Λ as a function of n, M, and N. We will discus some properties of $\Lambda(n, M, N)$ that are especially useful in Chapter 6.

Theorem 4.1.1. *The following properties hold for* $\Lambda(n, M, N)$*:*

1. $\Lambda(n, M, N) = \Lambda(M, n, N)$.

2. $\Lambda(n, M, N) = 1$ *if and only if* $M \le k_0$ *or* $n \le k_0$.

3. $\Lambda(n, M, N) = 0$ *if and only if* $M > N - n + k_0$.

4. *Let* $M \in \{0, \dots, N-1\}$*, then*

$$\Lambda(n, M+1, N) = \Lambda(n, M, N) - \frac{\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}}{\binom{N}{n}},$$

   or, equivalently,

$$\Lambda(n, M+1, N) = \Lambda(n, M, N) - \frac{n}{N} \cdot \mathrm{P}\{K = k_0 | n-1, M, N-1\}$$

   for $n \in \{1, \dots, N\}$.

5. *Let* $M \in \{0, \dots, N-1\}$*, then* $\Lambda(n, M+1, N) \le \Lambda(n, M, N)$*. The inequality is strict if and only if* $k_0 \le M \le N - n + k_0$ *and* $n > k_0$.

6. *Let* $n \in \{0, \dots, N-1\}$*, then* $\Lambda(n+1, M, N) \le \Lambda(n, M, N)$*. The inequality is strict if and only if* $k_0 \le n \le N - M + k_0$ *and* $M > k_0$.

Proof. Parts 1, 2, and 3 immediately follow from Property 4.1.2, (4.1.2), and the definition of the hypergeometric distribution, respectively. Using Pascal's triangle we obtain

$$\begin{aligned}
\binom{N}{n}\Lambda(n, M+1, N) &= \sum_{k=0}^{k_0} \binom{M+1}{k}\binom{N-M-1}{n-k} \\
&= \sum_{k=0}^{k_0} \left(\binom{M}{k} + \binom{M}{k-1}\right)\binom{N-M-1}{n-k} \\
&= \sum_{k=0}^{k_0} \binom{M}{k}\binom{N-M-1}{n-k} + \sum_{k=0}^{k_0-1} \binom{M}{k}\binom{N-M-1}{n-k-1}
\end{aligned}$$

and hence

$$\begin{aligned}\binom{N}{n}\varLambda(n, M, N) &= \sum_{k=0}^{k_0}\binom{M}{k}\binom{N-M}{n-k}\\ &= \sum_{k=0}^{k_0}\binom{M}{k}\left(\binom{N-M-1}{n-k}+\binom{N-M-1}{n-k-1}\right)\\ &= \sum_{k=0}^{k_0}\binom{M}{k}\binom{N-M-1}{n-k}+\sum_{k=0}^{k_0}\binom{M}{k}\binom{N-M-1}{n-k-1}\\ &= \binom{N}{n}\varLambda(n, M+1, N)+\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}.\end{aligned}$$

Summations with empty index sets are equal to zero by definition. This proves the first result of part 4. Its second result is obvious. Part 5 follows immediately from part 4. Part 6 follows from part 5 by applying the result from part 1. □

Theorem 4.1.1, part 5 shows that the probability of accepting the population is decreasing in M. Part 6 shows that this probability also decreases if a larger sample is taken. These facts are in accordance with intuition.

4.1.2 Properties of $\lambda(n, M, N)$

This subsection will focus on the quotient of $\varLambda(n, M+1, N)$ and $\varLambda(n, M, N)$, which plays a key role in proving some of the properties in Chapter 5. This quotient is defined by

$$\lambda(n, M, N) = \begin{cases} \frac{\varLambda(n,M+1,N)}{\varLambda(n,M,N)} > 0 & \text{if } M < N-n+k_0, \\ 0 & \text{if } N-n+k_0 \le M \le N-1\ . \end{cases} \tag{4.1.3}$$

According to Theorem 4.1.1, part 3 this ratio is well-defined for $M \le N-n+k_0$, and in the special case $M = N-n+k_0$ it is equal to zero. Obviously $0 \le \lambda \le 1$, according to Theorem 4.1.1, part 4. A number of properties of λ are collected in the following theorem.

Theorem 4.1.2. *The following properties hold for* $\lambda(n, M, N) \in [0, 1]$.

1. $\lambda(n, M, N) = 1$ *if and only if* $M < k_0$ *or* $n \le k_0$.
2. $\lambda(n, M, N) = 0$ *if and only if* $M \ge N-n+k_0$.

3. *Let $n > k_0$, if $k_0 \leq M \leq N - n + k_0$, then it can be written*
$$\lambda(n, M, N) = 1 - \frac{1}{g(n, M, N)},$$
where
$$g(n, M, N) = \frac{\binom{N}{n}\Lambda(n, M, N)}{\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}} > 0.$$

4. *Let $M \in \{0, \dots, N-1\}$, then $\lambda(n, M, N) \geq \lambda(n, M+1, N)$. The inequality is strict if and only if* $\max(0, k_0 - 1) \leq M \leq N - n + k_0 - 1$ *and* $n > k_0$.

5. *Let $n, M \in \{0, \dots, N-1\}$ then $\lambda(n, M, N) \geq \lambda(n+1, M, N)$. The inequality is strict if and only if $k_0 \leq n \leq N - M + k_0 - 1$ and $M > k_0$.*

6. *If $M \geq k_0$, $n > k_0$ and $N \geq n + M - k_0$, then*
$$\lambda(n, M, N) < \lambda(n, M, N+1).$$

Proof. Part 1 follows from (4.1.2) and Theorem 4.1.1, parts 2 and 5. Part 2 is obvious. Part 3 follows from Theorem 4.1.1, parts 3, 4, and 5. Now we prove part 4. For $n \leq k_0$, part 4 follows trivially from part 1. Therefore, we assume $n > k_0$. Using part 3 we derive for $k_0 \leq M \leq N - n + k_0$ that

$$g(n, M, N) = \frac{\binom{N}{n}\Lambda(n, M, N)}{\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}} = \sum_{k=0}^{k_0} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}} =$$
$$= \sum_{k=\max(0,n+M-N)}^{k_0} \frac{k_0!}{k!} \cdot (N-M) \cdot \prod_{h=k+1}^{k_0} \frac{N-M-n+h}{M-h+1} \cdot \prod_{j=k}^{k_0} \frac{1}{n-j}. \tag{4.1.4}$$

Notice that from Theorem 4.1.1, part 5 and the parts 1 and 2 just established it follows that $0 < \lambda(n, M, N) < 1$ for $k_0 \leq M \leq N - n + k_0 - 1$, and

$$1 = \lambda(n, 0, N) = \ldots = \lambda(n, k_0 - 1, N) > \lambda(n, k_0, N) \text{ for } k_0 \geq 1,$$

and

$$\lambda(n, N - n + k_0 - 1, N) > \lambda(n, N - n + k_0, N) = \ldots = \lambda(n, N - 1, N) = 0.$$

For $k_0 = 0$ there is no M such that $\lambda(n, M, N) = 1$. Now it remains to prove that $\lambda(n, M, N) > \lambda(n, M + 1, N)$ or, equivalently $g(n, M, N) > g(n, M + 1, N)$, for $k_0 \leq M \leq N - n + k_0 - 2$. This follows from (4.1.4), because $g(n, M, N)$ is a decreasing function of M on this interval. This concludes the proof of part 4. Notice that for $M < k_0$, part 5 follows trivially from part 1. Therefore, we assume $M \geq k_0$. From Theorem 4.1.1, part 6 and the parts 1 and 2 just proved, it follows that $0 < \lambda(n, M, N) < 1$ for $k_0 + 1 \leq n \leq N - M + k_0 - 1$, and

$$1 = \lambda(0, M, N) = \ldots = \lambda(k_0, M, N) > \lambda(k_0 + 1, M, N),$$

and

$$\lambda(N - M + k_0 - 1, M, N) > \lambda(N - M + k_0, M, N) = \ldots = \lambda(N - 1, M, N) = 0.$$

To complete the proof of part 5 we have to prove that $g(n, M, N) > g(n + 1, M, N)$ for $k_0 + 1 \leq n \leq N - M + k_0 - 2$. This follows from (4.1.4), because $g(n, M, N)$ is a decreasing function of n on this interval. This concludes the proof of part 5. From (4.1.4) we can see that $g(n, M, N) < g(n, M, N + 1)$ and hence $\lambda(n, M, N) < \lambda(n, M, N + 1)$ for $M \geq k_0$, $n > k_0$ and $N \geq n + M - k_0$. This concludes the proof of part 6. □

Remark 4.1.1. *Theorem 4.1.2, parts 4 and 5 imply logconcavity of the cumulative hypergeometric distribution function in the arguments n and M in all possible points, and even strict logconcavity on a relevant subset. Here, logconcavity of a function f on the non-negative integers is defined as $f(x + 2) \cdot f(x) \leq [f(x + 1)]^2$ with $x = 0, 1, \ldots$. Strictness occurs if the inequality is strict.*

4.2 Confidence sets

The value of M is not known to us, but after taking a random sample of size n we can give a point estimate and construct a confidence interval for M. Suppose we observe k items with the characteristic of interest in the sample. The maximum likelihood estimator for M is then given by the largest integer not exceeding $K \cdot \frac{N+1}{n}$, i.e. $\lfloor K \cdot \frac{N+1}{n} \rfloor$. If $K \cdot \frac{N+1}{n}$ is an integer, then $K \cdot \frac{N+1}{n} - 1$ and $K \cdot \frac{N+1}{n}$ both maximize the likelihood. This is not an unbiased estimator. The unbiased estimator is given by $K \cdot \frac{N}{n}$ and an unbiased estimator for its variance is

$$\frac{N \cdot (N-n)}{n-1} \cdot \frac{K}{n} \cdot \left(1 - \frac{K}{n}\right).$$

Only providing point estimators for M will not suffice. To quantify the uncertainty, we would also like to give confidence interval estimators. We prefer to give exact confidence intervals. Here, with exact we mean that we use the underlying hypergeometric distribution and not some approximation of this distribution. Due to the discrete character of the hypergeometric distribution it is possible to construct confidence sets instead of confidence intervals. Although from a practical view we prefer confidence intervals, we cannot exclude the possibility of confidence sets that are not confidence intervals.

If we observe $K = k$, where $k \in \{0, \dots, n\}$, we would like to find a way to associate to this value of k for $\alpha \in (0, 1)$, a subset of possible values of $M \in \{0, \dots, N\}$, we call this subset $M_C(k)$, to state that $M_C(K)$ contains the true value of M with probability of at least $1 - \alpha$, or, in symbols,

$$\mathrm{P}\{M_C(K) \ni M | M\} \geq 1 - \alpha, \text{ for every } M \in \{0, \dots, N\}. \tag{4.2.1}$$

The quantity $1 - \alpha$ is called the confidence level. The probability in equation (4.2.1) is called the coverage probability for M. Due to the discrete character of M it is not possible to exactly attain the nominal confidence level $1 - \alpha$ without using randomized methods (Wright, 1997). These methods will always attain the exact nominal confidence level. We will not consider these methods. The methods discussed here are conservative, meaning that the confidence level will be at least $1 - \alpha$.

We have to construct $M_C(K)$, i.e. $M_C(0), \dots, M_C(n)$ in such a way that (4.2.1) is satisfied for every $M \in \{0, \dots, N\}$. We first notice that given the true value of M the probability that $M_C(K)$ contains M is the same as the total probability of observing those values of k for which $M_C(k)$ contains M. Let $R(M)$ be the set containing all these values of k, i.e.

$$R(M) = \{k | M_C(k) \ni M\},$$

then we can rewrite the left-hand side of (4.2.1) in the following way

$$\mathrm{P}\{M_C(K) \ni M | M\} = \sum_{k \in R(M)} \mathrm{P}\{K = k | n, M, N\}. \tag{4.2.2}$$

Remember that $K \sim \mathcal{H}(n, M, N)$.

Now, suppose we construct for every $M \in [0, \ldots, N]$ sets $R'(M)$ with values of k such that $\sum_{k \in R'(M)} \mathrm{P}\{K = k | n, M, N\} \geq 1 - \alpha$ and let $M'_C(k)$ be the set of all values of M for which $k \in R'(M)$. It is obvious from (4.2.2) that by using $M_C(K) = M'_C(K)$, and also $R(M) = R'(M)$, we have found a way to define $M_C(K)$ such that equation (4.2.1) is satisfied for every $M \in \{0, \ldots, N\}$. There are various methods to define $R'(M)$ to construct confidence sets. We will discuss two of these methods here.

4.2.1 Test-method

We call $M_C(k)$ a confidence set. In those cases where the confidence sets $M_C(K)$ actually turn out to be confidence intervals $[M_L(K), M_U(K)]$, we speak of a $100(1-\alpha)\%$ two-sided confidence interval with lower confidence bound $M_L(K)$ and upper confidence bound $M_U(K)$.

Since we know that the hypergeometric distribution function is a unimodal function of k, we can construct $R(M)$ in the following way. For every $M \in [0, \ldots, N]$ the set $R(M)$ contains all values of k for which

$$\mathrm{P}\{K \leq k | M\} > \beta \text{ and } \mathrm{P}\{K \geq k | M\} > \gamma,$$

with $\beta + \gamma = \alpha$. First, we will consider the case $\beta = \gamma = \frac{\alpha}{2}$. Note that $\min(R(M))$ and $\max(R(M))$ are non-decreasing functions of M. This ensures that the confidence set $M_C(K) = \{M | K \in R(M)\}$ will always be a confidence interval. If we observe $K = k$, then the lower and upper confidence interval limits are given by

$$M_L(k) = \text{smallest integer } M \text{ s.t. } \mathrm{P}\{K \geq k\} > \tfrac{\alpha}{2}$$

and

$$M_U(k) = \text{largest integer } M \text{ s.t. } \mathrm{P}\{K \leq k\} > \tfrac{\alpha}{2}.$$

This method coincides with generating a confidence interval by inverting a family of hypothesis tests for M. That is why this method is called the test-method. It also appears to be the same method as described by Katz (1953), Konijn (1973) and Wright (1991).

Buonaccorsi (1987) showed that this method is always superior to the one described by Cochran (1977) in the sense that this method always delivers confidence intervals that are shorter than the confidence intervals that were suggested by Cochran. Cochran's intervals were the finite population analog of the method by Clopper and Pearson (1934) for the construction of confidence intervals for a binomial fraction.

Also other values of β and γ could be considered. A very interesting case is the case of $\beta = 0$ and $\gamma = \alpha$. This is the case of only giving an upper confidence bound. Bickel and Doksum (1977) showed that this bound will be uniformly most accurate, because if the inverse test method is used, then the corresponding tests are uniformly most powerful.

4.2.2 Likelihood-method

We could also construct $R(M)$ in the following way. For every $M \in [0, \ldots, N]$ we sort the values of k according to the size of the accompanying probabilities. Therefore, $k_{(1)}$ has the largest probability, $k_{(2)}$ has the next largest and so forth. If ties occur between $k_{(i)}$ and $k_{(i+1)}$, then the ordering is not strict. We deal with this issue later. This means that

$$\mathrm{P}\big\{K = k_{(1)}|M\big\} \geq \mathrm{P}\big\{K = k_{(2)}|M\big\} \geq \ldots \geq \mathrm{P}\big\{K = k_{(n)}|M\big\}\,.$$

Now, for every $M \in [0, \ldots, N]$ we construct $R(M)$ in such a way that it consists of the smallest possible number of elements, say $n_k(M)$, such that

$$\sum_{i=1}^{n_k(M)} \mathrm{P}\big\{K = k_{(i)}|M\big\} \geq 1 - \alpha.$$

Because the elements are selected based on their likelihood, we call the confidence set $M_C(K) = \{M|K \in R(M)\}$ obtained in this way a likelihood confidence set. This method was first described by Wendell and Schmee (2001). We will call $\min\big(M_C(K)\big)$ the lower confidence bound $M_L(K)$ and $\max\big(M_C(K)\big)$ the upper confidence bound $M_U(K)$. Using this method it is possible that the confidence sets produced are not confidence intervals, gaps can occur. A practical solution is to take the interval $[M_L(K), M_U(K)]$. Some theoretical solutions are suggested by Wendell and Schmee. They also show that the occurrence of these gaps is seldom.

Using this method ties can occur. Ties occur when

$$\mathrm{P}\big\{K = k_{(n_k(M))}|M\big\} = \mathrm{P}\big\{K = k_{(n_k(M)+1)}|M\big\}\,.$$

These ties often occur when the hypergeometric distribution is symmetric for lower and upper tail probabilities. Suppose $k_{(n_k(M))} < k_{(n_k(M)+1)}$, then if we choose $k_{(n_k(M))}$ to add to $R(M)$ this means that $M_U(k_{(n_k(M))})$ is less tight and that $M_L(k_{(n_k(M)+1)})$ is tighter compared to the choice of $k_{(n_k(M)+1)}$. Of course this choice has to be made before we start sampling.

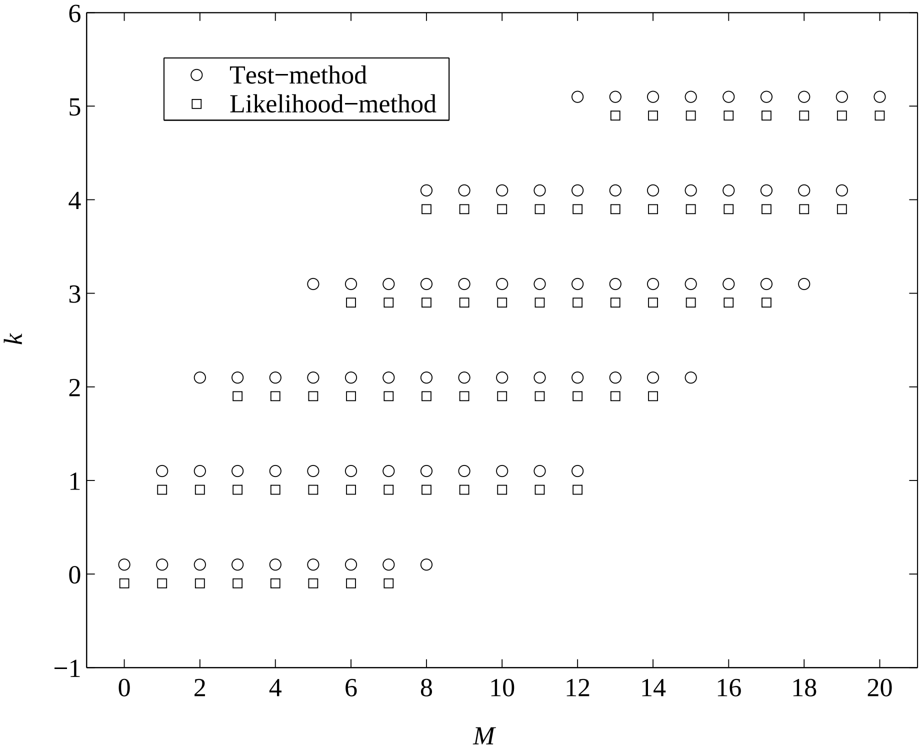

Figure 4.1. Comparison of the 90%-confidence intervals of the test-method and the likelihood-method for $n = 5$ and $N = 20$.

Wendell and Schmee also showed by simulation studies that this method performs well in comparison to test-method. Figure 4.1 gives a comparison of the two methods for a 90%-confidence interval with $n = 5$ and $N = 20$. Notice that in this case for $k = 1$ and $k = 4$ the confidence intervals are equally long. In all other cases the likelihood-method produces shorter intervals. It is also possible that the test-method produces shorter intervals, but study of Wendell and Schmee shows that this will not occur very often.

4.2.3 Approximate confidence sets

Instead of using the exact hypergeometric distribution to obtain confidence sets for M, also in certain cases approximations of this distribution can be used. We use these approximations to find confidence intervals for $p = \frac{M}{N}$. Of course confidence intervals for M can be obtained by multiplying with the population size N. We will describe three approximations, that is the approximation by the binomial distribution, the approximation by the Poisson distribution, and the approximation by the normal distribution.

The question arises when we are allowed to use a certain approximation. Text books give so-called rules of thumb. However, these rules differ among text books, and are almost always given without any quantitative assessment of the quality of such approximations. Therefore, we should not pay too much attention to rules of thumb. Schader and Schmid (1992) showed that two rules of thumb for approximating the binomial distribution by the normal distribution are of dubious quality in numerical accuracy. Leemis and Kishor (1996) investigated rules of thumb for normal and Poisson approximations of the binomial distribution. From their article we can see, especially when we look at it from an auditing point of view (in which the proportions are usually very small), that using rules of thumb without any quantitative assessment of the quality of the approximations should be avoided. Therefore, if possible we should use an exact approach.

We will apply these approximations to the test-method with $\beta = \gamma = \frac{\alpha}{2}$. Therefore, in terms of p our problem focusses on solving the following equations to find the smallest integer value of $N \cdot p_L$ such that

$$\mathrm{P}\{K \geq k | p = p_L\} = \sum_{i=k}^{n} \frac{\binom{Np_L}{i}\binom{N(1-p_L)}{n-i}}{\binom{N}{n}} > \frac{\alpha}{2},$$

and the largest integer value of $N \cdot p_U$ such that

$$\mathrm{P}\{K \leq k | p = p_U\} = \sum_{i=0}^{k} \frac{\binom{Np_U}{i}\binom{N(1-p_U)}{n-i}}{\binom{N}{n}} > \frac{\alpha}{2}.$$

Note that, p_L and p_U are elements of $\{0, 1/N, 2/N, \ldots, 1\}$. Our $(1-\alpha)$-confidence interval for p becomes $[p_L, p_U]$. In certain cases we can approximate the hypergeometric distribution by another discrete or even continuous distribution

Binomial approximation

For relatively small values of p and large values of N we can approximate the hypergeometric distribution by the binomial distribution. As a rule of thumb $p < 0.1$ and $N \geq 60$ is sometimes used. Now, p_L and p_U are elements of [0, 1], and we have to solve the following problem. Find p_L and p_U such that

$$\mathrm{P}\{K \geq k | p = p_L\} = \sum_{i=k}^{n} \binom{n}{i} p_L^i (1-p_L)^{n-i} = \frac{\alpha}{2},$$

and

$$\mathrm{P}\{K \leq k | p = p_U\} = \sum_{i=0}^{k} \binom{n}{i} p_U^i (1-p_U)^{n-i} = \frac{\alpha}{2}.$$

This confidence interval is known as the Clopper-Pearson confidence interval for p (Clopper and Pearson, 1934). The following relationship relates the tail of a binomial distribution with the tail of an F-distribution

$$\sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} = \mathrm{P}\left\{Y \leq \frac{(1-p)(k+1)}{p(n-k)}\right\}$$

with $Y \sim F(2(n-k), 2(k+1))$. A proof can be found in Leemis and Kishor (1996). Now, it follows immediately that

$$p_L = \frac{1}{1 + \frac{n-k+1}{k} F_{1-\frac{\alpha}{2}}(2(n-k+1), 2k)}$$

and

$$p_U = \frac{1}{1 + \frac{n-k}{k+1} F_{\frac{\alpha}{2}}(2(n-k), 2(k+1))},$$

where $F_{1-\frac{\alpha}{2}}(\cdot,\cdot)$ and $F_{\frac{\alpha}{2}}(\cdot,\cdot)$ denote the $100{\cdot}(1-\alpha/2)$th and the $100{\cdot}(\alpha/2)$th percentile of the F-distribution. Many statistical software packages provide the percentiles of the F-distribution. For large degrees of freedom numerical problems can occur, then approximate methods could be used. Vollset (1993) compared thirteen methods that produce two-sided confidence intervals for the binomial proportion. Newcombe (1998) further examined seven of these methods. The Clopper-Pearson method is known to be rather conservative, meaning that the coverage probabilities usually exceed $1-\alpha$. Very often approximate methods as adjusted Wald intervals or continuity corrected score intervals are suggested to tackle this problem (e.g. Vollset, 1993; Leemis and Kishor, 1996). Blyth and Still (1983) remark that the Clopper-Pearson method is only an approximation of the exact interval and consider procedures with correct confidence coefficient. These methods give numerical results that are very similar to the approach with the acceptability function of Blaker and Spjøtvoll (2000).

Poisson approximation

For small values of p and extremely large values of n the Poisson approximation can be used. As a rule of thumb ($p < 0.01$) and ($n \geq 1000$) is sometimes used. Now, p_L and p_U are elements of [0, 1] again, and we have to solve the following problem. Find p_L and p_U such that

$$\mathrm{P}\{K \geq k | p = p_L\} = \sum_{i=k}^{\infty} \frac{e^{-np_L}(np_L)^i}{i!} = \frac{\alpha}{2},$$

and

$$\mathrm{P}\{K \leq k | p = p_U\} = \sum_{i=0}^{k} \frac{e^{-np_U}(np_U)^i}{i!} = \frac{\alpha}{2}.$$

The following relationship relates the tail of a Poisson distribution with the tail of a χ^2-distribution.

$$\sum_{i=0}^{k-1} \frac{e^{-np}(np)^i}{i!} = \mathrm{P}\{Y > 2np\}$$

with $Y \sim \chi^2(2k)$. A proof can be found in Johnson et al. (1992). Now, it follows immediately that

$$p_L = \frac{1}{2n}\chi^2_{\frac{\alpha}{2}}(2k)$$

and

$$p_U = \frac{1}{2n}\chi^2_{1-\frac{\alpha}{2}}(2(k+1)),$$

where $\chi^2_{\frac{\alpha}{2}}(\cdot)$ and $\chi^2_{1-\frac{\alpha}{2}}(\cdot)$ denote the $100 \cdot (\alpha/2)$th and the $100 \cdot (1-\alpha/2)$th percentile of the χ^2-distribution. Also this confidence interval is conservative. It is possible to increase some of the lower endpoints and decrease some of the higher endpoints and still satisfy the coverage requirement. Examples can be found in Crow and Gardner (1959), Casella and Robert (1989), and Kabaila and Byrne (2001).

Normal approximation

We can also use the normal distribution to approximate the hypergeometric distribution. To do so the rule of thumb $np \geq 4$ is sometimes used. We can approximate the hypergeometric distribution by a normal distribution with mean and variance equal to mean and variance of K. Therefore, p_L and p_U are elements of [0, 1] again, and using continuity corrections we have to solve the following problem. Find p_L and p_U such that

$$\mathrm{P}\{K \geq k | p = p_L\} = 1 - \Phi\left(\frac{k - 0.5 - np_L}{\sqrt{np_L(1-p_L)\frac{N-n}{N-1}}}\right) = \frac{\alpha}{2},$$

and

$$\mathrm{P}\{K \leq k | p = p_U\} = \Phi\left(\frac{k + 0.5 - np_U}{\sqrt{np_U(1-p_U)\frac{N-n}{N-1}}}\right) = \frac{\alpha}{2}.$$

Solving these equations gives the following confidence interval

$$[p_L, p_U] = \frac{1}{2u}\Bigg[u + (2k - n \pm 1) \\ \pm \sqrt{u^2 - \frac{2u}{n}\big((k \pm 0.5)^2 + (n - k \mp 0.5)^2\big) + (2k - n \pm 1)^2}\,\Bigg]$$

where

$$u = n + \frac{N-n}{N-1}Z^2_{1-\frac{\alpha}{2}},$$

with $Z^2_{1-\frac{\alpha}{2}}$ the $100 \cdot (1-\alpha/2)$th percentile of the standard normal distribution. More simplified versions of this approximation are also used.

Ling and Pratt (1984) compared several normal approximations for the hypergeometric distribution. They show that especially the so-called Peizer approximations turn out to be very accurate. These complicated approximations originate from an unpublished paper by Peizer. However, these approximations are not invertible in closed form. Molenaar (1973) gave two relatively simpler normal approximations that are invertible in closed form, but still give very complicated solutions. These approximations will probably give more accurate bounds than the method described above. A crude approximation can be obtained by using the approximate normality of p with mean equal to the unbiased estimator for p, i.e. $\frac{K}{n}$, and variance equal to the unbiased estimator for the variance of this estimator, i.e.

$$\frac{N-n}{N(n-1)}\left(\frac{K}{n}\right)\left(1-\frac{K}{n}\right).$$

If we also correct for continuity, then we find the following confidence interval

$$[p_L, p_U] = \left[\left(\frac{k}{n}\right) \pm Z^2_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N(n-1)}\left(\frac{k}{n}\right)\left(1-\frac{k}{n}\right)+\frac{1}{2n}}\,\right].$$

4.3 Computing the hypergeometric distribution

Theorem 4.1.1, part 4 can be used to find some recursive properties that we will use in calculating the hypergeometric distribution. It shows that we can compute $\Lambda(n, M, N)$ from $\Lambda(n, M+1, N)$, by using the hypergeometric probability $\mathrm{P}\{K = k_0 | n-1, M, N-1\}$. But suppose that we already calculated $\Lambda(n, M+1, N)$ from $\Lambda(n, M+2, N)$, then we can use this step to facilitate the computation of $\mathrm{P}\{K = k_0 | n-1, M, N-1\}$. Property 4.3.1 gives a few examples of this.

Property 4.3.1. *The following recursive properties facilitate the computation of the hypergeometric distribution.*

1. *If* $k_0 \le M \le N-n+k_0-1$ *and* $k_0+1 \le n \le N-1$, *then*

$$\Lambda(n, M, N) = \Lambda(n, M+1, N) + C_1(n, M, N)$$

with

$$C_1(n, M, N) = \frac{M - k_0 + 1}{M + 1} \cdot \frac{N - M - 1}{N - M - n + k_0} \cdot C_1(n, M + 1, N),$$

If $k_0 + 1 \leq n \leq N - 1$*, then*

$$C_1(n, N - n + k_0, N) = \frac{n!\,(N - n + k_0)!}{k_0!\,N!}.$$

2. *If* $k_0 + 1 \leq M \leq N - n + k_0 + 1$ *and* $k_0 + 2 \leq n \leq N$*, then*

$$\Lambda(n, M, N) = \Lambda(n - 1, M, N) - C_2(n, M, N)$$

with

$$C_2(n, M, N) = \frac{N - M}{N - n + 1} \cdot \frac{M - k_0}{n - k_0 - 1} \cdot C_1(n - 1, M, N).$$

3. *If* $k_0 \leq M \leq N - n + k_0$ *and* $k_0 + 1 \leq n \leq N$*, then*

$$\Lambda(n, M, N) = \Lambda(n, M + 1, N) + C_1(n, M, N)$$

with

$$C_1(n, M, N) = \frac{n}{M + 1} \cdot C_2(n, M + 1, N).$$

4. *If* $k_0 + 1 \leq M \leq N - n + k_0$ *and* $k_0 \leq n \leq N - 1$*, then*

$$\Lambda(n, M, N) = \Lambda(n + 1, M - 1, N) + C_3(n, M, N)$$

with

$$C_3(n, M, N) = \frac{M - n - 1}{n + 1} \cdot C_1(n + 1, M - 1, N).$$

Proof. First we prove part 1. Using Theorem 4.1.1, part 4 we find

$$\Lambda(n, M, N) = \Lambda(n, M + 1, N) + C_1(n, M, N)$$

with

$$C_1(n, M, N) = \frac{\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}}{\binom{N}{n}},$$

and

$$\Lambda(n, M + 1, N) = \Lambda(n, M + 2, N) + C_1(n, M + 1, N)$$

with

$$C_1(n, M+1, N) = \frac{\binom{M+1}{k_0}\binom{N-M-2}{n-k_0-1}}{\binom{N}{n}}.$$

From Theorem 4.1.1, part 4 we notice that $C_1(n, M, N) > 0$ and $C_1(n, M + 1, N) > 0$ if $k_0 \leq M \leq N - n + k_0 - 1$ and $k_0 + 1 \leq n \leq N - 1$. Combining the expressions for $C_1(n, M, N)$ and $C_1(n, M + 1, N)$ gives

$$C_1(n, M, N) = \frac{M - k_0 + 1}{M + 1} \cdot \frac{N - M - 1}{N - M - n + k_0} \cdot C_1(n, M + 1, N).$$

For $M = N - n + k_0$ and $k_0 + 1 \leq n \leq N - 1$ we find

$$\begin{aligned} C_1(n, N - n + k_0, N) &= \Lambda(n, N - n + k_0, N) - \Lambda(n, N - n + k_0 + 1, N) \\ &= \Lambda(n, N - n + k_0, N) \\ &= \frac{n!\,(N - n + k_0)!}{k_0!\,N!}. \end{aligned}$$

In proving part 2 we again use Theorem 4.1.1, part 4 in combination with part 1. Using this theorem we find

$$\Lambda(n, M, N) = \Lambda(n - 1, M, N) - C_2(n, M, N)$$

with

$$C_2(n, M, N) = \frac{\binom{n-1}{k_0}\binom{N-n}{M-k_0-1}}{\binom{N}{M}},$$

and

$$\Lambda(n - 1, M, N) = \Lambda(n - 1, M + 1, N) + C_1(n - 1, M, N)$$

with

$$C_1(n-1, M, N) = \frac{\binom{M}{k_0}\binom{N-M-1}{n-k_0-2}}{\binom{N}{n-1}}.$$

Observe that $C_2(n, M, N) > 0$ and $C_1(n - 1, M, N) > 0$ if $k_0 + 1 \leq M \leq N - n + k_0 + 1$ and $k_0 + 2 \leq n \leq N$. Combining the expressions for $C_2(n, M, N)$ and $C_1(n - 1, M, N)$ gives

$$C_2(n, M, N) = \frac{N - M}{N - n + 1} \cdot \frac{M - k_0}{n - k_0 - 1} \cdot C_1(n - 1, M, N).$$

To prove part 3 we use the previous results

$$\Lambda(n, M, N) = \Lambda(n, M + 1, N) + C_1(n, M, N)$$

with

$$C_1(n, M, N) = \frac{\binom{M}{k_0}\binom{N-M-1}{n-k_0-1}}{\binom{N}{n}},$$

and

$$\Lambda(n, M+1, N) = \Lambda(n-1, M+1, N) - C_2(n, M+1, N)$$

with

$$C_2(n, M+1, N) = \frac{\binom{n-1}{k_0}\binom{N-n}{M-k_0}}{\binom{N}{M+1}}.$$

Note that $C_1(n, M, N) > 0$ and $C_2(n, M+1, N) > 0$ if $k_0 \leq M \leq N - n + k_0$ and $k_0+1 \leq n \leq N$. Combining the expressions of $C_1(n, M, N)$ and $C_2(n, M+1, N)$ gives

$$C_1(n, M, N) = \frac{n}{M+1} \cdot C_2(n, M+1, N).$$

In proving part 4 we use Theorem 4.1.1, part 4 in combination with part 1 and find

$$\Lambda(n, M, N) = \Lambda(n+1, M, N) + C_2(n+1, M, N)$$

with

$$C_2(n+1, M, N) = \frac{\binom{n}{k_0}\binom{N-n-1}{M-k_0-1}}{\binom{N}{M}}.$$

We again use Theorem 4.1.1, part 4 to find

$$\Lambda(n+1, M, N) = \Lambda(n+1, M-1, N) - C_1(n+1, M-1, N)$$

with

$$C_1(n+1, M-1, N) = \frac{\binom{M-1}{k_0}\binom{N-M}{n-k_0}}{\binom{N}{n+1}}.$$

Note that $C_1(n+1, M-1, N) > 0$ and $C_2(n+1, M, N) > 0$ if $k_0+1 \leq M \leq N - n + k_0$ and $k_0 \leq n \leq N-1$. Combining the expressions of $C_1(n+1, M-1, N)$ and $C_2(n+1, M, N)$ gives

$$C_2(n+1, M, N) = \frac{M}{n+1} \cdot C_1(n+1, M-1, N).$$

Using the previous results we find

$$\begin{aligned}
\Lambda(n, M, N) &= \Lambda(n+1, M, N) + C_2(n+1, M, N) \\
&= \Lambda(n+1, M, N) + \frac{M}{n+1} \cdot C_1(n+1, M-1, N) \\
&= \Lambda(n+1, M-1, N) - C_1(n+1, M-1, N) + \\
&\qquad\qquad + \frac{M}{n+1} \cdot C_1(n+1, M-1, N) \\
&= \Lambda(n+1, M-1, N) + \frac{M-n-1}{n+1} \cdot C_1(n+1, M-1, N)
\end{aligned}$$

□

Table 4.1. The values of $\Lambda(n, M, 8)$ for $k_0 = 2$.

$M \setminus n$	0	1	2	3	4	5	6	7	8
0	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1
3	1	1	1	0.9821	0.9286	0.8214	0.6429	0.375	0
4	1	1	1	0.9286	0.7571	0.5	0.2143	0	0
5	1	1	1	0.8214	0.5	0.1786	0	0	0
6	1	1	1	0.6429	0.2143	0	0	0	0
7	1	1	1	0.375	0	0	0	0	0
8	1	1	1	0	0	0	0	0	0

Table 4.1 gives the values of $\Lambda(n, M, 8)$ for all possible combinations of n and M with $k_0 = 2$. This table gives an illustration of Property 4.3.1. First we can use Theorem 4.1.1, parts 2 and 3. This gives us $\Lambda(n, M, 8) = 1$ if $n \leq 2$ or $M \leq 2$ and $\Lambda(n, M, 8) = 0$ for all combinations of n and M for which $M > 10-n$. We start computing this table with $\Lambda(3, 7, 8)$. Using Property 4.3.1, part 1 it immediately follows that

$$\Lambda(3, 7, 8) = C_1(3, 7, 8) = \frac{3!\,(8-3+2)!}{2!\,8!} = 3/8 = 0.375,$$

note that $\Lambda(3, 8, 8) = 0$. Again by using Property 4.3.1, part 1 we can calculate

$\Lambda(3, 6, 8)$:

$$\begin{aligned}\Lambda(3,6,8) &= \Lambda(3,7,8) + C_1(3,6,8)\\ &= 3/8 + \frac{6-2+1}{6+1} \times \frac{8-6-1}{8-6-3+2} \times 3/8\\ &= 3/8 + 5/7 \times 3/8 = 3/8 + 15/56 = 9/14 \approx 0,6429.\end{aligned}$$

We can repeat this procedure until we have found $\Lambda(3, 3, 8)$ and by then we have found $\Lambda(3, M, 8)$ for all possible values of M. We can calculate $\Lambda(4, 6, 8)$, $\Lambda(4, M, 8) = 0$ for $M > 6$, by using Property 4.3.1, part 2:

$$\begin{aligned}\Lambda(4,6,8) &= \Lambda(3,6,8) - \frac{8-6}{8-4+1} \times \frac{6-2}{4-2-1} \times C_1(3,6,8)\\ &= 9/14 - 2/5 \times 4 \times 15/56 = 9/14 - 3/7 = 3/14 \approx 0.2143.\end{aligned}$$

Since $\Lambda(4, 7, 8) = 0$, it follows that $C_1(3, 6, 8) = 3/14$. Now we can apply Property 4.3.1, part 1 to find the remaining values of $\Lambda(4, M, 8)$. By repeating the procedure above the table can be completed.

Sometimes we have to use the terms of Λ to find a recursive expression. For instance if we would like calculate $\Lambda(n, M, N)$ from $\Lambda(n, M, N-1)$ or from $\Lambda(n, M-1, N-1)$. We introduce the following notation. We write $P(n, M, N)$ as a $(k_0 + 1)$-vector, with elements

$$P_j(n, M, N) = \mathrm{P}\{K = j\} = \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}, \quad j = 0, \ldots, k_0$$

and $\iota' = (1, \ldots, 1)$ a $(k_0 + 1)$-vector. Now, it follows that

$$\Lambda(n, M, N) = \iota' \cdot P(n, M, N). \tag{4.3.1}$$

How we compute the probabilities $P_j(n, M, N)$ from $P_j(n, M, N-1)$ will be shown in the following property.

Property 4.3.2. *If $M \geq k_0$, $n \geq k_0$ and $N > n$, then for $j = 0, \ldots, k_0$*

$$P_j(n, M, N) = \begin{cases} 0 & \text{if } j < n + M - N\\ \frac{j+1}{N} \cdot P_{j+1}(n, M, N-1) & \text{if } j = n + M - N < k_0\\ \binom{M}{j}/\binom{N}{n} & \text{if } j = n + M - N = k_0\\ \frac{(N-n)\cdot(N-M)}{N\cdot(N-M-n+j)} \cdot P_j(n, M, N-1) & \text{if } j > n + M - N.\end{cases}$$

Proof. The cases $j < n + M - N$, $j > n + M - N$ and $j = k_0 = n + M - N$ follow immediately from the definition of the hypergeometric probability. Note that if $M \geq k_0$, $n \geq k_0$ and $N > n$, then $P_j(n, M, N - 1) > 0$ implies that $P_j(n, M, N) > 0$. For $j = n + M - N < k_0$ the probability $P_j(n, M, N - 1)$ equals zero, but the probability $P_{j+1}(n, M, N - 1)$ does have a positive value. It is not difficult to see that for

$$P_j(n, M, N) = \frac{\binom{M}{j}}{\binom{N}{n}} = \frac{j+1}{N} \cdot \frac{\binom{M}{j+1}}{\binom{N-1}{n}} = \frac{j+1}{N} \cdot P_{j+1}(n, M, N - 1).$$

□

Notice that once $n + M - N \leq 0$, all elements of $P(n, M, N)$ are positive. We can find a similar property if we would like to compute the probability $P_j(n, M, N)$ from the probability $P_j(n, M - 1, N - 1)$.

Property 4.3.3. *If $M > k_0$, $n \geq k_0$ and $N > n$, then for $j = 0, \ldots, k_0$*

$$P_j(n, M, N) = \begin{cases} 0 & \text{if } j < n + M - N \\ \frac{M \cdot (N-n)}{N \cdot (M-j)} \cdot P_j(n, M - 1, N - 1) & \text{if } j \geq n + M - N. \end{cases}$$

Proof. This follows immediately from the definition of the hypergeometric probability. If $M > k_0$, $n \geq k_0$ and $N > n$, then $P_j(n, M - 1, N - 1) > 0$ implies that $P_j(n, M, M) > 0$. □

The properties we derived here will be essential in the developing of the algorithms that we will describe in Chapter 5 and 6. These properties enable the algorithms to be efficient and accurate.