Chapter 1

Introduction

1.1 Auditing

In ancient Rome farmers went to a sort of committee, called Forum, to publicly render account of their earnings. On the basis of this oral report the Forum decided how much taxes had to be paid by the farmer. The auditor listened to the report of the farmer and based on his knowledge of the circumstances, he then judged whether the report agreed with the established rules and reality. If the farmer’s report did not agree with the established rules and reality, it was his task to say so. For the history of the profession of ‘auditor’ we refer to e.g. Schilder, Gortemaker, Van Manen and Waardenburg (1998, p. 2).

In Latin ‘auditor’ literally means ‘listener’. Nowadays, not that much has changed in the essential elements of auditing if we follow the definition from Arens and Loebbecke (1997, p. 2):

“Auditing is the accumulation and evaluation of evidence about information to determine and report on the degree of correspondence between the information and established criteria. Auditing should be done by a competent independent person.”

There are different types of tests that can be performed by an auditor, but the two types we deal with in this thesis are substantive tests of details and compliance tests.
Substantive test of details

Collecting data to find the aggregate monetary error amount is called a substantive test of details. This error amount in the financial statements is considered to be material if knowledge of this error would affect a decision of a reasonable user of the statements. It is obvious that unambiguous quantitative criteria to find the materiality are not always available and subject to discussion. Materiality is a very important subject in auditing because it is used to decide on which methods are to be used to audit the financial statements and it is also used to evaluate potential and actual errors in the financial statements. If these errors exceed the materiality, then the auditor will ask the client to modify the financial statements.

The auditor has to assess the risk of material errors being made and to develop procedures to reduce this risk. A model that is widely used among auditors when dealing with risk is the audit risk model. This model is a more qualitative than quantitative model that is mainly used for planning purposes. This model helps to decide how much evidence should be collected at different stages of the audit process. For a more elaborate discussion of this model we refer to Arens and Loebbecke (1997). This audit risk model and the quantification of the risks involved are based on the auditor’s experience and insights. The process of auditing also involves making a lot of decisions under uncertainty. It is obvious that statistics could play an important role in dealing with these decisions under uncertainty. Oddly enough statistical procedures are hardly used and most of them use a classical approach to statistics. In his dissertation Wille (2003) stresses that using Bayesian decision analysis can be very helpful in dealing with these decisions under uncertainty in auditing.

Compliance test

A compliance test aims to check whether the rules followed by the auditee (for instance an organisation, a department or a person under inspection) comply with the rules set by a certain higher authority (for instance government or management). Recent developments in internal control emphasize the importance of trustworthy and reliable internal control procedures. Compliance tests are among these procedures.

Fraudulent reporting by a few large American and European companies which
were quoted on the stock exchange damaged the trust people had in capital markets, company directors, supervisors and other links of the reporting chain. In an attempt to regain this trust several countries developed legislation with respect to the way stock exchange quoted companies should report about internal control. In the USA for example, the Sarbannes-Oxley Act (SOx), developed by senators Sarbannes and Oxley, was imposed on stock exchange quoted companies. At an earlier stage, the COSO (Committee of Sponsoring Organizations of the Treadway Commission) report was developed for the same purpose, i.e. to prevent fraudulent reporting. This report gives a common definition of internal control and presents a framework to judge and improve internal control procedures. Apparently the recommendations of this report were not obligatory enough and more stringent measures were necessary. The new legislation mainly covers the part of internal control that concerns the financial reporting (ICFR). On the basis of the COSO report, De Koning (2004) defines internal control as a process that ensures that the company objectives are reached. The internal control of financial reporting is a component of this and tries to guarantee the reliability of the financial information. This is essential to shareholders and other users of the financial statement to assess if indeed the company objectives are reached and to management to render account about this. According to the legislation of SOx, companies should declare by means of quarterly and yearly reports that the acquirement of financial information is reliable and that it creates a fair and true view with respect to ICFR. These tasks do not belong exclusively to the external auditor anymore, but should be performed by the company itself as well. According to Emanuels, Van Leeuwen and Wallage (2004), this means that companies do not only have to define their internal control procedures, but they also have to construct a system which describes the assessment of the effectiveness of the internal control structure of procedures in order to enable monitoring.

Besides the attestation about the financial statement, the auditor has to report about the ICFR. The PCAOB (Public Company Accounting Oversight Board) formulates the standards for auditors. They not only demand an examination of the evaluation performed by management, but also an independent opinion by the auditor about the effectiveness of ICFR is required.
1.2 Statistical Auditing

In society, industry, etc. we like to see that certain expectations that we reasonably may have are fulfilled. We expect consumer products to have a certified quality. Similarly we expect that the financial reports of companies show reliable pictures of profits and financial positions. In quality control there are procedures for certifying the quality of products (e.g. Ryan, 1989). One method is to divide the production in batches and to take samples from these batches. If a sample does not give enough evidence that the batch is of the desired quality, it has to be reworked, defective items are repaired or replaced, or the whole batch will be destroyed. The cause of the rejection of this batch, the special cause of variation, can only be determined afterwards and has already led to severe consequences. This method can be refined by combining it with a method where the production process is monitored over time. Control charts on which we plot a function of the observations against time, are used to decide whether the process is still in control, meaning that there is no reason to believe that there is a special cause of variation by which items of undesired quality are being produced. If the plotted points are inside the bandwidth of certain predefined limits, then the process is considered to be in control. If one or more points are outside the bandwidth of these limits there is reason to suspect that there is a special cause of variation and the source of this variation has to be determined. In this way the variation in process outcomes can be reduced significantly, and hence less batches will be rejected. In most cases the batches will automatically be judged to be of good quality, because the process will lead to good products if it is in control.

In auditing, the financial statements have to be of a certain specified quality, to protect the users of the financial statements against losses that follow from using these statements. Analogously to the batch in quality control the financial statements could be inspected afterwards with the aid of statistical methods and either be judged to be of the desired quality or not (financial statements audit). The use of statistical quality control procedures within auditing and accounting was already promoted by Deming (1991), and others before him, in his landmark book ‘Out of the Crisis’ in which he stresses that managers should pay more attention to increasing the quality of the systems of people and machines that they manage. The processes which constitute the financial statements can also
be monitored over time. This will help to find and remove special causes of variation sooner. A firm can do this by setting up a good accounting organisation (AO) and internal control (IC) in which compliance tests play an important role. This will improve the overall quality of the financial statements. An auditor judging the financial statements will always take this into account. The field that uses statistical methods in auditing is called statistical auditing. As said before, the two type of tests we will examine are substantive tests of details and compliance tests.

**Substantive test of details**

Interesting statistical problems occur while an auditor performs a substantive test of details. To estimate the aggregate monetary error amount from a population under inspection a random sample is taken. The book value and the audited value of the sample items are compared. Based on the information from the sample an estimate of the aggregate monetary error amount can be made. Because this estimate is based on a random sample this estimate is accompanied with uncertainty. To express this uncertainty auditors use the upper confidence bound instead of the point estimator to judge if the aggregate monetary error amount exceeds materiality. The populations checked by auditors do usually not contain many items for which the book value and audited value differ. Therefore the distribution of the error amounts is a mixture of two distinguishable distributions. One with discrete probability mass at an error amount of zero, and the other with a ‘continuous’ distribution with non-zero positive and/or negative error amounts. The object of statisticians is to find confidence bounds for the total error amount in the population. Because samples from auditing populations usually contain only a few errors, standard statistical methods do not provide satisfactory inferences for the aggregate monetary error amount in the population. The last few decennia a range of statistical methods have been developed to solve this problem and many of them are treated in this thesis.

Non-standard mixtures of the kind we described above are not only found in auditing, they arise in all kinds of disciplines. These disciplines vary from applications in medicine to applications in forestry (Vännman, 1995). For instance in medicine the discrete variate could be the absence or presence of a tumor and
the continuous variate could be tumor size. In forestry, during the drying process checks can appear on the boards when the boards are dried too fast. Here, the discrete variate is the presence of checks, and the continuous factor is the total area of the checks.

**Compliance test**

Especially useful in compliance tests to control administrative processes is the Average Outgoing Quality Limit (AOQL) method. This method, among others, will be given special attention in this thesis. The research on this method is motivated by a practical case at the IB-Groep in Groningen. We applied the AOQL-method on administrative processes linked to the organising of national exams by the IB-Groep. While using this method certain questions arose. Because we dealt with relatively small populations, we preferred to use the underlying hypergeometric distribution instead of the approximation by the Poisson distribution to find the optimal sample sizes. Therefore, finding an efficient algorithm to find the optimal sample sizes had our special attention. This thesis also addresses other issues connected with the AOQL-method.

### 1.3 Statistician and Auditor

An auditor has to judge whether the financial statements are in accordance with certain specified rules. But this is often more difficult than it sounds. When judging the financial statements of a company an auditor also has to deal with soft numbers. Soft numbers are based on predictions and estimations. There can be many reasons why a company would come to other values of these numbers than the auditor would. But usually the auditor and the company will come to a compromise and the audit report will say that the specified rules are met. But a user of the financial statements does not know about the discussions between the auditor and the company and this could have great consequences. This problem and possible solutions are dealt with in Van Manen (1990). The auditor has to create an image of the truth based on the evidence that is available. A statistician is an expert in trying to uncover truth from data that are uncertain. Auditor and statistician both have to deal with soft numbers. It is obvious that the statistician
could use his expertise in some cases to help the auditor in his search for the truth.

Another aspect is that populations to be audited are so large that it is not possible to inspect these populations entirely. One can only partially inspect these populations by taking samples. Statisticians are experts on the subject of taking samples, based on the role of chance (sample design) they can examine whether deviations are due to chance or to systematical errors.

1.4 Outline

The remainder of this thesis consists of five chapters. Chapter 2 gives an overview of the accepted statistical methods that are used in auditing. This chapter will give an overview of the statistical methods used in statistical auditing and the problems that auditors encounter in using these methods. A distinction is made between the methods used in compliance testing and the methods an auditor uses in a substantive test of details. This chapter will deal with methods that are well-known within auditing as well as more recently developed methods. Chapter 3 describes the use of the AOQL-method, which will play an important role in subsequent chapters and in a practical case at the IB-Groep. Chapter 4 shows some properties of the hypergeometric distribution and how to find confidence limits for a proportion of items with a certain characteristic in a population. It also shows how we can calculate hypergeometric probabilities in an efficient and accurate way. These properties will be very helpful in Chapter 5 and 6. Chapter 5 discusses the AOQL-method in more detail. We will discuss a modified version of AOQL, i.e. the Expected Outgoing Quality Limit (EOQL) method. To find the optimal sample size associated with the EOQL-method often a Poisson approximation for the underlying hypergeometric distribution is used, which is allowed if the sample size is small compared to the population size. However, this chapter develops properties of EOQL that use the exact underlying distribution, i.e. the hypergeometric distribution, and hence can always be applied. These properties will be essential to Chapter 6, because this chapter develops an algorithm to find the optimal sample sizes in specific cases, but also an algorithm is developed to generate tables from which the optimal sample size can be determined quite easily. The method based on the hypergeometric distribution
is called the Exact Expected Outgoing Quality Limit (EEOQL) method. Some technical aspects about the choice of some of the parameters of the EEOQL-method are addressed and finally a comparison between the AOQL-method, the EOQL-method, and the EEOQL-method is given.
Chapter 2

Statistical Auditing

This chapter will give an overview of the statistical methods used in statistical auditing and the problems that auditors encounter in using these methods. A distinction is made between the methods used in compliance testing and the methods an auditor uses in a substantive test of details. This chapter will deal with methods that are well-known within auditing as well as more recently developed methods.

2.1 Compliance Testing

A compliance test aims to check whether the rules followed by the auditee (for instance an organisation, a department or a person under inspection) comply with the rules set by a certain higher authority (for instance government or management), hence the name compliance testing. Therefore, in compliance testing the auditor is interested in a certain characteristic of a population. This characteristic is uniquely defined and population items either possess this characteristic or they do not. For instance, such a characteristic can be a set of rules which have to be obeyed. Chapter 3 provides examples we used in our research at the IB-Groep, but one can also think of purchase and cost invoices that have to comply with certain rules. Sometimes it is very important that items do comply with this set of rules and no items are allowed that do not comply. The sampling method used to check whether a population contains no errors is called discovery sampling. Let us consider a population of \( N \) items. Suppose \( M \) is the number of incorrect items in this population that do not comply with a set of certain rules, then the
following hypothesis testing problem can be formulated: test the null-hypothesis
\[ H_0 : M = 0 \]
versus the alternative hypothesis
\[ H_A : M > 0. \]

Notice that the probability of making a type I-error, i.e. incorrectly rejecting \( H_0 \),
equals zero because populations that contain no errors must result in samples
that contain no errors. Let \( K \) denote the number of errors in the sample. We
can use the sample size, \( n \), to restrict the probability of making a type II-error,
i.e. incorrectly not rejecting \( H_0 \). In case all items of the population would be
inspected, this probability would be zero, under the assumption that no mistakes
are made while inspecting the items. But usually a certain probability of making
a type II-error is accepted. Suppose \( M > 0 \) is the number of incorrect items in
the population, then the probability of making a type II-error is given by
\[ \beta = P\{K = 0|n, M\}. \]

Dependent on the choice of the critical value \( M^* \), we accept for all \( M \geq M^* \)
a probability of at most \( \beta_0 \) of making a type II-error. After we have chosen the
values of \( \beta_0 \) and \( M^* \), at least in principle we can calculate the sample size that
fulfils the condition above. Here the underlying hypergeometric distribution
can be used or we can use either binomial or Poisson approximations to calculate the
sample size. In case of the hypergeometric distribution we can use an iterative
procedure to calculate the sample size by finding the smallest value of \( n \) that
satisfies the following condition
\[ \frac{(N-M^*)}{n} \cdot \frac{1}{\binom{N}{n}} \leq \beta_0. \]

Using the binomial approximation we get a sample size of
\[ n \geq \frac{\log \beta_0}{\log (N - M^*) - \log N}, \]
and using the Poisson approximation gives
\[ n \geq -\frac{N \cdot \log \beta_0}{M^*}. \]
If any errors are found while performing the discovery sampling, then we have to assess the seriousness of the situation. We can calculate an exact upper confidence limit for the fraction $p$ of errors in the population. Again the underlying hypergeometric distribution can be used or an approximation (binomial, Poisson or normal). Chapter 4 extensively describes how to find confidence sets for a proportion. For stratified samples the computations are more complicated. Wendell and Schmee (1996) propose a method for exact inference for proportions from a stratified finite population.

In discovery sampling we could use the sequential method that stops as soon as we find one error or we stop as soon as we have inspected $n$ items, where $n$ is determined in the way described above. This method will give the lowest expected sample size for all possible values of $M^*$ and $\beta_0$ (Ghosh, 1970). In discovery sampling the value of $M^*$ is usually small in practice, hence the expected sample size of this sequential procedure will be close to $n$. Therefore, in discovery sampling this sequential procedure will only be a little bit more efficient compared to the procedure with fixed sample size.

While performing discovery sampling also other (minor) errors may be found. Confidence intervals based on these errors provide insight in the number of such errors in the population. A separate check could be performed on these minor errors or other characteristics considering the following testing problem with the null-hypothesis

$$H_0 : M \leq M_l$$

versus the alternative hypothesis

$$H_A : M \geq M_u,$$

with $M_l \leq M_u$. Because minor errors occur more frequently than major errors and minor errors are probably always present it is not sensible to use the null-hypothesis of absence of errors in the population that we used in case of testing for major errors. We restrict the probability of the type I-error and the probability of the type II-error to $\alpha_0$ and $\beta_0$. Notice that here the null-hypothesis and the alternative hypothesis are not collectively exhaustive. For values of $M$ between $M_l$ and $M_u$, the so-called indifference region, we cannot make a type I-error or a type II-error but the probability of acceptance decreases for increasing values
of $M$ and lies between $\beta_0$ and $1 - \alpha_0$. To perform this test we have to set the values of four parameters, namely $M_l$, $M_u$, $\alpha_0$ and $\beta_0$, and we have to choose $n$ and the number of erroneous items allowed in the sample, $c$, in such a way that the restrictions on the probabilities of type I- and type II-error are satisfied. To give a meaningful basis for setting the values of these parameters is difficult in practice. That is why testing for minor errors is usually done in combination with the testing of major errors or by providing confidence intervals for the number of (or proportion of) minor errors in the population.

The method above can also be interpreted in terms of quality control (see among others Wetherill and Brown (1991)). The type I-error, i.e. incorrectly rejecting the null-hypothesis, is called the consumer risk and the type II-error, i.e. not rejecting the null-hypothesis if it is false, is called the producer risk. The level $M_l$ is called the producer quality level and $M_u$ is called the consumer quality level. For each sampling plan $(n, c)$ we can plot the probability of not rejecting the null-hypothesis, $P\{K \leq c|n, c\}$, against the number of items in error in the population. By not rejecting the null-hypothesis, we conclude that the population is of sufficient quality. The curve we find is called the OC-curve (operating characteristic). A sampling plan $(n, c)$ is determined by guaranteeing that the probability of accepting the population is at least equal to $1 - \alpha_0$ if $M$ does not exceed $M_l$ (the producer risk does not exceed $\alpha_0$), the acceptance region, and by guaranteeing that the probability of accepting the population does not exceed $\beta_0$ if $M$ equals the consumer quality level, $M_u$, or exceeds it (the consumer risk does not exceed $\beta_0$), the rejection region. If the number of items in error in the population lies between the producer quality level and the consumer quality level, the indifference region, then the probability of accepting the population decreases as $M$ moves from $M_l$ to $M_u$. A graphical summary of the above can be found in Figure 2.1.

The procedures considered above are often used to test for errors of a qualitative nature, but with these procedures we can also acquire insight in errors of a quantitative nature. For instance we could apply these procedures to errors that exceed a certain amount. If this error is considered to be a major error we can test for the occurrence of this error and else we could provide confidence intervals for the number of these quantitative errors in the population.

Of course also Bayesian methods can be used in compliance testing. A
2.1. Compliance Testing

Comparison of classical and Bayesian interpretations of statistical compliance tests in auditing can be found in Johnstone (1997). Also Steele (1992) discusses Bayesian methods that can be used in compliance testing in auditing. Meeden (2003) uses a hierarchical Bayes model to find an upper confidence limit for the proportion of items that are in error in a stratified finite audit population.

Using the procedures above, the quality of a population can be established after the population was produced. Auditors often prefer procedures that enable them to inspect and improve the quality of the population while the population is being produced. A popular method that auditors use is the AOQL-method. This method guarantees that the population is of a certain predefined quality by taking samples and find and improve errors in the sample. If a sample contains too many errors, then the entire population is inspected and all errors in the population are corrected. Thus, after inspection of the entire population the population
contains no errors, where perfect inspection is tacitly assumed. The main part of this thesis discusses the AOQL-method. Chapter 3 discusses a practical case at the IB-Groep. Chapter 6 discusses some theoretical issues connected with the AOQL-method. How to find the optimal sample size without using a Poisson approximation for the underlying hypergeometric distribution plays an important role.

2.2 Substantive Test of Details

In a substantive test of details data is collected to find the aggregate monetary error amount. This error in the financial statements is considered to be material if knowledge of this error would affect a decision of a reasonable user of the statements. It is obvious that unambiguous quantitative criteria to find the materiality are not always available and subject to discussion. Materiality is a very important subject in auditing because it is used to decide on which methods are to be used to audit the financial statements and it is also used to evaluate potential and actual errors in the financial statements. If these errors exceed the materiality, then the auditor will ask the client to modify the financial statements.

2.2.1 Classical methods

This subsection will give a concise description of some classical methods that are used in auditing to find the aggregate monetary error; for a more detailed discussion see Cochran (1977). According to these methods items are sampled from the population. Each of the population items contains a certain number of monetary units which add up to a certain amount. This amount is called the book amount or recorded amount. The items in the sample are inspected by the auditor and the auditor establishes the number of monetary units this item should have consisted of. This amount is called the audited amount or correct amount. The difference between the book amount and the audited amount is called the error amount. The sum of the book values of all items in the population is denoted by $Y$ and the sum of the audited values of these items is denoted by $X$. An auditor is interested in the total error amount found by taking the difference between the total recorded amount and the total audited amount. Therefore, this difference is
given by

\[ D = Y - X. \]  \hspace{1cm} (2.2.1)

Notice that the total book amount is known to the auditor. Finding point estimators and confidence intervals for the total audited amount will also supply us with point estimators and confidence intervals for the total error amount. The same notation in lower cases is used in denoting the audited amount, the book amount and the error amount of the \( k \)th item in the sample, namely, \( x_k \), \( y_k \), and \( d_k \), respectively.

**Mean-per-unit estimator**

A sample of \( n \) items is taken and the auditor establishes the audited amount of these items, \( x_1, \ldots, x_n \). The sample mean of these audited amounts can be used as an estimator for the population mean. Hence, the audited amount of the population can be estimated by multiplying the number of items in the population, \( N \), by the sample mean, \( \bar{x} \). This gives the unbiased estimator

\[ \hat{X} = N \bar{x}. \]

An unbiased estimator of the variance of \( \hat{X} \) is

\[ s^2_\hat{X} = \frac{N(N-n)}{n} s^2_x, \]

where \( s^2_x \) is the sample variance of the \( x_i \)'s. A \((1 - \alpha)\)-confidence interval can be found by using the asymptotic normality of the sample distribution. This estimator is imprecise, because the variance of the audited amount can be large. Also a problem arises when the sample contains no errors. In this case all distinct samples of the same size will provide a different estimate and confidence interval. An auditor would like to see that every sample of the same size, containing no errors, gives the same results.

**Ratio estimator**

To improve the precision of this mean-per-unit estimator we can include an auxiliary variable that is correlated with the audit value. Here, the book value is the most logical candidate. Not only are the book amount and the audited amount of
an item highly correlated, very often they are the same. We could use the ratio of
the mean of the sampled audit values and the mean of the sampled book values to
estimate the total audit amount of the population. This gives the so-called ratio estimator

\[ \hat{X} = \frac{\bar{x}}{\bar{y}} = \hat{Q}Y, \]

which is generally biased. This bias, however, becomes negligible as \( n \) becomes
large. If the relation between \( x_i \) and \( y_i \) is reflected by a straight line through
the origin and the variance of \( x_i \) given \( y_i \) about this line is proportional to \( y_i \),
then it can be shown that the ratio estimator is a best linear unbiased estimator.

The following biased estimator can be used to estimate the variance of the ratio estimator

\[ s^2_{\hat{X}} = \frac{N(N-n)}{n}(s^2_x - 2\hat{Q}s_{xy} + \hat{Q}^2s^2_y), \]

in which \( s^2_y \) equals the sample variance of the \( y_i \)'s, and \( s_{xy} \) is the sample covari-
ance. Especially in relatively small samples the biases in \( s^2_{\hat{X}} \) turn out to be more
serious than the biases in \( \hat{X} \). To reduce this bias the jackknife method can be
applied. Frost and Tamura (1982) showed that when error rates are not too small
the jackknife method gives a better performance.

**Regression estimator**

Suppose an approximate linear relation between book value and audit value ex-
ists in which the line does not go through the origin. In this case we could try an
estimate based on the linear regression of \( x_i \) on \( y_i \), instead of using the ratio of
the two variables. This linear regression estimate is given by

\[ \hat{X} = N\bar{x} + b(Y - N\bar{y}) \quad (2.2.2) \]

in which \( b \) is the estimate of the change in the audit value when the book value
is increased by unity. If we choose the value of \( b \) beforehand, then this estimator
is unbiased and an unbiased estimator of its variance is given by

\[ s^2_{\hat{X}} = \frac{N(N-n)}{n}(s^2_x - 2bs_{xy} + b^2s^2_y). \]

For \( b = 1 \) the regression estimator yields the so-called difference estimator. If
\( b = 0 \), then the regression estimator gives the mean-per-unit estimator and for
2.2. Substantive Test of Details

\[ b = \bar{x} / \bar{y} \] it coincides with the ratio estimator. For values of \( b \) between zero and one (2.2.2) is a weighted average of the mean-per-unit estimator and the difference estimator, and the weight assigned to the difference estimator equals \( b \). The variance of the regression estimator is minimized by

\[
b = \frac{\sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{N} (y_i - \bar{Y})^2},
\]

in which \( \bar{X} \) and \( \bar{Y} \) are the population means of the audit values and the book values, respectively. An effective estimator is likely to be the familiar least squares estimate of \( b \). This is given by

\[
\hat{b} = \frac{s_{xy}}{s_y^2}.
\]

By substituting this estimator in (2.2.2) a biased linear regression estimator is obtained. Standard linear regression theory assumes that the population regression of \( x \) on \( y \) is linear, that the population is infinite, and that the residual variance of \( x \) given \( y \) about the regression line is constant. Here, we are not willing to make these assumptions, and by only using large sample results the following biased estimator for the variance of the regression estimator can be found

\[
s_x^2 = \frac{N(N-n)}{n} s_y^2 (1 - r_{xy}^2),
\]

where \( r_{xy} \) is equal to the sample correlation coefficient (\( r_{xy} = \frac{s_{xy}}{s_x s_y} \)). Cochran (1977) suggests multiplying this by \((n - 1)/(n - 2)\), because the factor \( n - 2 \) is used in standard regression theory and gives an unbiased estimate of the residual variance. When the assumptions from standard linear regression theory are made, an expression of the estimated variance of the linear regression estimator can be found in Kleijnen, Kriens, Timmermans and Van den Wildenberg (1988). They also make a comparison between different ways of computing this variance. Among this methods is also the jackknife, which reduces the bias of the variance estimator. They show that jackknifing enhances the performance, but Frost and Tamura (1982) showed for the ratio estimator that the coverage of the confidence intervals computed with these variances is not acceptable if the error percentages are too low or the error distributions are too skew.
Stratification

Stratification can be used to improve the performance of the estimators described previously. After the auditor has decided on the number of strata and has determined the stratum boundaries, the auditor allocates the sample to the different strata. Auditors will use the recorded amounts of the items in the population in the stratification procedure. These amounts have the advantage of being known and are often closely related to the audited amounts. To allocate the sample items to the different strata, the auditor often uses stratification proportional to the sum of recorded amounts in the strata or Neyman allocation is used. When using stratification the items with a large book value are more likely to be in the sample than items with a smaller book amount. Auditors prefer this, because items with large book amounts are more likely to contain larger deviations. Very often all the items in the top stratum are included in the sample. Other sampling methods, for instance cell sampling and PPS, which also favour the larger book amounts, will be discussed later on in this chapter. Stratification dramatically improves the performance of the mean-per-unit estimator, but not the performance of the confidence interval estimators that use the audited amounts as auxiliary information.

PPS estimator

To favour the items with larger book values to be included in the sample also sampling proportional to book values can be used. This means that not all items have an equal probability of being selected in the sample, but items are included with a probability proportional to their book amount. This method of sampling is also known as probability-proportional-to-size (PPS) sampling. Auditors often refer to this method as monetary-unit-sampling or dollar-unit-sampling (MUS or DUS). MUS has the advantage that items which contain more monetary units have a greater chance of selection than items which contain less monetary units. This is from an auditor’s point of view a pleasant characteristic. However, items with zero book values have no chance of being selected in the sample and should be inspected separately. Also items with audited values that exceed the book values have a smaller chance of selection than preferred. Therefore, if a population contains such items, MUS may not be the correct sampling procedure.
Using MUS, the auditor samples individual monetary units instead of items. To provide an estimator for the total audited amount of the population the auditor inspects the item in which the \( i \)th monetary unit of the sample falls. The auditor uses the proportion of recorded monetary units of the item that are in error to determine by how much this sampled monetary unit is tainted. This proportion is called the taint of this monetary unit. The taint of a sampled item, \( t_i \), is given by

\[
t_i = \frac{y_i - x_i}{y_i}.
\]

Using these taints the following unbiased estimator can be found for the total audited amount of the population

\[
\hat{X} = \frac{Y}{n} \sum_{i=1}^{n} \frac{x_i}{y_i} = Y(1 - \bar{t}).
\]

Its variance can be estimated by

\[
s^2_{\hat{X}} = \frac{Y(Y - n)}{n} s^2_t,
\]

with \( s^2_t \) equal to the sample variance of the taints. If the number of taints that differ from one is not too small, then the sampling distribution is approximately normal. But also this estimator performs poorly in an auditing context. Rohrbach (1993) claims that variance estimates for auxiliary variable estimators consistently underestimate the variance and lead to unreliable upper confidence bounds in audit populations. According to Rohrbach this underestimation is caused by overestimating the correlation between book and audit values. This overestimating of the correlation is caused because audit populations usually contain few errors and if an audit population does have a relatively high error rate, then most errors are of an insignificant amount. Rohrbach proposes a variance augmentation method, which in a simplified form (Swinamer, Lesperance and Will, 2004) gives the following augmented variance estimate for the total audited amount of the population

\[
s^2_{\hat{X}}^* = \frac{Y(Y - n)}{n} \left( \frac{\sum(1 - t_i)^2}{n} - (2 - \Delta / n) \left( \frac{1}{2} \left( \frac{\sum(1 - t_i)^2}{n} - s^2_t \right) \right) \right),
\]

where \( \Delta \) is an adjustment factor. Empirical study of Rohrbach determined \( \Delta = 2.7 \) as the smallest value that consistently provided nominal coverage for a certain error distribution.
2.2.2 Non-standard mixture

The methods that are based on auxiliary information fail to provide confidence levels as planned. What causes this poor performance? In auditing populations there is a proportion of items in which the book value and audit value coincide. There is also a proportion of items $p$ in which the book value and audit value do not coincide. The amount by which they differ can be modelled by a random variable $Z$. Therefore, the error $d$ of an item can be modelled in the following way

$$
d = \begin{cases} 
0 & \text{with probability } 1 - p \\
Z & \text{with probability } p.
\end{cases} \quad (2.2.3)
$$

The distribution of the error amount is a so-called nonstandard mixture of a degenerate distribution and a continuous distribution. The poor performance of auxiliary information interval estimation is caused by this mixed nature of the audit population. The standardized statistic does not follow the $t$ distribution for an auxiliary information estimator. The mixture causes the error distribution to be highly skewed, especially when there are few errors and errors are overstatements (the book value is larger than the audit value). This causes a poor performance of the auxiliary information interval estimator based on the asymptotic normality of the sampling distribution for the sample sizes usually used in auditing. The confidence limits for the total monetary error tend to be small using statistical techniques based on the approximate normality of the estimator of the total monetary error (Frost and Tamura, 1986). The upper confidence limit tends to be unreliable and the lower confidence limit usually is too conservative.

2.2.3 Deviating levels of confidence

In auditing practice, auditors are often more interested in obtaining lower or upper confidence limits than in obtaining two-sided confidence intervals. Independent public accountants are very often concerned in estimating the lower confidence bound for the total audited amount. An auditor wants to avoid overestimating this bound because of the potential legal liability that may follow from this. Giving a lower confidence limit for the total audited amount coincides with giving an upper confidence limit for the total error amount. Classical methods have the tendency to give an upper confidence limit which is too tight and thus
provides a level of confidence which is actually lower than the level supposed. This means that the auditor takes a greater risk than intended. Governmental agencies on the other hand are primarily concerned about the lower confidence limit of the total error amount, because they do not wish to overestimate the difference between the costs reported and what should have been reported. Classical methods tend to give a lower confidence limit which is too low. The true level of confidence is higher than it is supposed to be; the agency is assuming a lower risk than allowed by the policy. Internal auditors can be interested in providing a two-sided confidence interval, and they also would like to provide the nominal confidence level. The examples above show that research was and is still needed to provide better confidence bounds.

2.2.4 Methods using attribute sampling

Methods that take the mixed nature of audit populations into consideration have been developed in the last few decades. These methods use several approaches varying from attribute sampling theory, Bayesian inference, bootstrapping, modelling of the sampling distribution, inequalities for tail probabilities, or combinations of these methods. This subsection will give an overview of some methods that use attribute sampling.

We will begin to describe some methods that use attribute sampling. Under the assumption that the population only contains overstatements and the error amount never exceeds the book amount \((0 \leq D_i \leq Y_i)\), taints will take on values between zero and one \((0 \leq T_i \leq 1)\). Especially populations of accounts receivable are contaminated by overstatements. The maximum of all book values in the population is denoted by \(Y_{\text{max}}\). An auditor is interested in giving an upper confidence bound for the total error amount of these overstatements. For this purpose a sample of \(n\) items is taken. The observed error for any item in the sample is given by (2.2.3), using this model we find for the total error amount

\[
D = NpE(Z) \leq NpY_{\text{max}}.
\]

This inequality holds because the error amount of an item cannot exceed the book value of this item. Suppose the number of errors in the sample is denoted
by \( K \), then a \((1 - \alpha)\)-upper confidence limit is given by
\[
D_u(K; 1 - \alpha) = N p_u(K, 1 - \alpha) Y_{\text{max}},
\]
in which \( p_u(K, 1 - \alpha) \) is a \((1 - \alpha)\)-upper confidence bound for \( p \). This upper bound can be based on either the hypergeometric, binomial, Poisson or, even, normal distribution (see Chapter 4). In case of the hypergeometric distribution also the population size could be included as a parameter for the upper bound, but for the sake of notation we neglect this. Because the real outcomes of \( Z \) are not used in this upper bound for \( D \), this bound can be too conservative and may be improved by using a stratification procedure on the book amount.

Instead of sampling the items, we could also take a sample by sampling monetary units. Stringer (1963) proposed the sampling of monetary units and the Stringer bound resulted from this, but Van Heerden (1961) was the first to propose a method of finding the total error amount using monetary unit sampling (MUS). He assumed that if a sampled monetary unit fell into the first \( X_i \) monetary units, which were considered to be correct, this sampled monetary unit was correct, and if it fell into the last \( D_i \) dollars, which were considered to be the dollars in error, it was in error.

Under the assumptions we previously made, we again use (2.2.3), but \( d \) is now considered to be the taint of a monetary unit instead of being the error amount of an item. Notice that \( 0 \leq Z \leq 1 \), and \( p \) is the proportion of monetary units in error in the population. The total error amount is given by
\[
D = Y p E(Z) \leq Y p.
\]
So, a \((1 - \alpha)\)-upper confidence limit is given by
\[
D_u = Y p_u(K, 1 - \alpha),
\]
with \( K \) equal to the number of non-zero taints in the sample. Although this bound is less conservative than the bound we found by sampling items, it still has the tendency to be too conservative, because it assumes all taints to be equal to one.

### 2.2.5 The cell bound

Methods that try to use the information that not all taints are equal to one, are called combined attributes and variables (CAV) estimation methods. An example
of such a bound is the cell bound, which was introduced by Leslie, Teitlebaum and Anderson (1980). This method assumes that errors are overstatements and that the sample is taken by using MUS with cell sampling. What is cell sampling? Instead of random sampling of \( n \) monetary units, auditors often divide the population in parts, cells, consisting of \( Y/n \) monetary units, and they select randomly from each cell one monetary unit. By doing so the minimal distance between two consecutive sampled monetary units is one, and the maximum distance is \( (2Y/n) - 1 \). Sampling items which are certain to be sampled, items of \( 2Y/n \) monetary units or larger, are inspected separately. This so-called cell sampling gives a more even distribution of the sampled monetary units over the population. Auditors prefer this. It also helps to find the smartest fraud, because the smartest fraud will minimize the chance of discovery by spreading the errors evenly over the population. The smartest auditor maximizes this minimal chance by spreading the sample evenly over the population. Cell sampling has the disadvantage of not being one sample of \( n \) items, but in fact when using cell sampling the auditor takes \( n \) random samples each consisting of one item. Hoeffding (1956) showed that the probability distribution of sample errors can still be approximated by a binomial distribution.

After taking a MUS sample with cell sampling, the \( K \) taints of the \( n \) sampled monetary units are ordered in a decreasing order, \( 1 \geq Z_{(1)} \geq \ldots \geq Z_{(K)} > 0 \). An upper limit for the total error amount in the population is calculated by

\[
D_u = \frac{Y}{n} \text{UEL}_K,
\]

in which \( \text{UEL}_K \), the upper error limit factor for the \( K \)th taint is calculated by an iterative procedure. Starting at \( \text{UEL}_0 = \lambda_u(1 - \alpha; 0) \) the upper error limit factors for the \( i \)th taint can be found with

\[
\text{UEL}_i = \max \left( \text{UEL}_{i-1} + Z_{(i)}, \frac{\lambda_u(1 - \alpha; i)}{i} \sum_{j=1}^{i} Z_{(j)} \right),
\]

in which \( \lambda_u(1 - \alpha; i) \) denotes the upper limit for the Poisson distribution parameter \( \lambda \) when \( i \) errors are observed. See Leslie et al. (1980) for a more detailed description of this bound.
2.2.6 The Stringer bound

A CAV method frequently used is the Stringer bound, which was introduced by Stringer (1963) and elaborated by, among others, Leslie et al. (1980). This heuristic procedure, which assumes that errors are overstatements, has never been satisfactorily explained and not even an intuitive explanation can be found in the literature. The first step in constructing the Stringer bound is to order the \( K \) taints in a decreasing order, \( 1 \geq Z(1) \geq \ldots \geq Z(K) > 0 \), then the Stringer bound is given by

\[
D_u = Y p_u(0, 1 - \alpha) + \sum_{j=1}^{K} Z(j) \left( p_u(j, 1 - \alpha) - p_u(j - 1, 1 - \alpha) \right).
\]

However, several simulation studies show that the Stringer bound is too large, i.e. the actual level of confidence exceeds the nominal confidence level. Indications of this conservatism of the Stringer bound can be found, for example, in Leitch, Neter, Plante and Sinha (1982), Plante, Neter and Leitch (1985), and Reneau (1987). Also Lucassen, Moors and Van Batenburg (1996) present a simulation study that confirms this conservatism, and they also examine several modifications of the Stringer bound based on different rankings of the taints. They examine methods in which the taints are ordered in an increasing order (ITO), a random order (RTO), and according to the corresponding error amounts (ATO) instead of ordering the taints in a decreasing order. They conclude that ITO does not give a proper confidence bound and ATO and RTO provide bounds that are less conservative than the original Stringer bound. Since ATO is a natural method of expressing the auditor’s ideas about the severeness of the misstatements, they prefer the ATO method. According to the authors this method is recommended as an alternative to the Stringer bound in the Deloitte Touche Tohmatsu International manual on Audit Sampling. Bickel (1992) gives some weak fixed sample support to the conservatism of the Stringer bound and claims that the Stringer bound is asymptotically always too large. Pap and Van Zuijlen (1996) show that the Stringer bound is asymptotically conservative for confidence levels \((1 - \alpha)\), with \(\alpha \in (0, 1/2]\), and asymptotically it does not have the nominal confidence level for \(\alpha \in (1/2, 1)\). They also propose a modified Stringer bound which asymptotically does have the nominal confidence level.
Here we only consider overstatements, but suppose that also understatements are present. Meikle (1972) and Leslie et al. (1980) present methods that provide upper confidence limits that take understatements into account. Meikle suggests to subtract the lower confidence limit for the understatements from the upper confidence limit for the overstatements. Leslie et al. suggest to subtract the estimated mean value of the understatements from the upper confidence limit for the overstatements. Grimlund and Schroeder (1988) found the adjustment by Leslie et al. to be reliable and uniformly more efficient than the adjustment of Meikle.

2.2.7 The multinomial bound

The Stringer bound calculates an upper bound for the total monetary error which takes the magnitude of the errors into account and does not require any assumptions about the error distribution. Confidence bounds based on the multinomial distribution also have these properties but these bounds also have distributional properties which are fully known. This bound was first described by Fienberg, Neter and Leitch (1977). They assumed errors to be overstatements and each sampled dollar was placed in one of 101 unit categories corresponding to the magnitude of the error (in cents). Each taint is rounded upward to the nearest whole percent and classified according to its value in cents. If the item has no error, then the sampled dollar is placed in the category of zero cents. This gives categories of 0, 1, ..., 99, and 100 cents, which coincides with taints of 0, 0.01, ..., 1. More generally, there will be \( r + 1 \) categories labelled 0, 1, ..., \( r \). Let \( p_i \) denote the population proportions of the \( i \)th category, with \( 0 \leq p_i < 1 \) and \( \sum p_i = 1 \), and \( W_i \) the observed number of sampled monetary units in category \( i \). If we sample with replacement, then the distribution of \( W = (W_0, \ldots, W_r) \) will be multinomial with parameters \( (n, \mathbf{p}) \), \( \mathbf{p} = (p_0, \ldots, p_r) \). The distribution is approximately multinomial if sampling is without replacement and the sample size is small compared to the total book value. In the original setup of Fienberg et al. the total error amount in the population is given by

\[
D = \frac{Y}{100} \sum_{i=1}^{100} ip_i,
\]  

(2.2.4)
and a point estimator for $D$ is given by

$$\hat{D} = \frac{Y}{100} \sum_{i=1}^{100} i \frac{W_i}{n}.$$ 

To find an upper confidence bound for $D$ a set $S$ is defined in the following way. It contains all possible outcomes which are at least as extreme as the sample outcome. There are many ways to define this set $S$, but taking computational simplicity into account, Fienberg et al. defined this so-called step down set $S$ such that the number of errors associated with an element of $S$ does not exceed the number of errors in the sample outcome and each error amount does not exceed any observed amount. Using $S$, a $(1 - \alpha)$ joint confidence set can be found for $p$. This confidence set consists of all $p$ such that

$$\sum_S \mathbb{P}(W_0 = v_0, \ldots, W_r = v_r) \geq \alpha, \sum v_i = n.$$ 

The confidence upper limit for $D$ is acquired by maximizing (2.2.4) over this confidence set for $p$. Although the true confidence level of this bound is unknown, Plante et al. (1985) showed that it was tighter than the Stringer and cell bounds. Plante et al. also showed that if the line items are in random order, then cell sampling does not have serious effect on the location and symmetry of the sampling distributions of the Stringer and multinomial bound, but it does tend to reduce the variability of the distributions. For the multinomial bound cell sampling tends to give improved coverage.

A lower bound for the overstatement error can be obtained by minimizing (2.2.4), under the same restrictions as above, with respect to a step-up set $S$ which contains outcomes that are as extreme as or more extreme than the observed outcome. This set is the set of outcomes for which the total number of errors is at least as large as the observed number of errors and any individual error cannot be less than the corresponding observed individual error (Plante, Neter and Leitch, 1984).

The multinomial bound can also be used to give a lower confidence bound for understatement errors. The only difficulty is setting the maximum value of the understatement error for a monetary unit. This is far more difficult than setting the maximum value of the overstatement error for a monetary unit. But
once this maximum has been set the procedure is the same as for calculating the multinomial bound for overstatement errors (Neter, Leitch and Fienberg, 1978).

Calculating the multinomial bound is a computational complex and intensive process and soon becomes unmanageable as the number of errors increases. A solution is to make a categorisation which is less refined than the one described here above. Each tainting within such a cluster will be regarded as a tainting with the maximum possible value within this cluster. This will reduce the computational complexity considerably and still the multinomial bound will be tighter than the Stringer bound (Leitch et al., 1982). Leitch et al. propose a clustering that performs very well. Their clustering is a modification of an optimal clustering procedure developed by Fisher (1958).

2.3 Bayesian methods

Many empirical studies have been carried out to discover more about the error distributions of various accounting populations, see e.g. Johnson, Leitch and Neter (1981). Such results, in combination with other knowledge the auditor possesses, can be used by an auditor to make a prediction about the error distribution of certain audit populations. Bayesian inference provides a useful framework to incorporate this knowledge.

2.3.1 Felix and Grimlund method

We will first discuss the parametric Bayesian model of Felix and Grimlund (1977). They applied this method to item sampling, but Menzefricke and Smieliauskas (1984) applied this method also to dollar unit sampling. Suppose a MUS-sample is collected according to the model given by (2.2.3), then the density of the non-zero taints $Z$ is assumed to be normal with mean $\mu_Z$ and variance $\sigma_Z^2$. The precision $h$ is defined as $\sigma_Z^{-2}$. The population proportion of monetary units $p$ is assumed to follow a prior beta distribution with parameters $n_0 p_0$ and $n_0 (1 - p_0)$. The precision follows a gamma prior distribution with parameters $\tau_0/2$ and $\tau_0 \psi_0/2$. The mean of the non-zero taints $\mu_Z$ has a normal prior distribution with mean $\mu_0$ and variance $(hr_0)^{-1}$. The prior expected values of respectively $\mu_Z$, $h$, and $p$ are equal to $\mu_0$, $\psi_0$, and $p_0$. The measure of confidence in the choices of these prior expected values is reflected by the values of $r_0$, $\tau_0$ and
The prior joint distribution of $(\mu_Z, h)$ is a normal gamma distribution and the prior distribution of $p$ is independent of $(\mu_Z, h)$. The posterior distribution of $(\mu_Z, h)$ is again a normal gamma distribution and the posterior distribution of $p$ is again a beta distribution and independent of the posterior distribution of $(\mu_Z, h)$. The marginal distribution of $\mu_Z$ is obtained by integrating out $h$. This gives a Student $t$ distribution. Because we can write the mean error amount per item $\mu_d$ as $\mu_d = p\mu_Z$, the transformation $\mu_Z = \mu_d / p$ can be substituted into the marginal distribution of $\mu_Z$ and by integrating out $p$ the posterior distribution of $\mu_d$ is obtained. This integration does not lead to an explicit solution and has to be done numerically. Expressions for the expectation and variance can be found (Menzefricke and Smieliauskas, 1984), and by using a Student $t$ distribution an approximate upper bound for $\mu_d$ can be found (Swinamer et al., 2004). Notice that this method can also deal with understatements. This method has the disadvantage that only certain prior parameter values can be used if a sample does not contain any errors.

### 2.3.2 Cox and Snell method

Cox and Snell (1979) used an exponential prior distribution for $Z$, with parameter $1/\mu_Z$. Since the exponential distribution can only take on positive values, this method can only deal with overstatements, but it also can handle the situation when no errors are contained in the sample. It is assumed that $1/\mu_Z$ has a gamma density with parameters $b$ and $(b - 1)\mu_0$, where $\mu_0$ is the prior mean of $\mu_Z$ and $b$ specifies the variance of the prior distribution. The probability that a monetary unit is in error, $p$, has a gamma distribution with parameters $a$ and $p_0/a$. The parameter $p_0$ is the prior expected value of $p$ and $a$ controls the variance of the prior distribution. Unlike the previous model, where the number of errors in the sample has a binomial distribution with parameters $n$ and $p$, here a Poisson distribution is used with parameter $np$. The posterior distributions of $\mu_d$ can be shown to be a scalar transformation of an $F$ distribution. Suppose $(z_1, \ldots, z_k)$ are the taints that are found in the sample, then a $(1 - \alpha)$ upper bound for $D$ is calculated by

$$
Y \left(\frac{(k + a)(k\bar{z} + (b - 1)\mu_0)}{(k + b)(n + a/p_0)} \right) \frac{F_{2(k+a),2(k+b);1-\alpha}}{1-\alpha}.
$$
where $\bar{z} = \sum z_i / k$ and $F_{u,v;1-\alpha}$ is the $(1-\alpha)$ quantile of the $F$ distribution with $u$ and $v$ degrees of freedom. Neter and Godfrey (1985) show that this bound is very sensitive to the choice of the prior parameters and this sensitivity does not disappear for larger sample sizes. They also show that it is possible to set these values conservatively in such a way that this bound is still reliable and that it is still significantly tighter than the Stringer bound.

### 2.3.3 Multinomial-Dirichlet bound

Tsui, Matsumura and Tsui (1985) described a Bayesian nonparametric method which has the same setup as the multinomial bound. Here, also only overstatements are assumed and each taint is rounded and classified according to its value in cents (0 to 100 cents). Tsui et al. assume a Dirichlet prior distribution for $p$, namely Dir$(K\alpha_0, \ldots, K\alpha_{100})$ with $K > 0$, $\alpha_i > 0$, $i = 0, \ldots, 100$ and $\sum \alpha_i = 1$. Appendix 2.A gives more detailed information and some properties of the Dirichlet distribution. The $\alpha_i$ represent the best prediction of the auditor of the unknown $p_i$, and large values of $K$ imply that the prior guesses are fairly sharp. $K$ is usually chosen to be much smaller than the sample size to reflect that the sample information is more reliable than the best predictions of the auditor.

The posterior distribution of $p$ is again a Dirichlet distribution with parameters $K'\alpha'_0, \ldots, K'\alpha'_{100}$, where

$$K' = K + n \quad \text{and} \quad \alpha'_i = \frac{K\alpha_i + W_i}{K'}.$$

The exact form of the posterior distribution of $D$ can be obtained but is very complicated to work with. Appendix 2.A shows how $p$ can be simulated by simulating 101 independent gamma distributed variables, with the $i$th variable $X_i$ having a gamma distribution with shape parameter $K'\alpha'_i$ and scale parameter equal to 1. By using this property the empirical distribution of $D$ can be obtained by simulating a large number (10,000) of these $p$'s and for each of these the total error amount $D$ can be calculated by

$$D = \frac{Y}{100} \sum_{i=1}^{100} i p_i = \frac{Y}{100} \sum_{i=1}^{100} \frac{X_i}{\sum_{j=0}^{100} X_j}. $$
The percentiles of this empirical distribution can be used as the true percentiles of the posterior distribution of $D$. As an upper bound the $100(1 - \alpha)$ percentile can be used. The associated interval is said to have credibility $(1 - \alpha)$. Notice that this does not mean that this interval automatically is a $100(1 - \alpha)$ confidence interval from a classical point of view. Tsui et al. show that if a prior Dirichlet distribution is used with $K = 5$, $\alpha_0 = 0.8$, $\alpha_1 = \alpha_2 = \ldots = \alpha_{99} = 0.001$, and $\alpha_{100} = 0.101$, then the acquired bounds have good repeated sampling performance, with good frequentist properties. Tsui et al. also show that a good approximate posterior distribution of $\mu_d = D/Y$, the mean dollar taint in the population, is obtained by taking a beta distribution with mean and variance equal to the theoretical posterior distribution of $\mu_d$. The mean of the theoretical posterior distribution of $\mu_d$ can be shown to be equal to

$$E(\mu_d) = \frac{1}{100} \sum_{i=1}^{100} i \alpha_i',$$

and

$$\text{Var}(\mu_d) = \frac{\sum_{i=1}^{100} i^2 \alpha_i' - \left(\sum_{i=1}^{100} i \alpha_i'\right)^2}{10^4(K' + 1)}.$$

This leads to an approximation of the theoretical posterior distribution of $\mu_d$ by a beta distribution $B(a, b)$ with parameters

$$a = E(\mu_d) \left(\frac{E(\mu_d)(1 - E(\mu_d))}{\text{Var}(\mu_d)} - 1\right), \quad (2.3.1)$$

and

$$b = (1 - E(\mu_d)) \left(\frac{E(\mu_d)(1 - E(\mu_d))}{\text{Var}(\mu_d)} - 1\right). \quad (2.3.2)$$

An upper bound for $D$ is obtained by finding the $100(1 - \alpha)$ percentile of this beta distribution and multiplying this with the total book value $Y$.

Matsumura, Tsui and Wong (1990) extended the multinomial-Dirichlet model to situations in which both understatement and overstatement errors are possible. They assume that the maximum over- and understatement error in a dollar unit is 100 percent. For overstatement errors this is by the definition of a tainting almost always the case, but for understatement errors this is somewhat different. Leslie et al. (1980) claim that in practice understatement errors rarely
exceed 100 percent. Matsumura et al. treat understatement errors that do exceed 100 percent as 100 percent errors. The model could also be modified to allow for larger errors. Here, errors are classified in one of 201 cells, namely $-100, -99, ... , 0, ... , 99, 100$ percent errors. The distribution of the vector $W = (W_{-100}, \ldots, W_{100})$ is again assumed to be multinomial with parameters $n, p$. A Dirichlet prior distribution for $p$ is assumed, namely $\text{Dir}(K\alpha_{-100}, \ldots, K\alpha_{100})$ with $K > 0$, $\alpha_i > 0$, $\sum \alpha_i = 1$, $i = -100, \ldots, 100$. Matsumura et al. recommend a Dirichlet distribution as prior distribution for $p$ with $K = 1.5$, $\alpha_{-100} = \ldots = \alpha_{-1} = 0.05/100$, and $\alpha_0 = \ldots = \alpha_{100} = 0.95/100$. The posterior distribution of $p$ is again a Dirichlet distribution with parameters $K'\alpha'_{-100}, \ldots, K'\alpha'_{100}$, where $K' = K + n$ and $\alpha'_{i} = \frac{K\alpha_{i} + W_{i}}{K'}$.

The exact form of the posterior distribution of $\mu_d$ can be obtained but is very complicated to work with. Since $(1 + \mu_d)/2$ takes values between 0 and 1, we can use a beta distribution $B(a, b)$ with expectation and variance that match the expectation and variance of the true posterior distribution of $(1 + \mu_d)/2$. The values of $a$ and $b$ can be found by replacing $E(\mu_d)$ by $0.5(1+E(\mu_d))$ and $\text{Var}(\mu_d)$ by $0.25\text{Var}(\mu_d)$ in equations (2.3.1) and (2.3.2). If the $100(1 - \alpha)$ percentile of this beta distribution is denoted by $B_{1-\alpha}$, then the approximate $100(1 - \alpha)$ percentile of the posterior distribution of $D$ is given by $Y(2B_{1-\alpha} - 1)$. This approximate percentile has a credibility of $100(1 - \alpha)$ percent, but Matsumura et al. state that further research is necessary to evaluate the achieved confidence levels of this procedure.

In auditing it can also be of importance to provide a lower bound for the overstatement error. Matsumura, Plante, Tsui and Kannan (1991) studied the performance of the multinomial bound and the multinomial-Dirichlet bound. They concluded that if computational considerations are not important, then the multinomial bound should be used, otherwise the multinomial-Dirichlet method should be used with a prior Dirichlet distribution with $K = 5$, $\alpha_0 = 0.8$, $\alpha_1 = \ldots = \alpha_{90} = 0.001$, and $\alpha_{100} = 0.101$.

Another nonparametric Bayesian method that uses the multinomial distribution as the data-generating model can be found in McGray (1984).
2.3.4 A distribution-free method

The Bayesian parametric approaches above specify a parametric model for the distribution function $F_Z$ of $Z$ for specification of the prior distribution for $\mu_Z$. Tamura (1988) treated $F_Z(z)$ as distribution-free with Ferguson’s Dirichlet process with parameter $\alpha(z)$ as the prior. Instead of using an exact form of the conditional error distribution as prior distribution, this method allows the auditor to describe the expected form of the error distribution. Let $F_{0,z}$ be the auditor’s best prior prediction of the conditional distribution of the error. This prediction can vary from a standard parametric distribution to observations without any formal structure. Using the Dirichlet process with parameter $\alpha(z) = \alpha_0 F_{0,z}$ as prior for $F_Z$, it follows that $P\{Z \leq z\} = F_Z(z)$ has a beta distribution $B(\alpha(z), \alpha_0 - \alpha(z))$. The auditor uses the finite weight $\alpha_0$ to reflect the uncertainty about his prediction. The prior expectation of $F_Z(z)$ is then given by $F_{0,Z}(z)$. Suppose $k$ errors are found, say $v = z_1, \ldots, z_k$, then given these observations the posterior distribution of $F_Z(z)$ is again a Dirichlet process with parameter $\alpha(z|v) = \alpha(z) + k F_{k,Z}(z)$, where $F_{k,Z}(z)$ is the empirical distribution function of $z$. The distribution of $\mu_d$ can be derived numerically from the distribution of $\mu_Z$, which can also be derived numerically. The exact derivation can be found in Tamura (1988, pp. 4-5). An upper bound can be found by multiplying $Y$ with the $100 \cdot (1 - \alpha)$ quantile of the distribution of $\mu_d$. Laws and O’Hagan (2000) extend the model by splitting up the error region into a number of error categories and a Dirichlet-multinomial model is used for the rates of errors. Independent Dirichlet process models are now used for the values of taints in these error categories. Using the distribution of the book values, Monte-Carlo simulation is applied to find the distribution of the total error amount. Laws and O’Hagan (2002) adapt this model for multilocation auditing.

2.3.5 Other methods

Dworin and Grimlund (1984) approximate the sampling distribution by a gamma distribution. To estimate the parameters of this gamma distribution the method of moments is applied. They introduce a hypothetical taint which is treated as an additional observed taint. The value of this hypothetical taint depends on the nature of the audit population. Using this hypothetical taint, the bound they
calculate tends to be larger than the bound calculated without using this hypothetical taint. Dworin and Grimlund state that this conservatism compensates for the lack of information about the error distribution. This method, which can deal with both over- and understatement errors, provides a confidence level close to the stated one and is about as tight as the multinomial bound. Another parametric bound is introduced by Tamura and Frost (1986). They model the tainting distribution by the power function density and they use a parametric bootstrap to find the upper bound of the total error amount.

A solution to gain additional information about the error distribution is bootstrapping. This method was used by Talens (1998) at the Informatie Beheer Groep to find the total error amount in an audit population which was related to organising exams in the Netherlands. This method gives very tight bounds but unreliability is an issue. Another solution is a reduction of the problem to inequalities for tail probabilities of certain relevant statistics. Bentkus and Van Zuijlen (2001) present results concerning this topic. According to them, Hoeffding (1963) inequalities are currently the best available. The bounds they find can be extended to non-i.d.d. cases and to settings with several samples. This method gives very reliable bounds but also this method tends to be conservative. Inequalities that give a better approximation of the tail probabilities will lead to tighter bounds. A combination of the bootstrap and the use of Hoeffding inequalities can be found in Howard (1994). Helmers (2000) developed a method, as part of a research project with the Statistical Audit Group of PricewaterhouseCoopers in Amsterdam, that provides a new upper confidence bound for the total error amount in an audit population, where line-item sampling is appropriate. His method consists of two stages, the first stage uses an empirical Cornish-Fisher expansion, and the second stage uses the bootstrap to calibrate the coverage probability of the resulting interval estimate.

Statistical analysis in auditing very often presumes that the auditor does not make any errors while inspecting. Despite the professional skill of auditors this presumption may not always hold. In a case where only qualitative errors were the subject of interest, an item in an audit population is either correct or incorrect and no interest is being paid to the possible error amount, Raats and Moors (2004) and Wille (2003) dealt with the problem of finding estimators and upper confidence limits for the error fraction in the audit population. They did assume
that the first check was followed by a second check by an expert auditor who is
infallible. By using Bayesian decision theory, but using different parameteriza-
tions, they found estimators and upper confidence limits for the error fraction in
the audit population.

2.4 Conclusions

This chapter gives an overview of statistical methods used in auditing. The 1989
National Research Council’s panel report on Statistical Models and Analysis in
Auditing already mentioned “...the generally scattered and ad hoc nature of the
existing methodology”. Today this description still is very appropriate and con-
firmed by Swinamer et al. (2004). They compared 14 different bounds currently
used in statistical auditing with each other and no bound was superior in terms
of reliability and efficiency. The fact that every profession, including auditing,
experiments, either motivated or not, with a variety of methods can possibly be
blamed for the generally scattered and ad hoc nature of the existing methodology.
This way a system with lack of structure evolves, but by feedback of experiences
this system seems to suffice in most situations. Of course, it is essential to ex-
amine if the methods do not only seem to suffice, but if they really do suffice.
Therefore, a good dialogue between auditor and statistician is very important.
Also the 1989 National Research Council’s panel report on Statistical Models
and Analysis in Auditing gives an outline of statistical problems in auditing that
need attention. At present many of these problems still need attention and further
research. Moreover, it stays a challenge for researchers to find the method that
is superior to all other methods.

2.A Dirichlet distribution

If \( p = (p_0, \ldots, p_r) \) follows a Dirichlet distribution \( \text{Dir}(K\alpha_0, \ldots, K\alpha_r) \) with
parameters \( K\alpha_0, \ldots, K\alpha_r \), with \( K > 0, \alpha_i > 0, i = 0, \ldots, r \) and \( \sum \alpha_i = 1 \),
then the probability density function of \( p \) has the form

\[
f(p) = \frac{\Gamma(K)}{\prod_{i=0}^{r} \Gamma(K\alpha_i)} \prod_{i=0}^{r} p_i^{K\alpha_i-1}.
\]
The distribution of $p_i$ is a beta distribution $B(K \alpha_i, K(1 - \alpha_i))$ with

$$E(p_i) = \alpha_i \quad \text{and} \quad \text{Var}(p_i) = \frac{\alpha_i(1 - \alpha_i)}{K + 1}.$$  

The family of Dirichlet distributions is a class of conjugate priors for the multinomial distribution. Let $W = (W_0, \ldots, W_r)$ have a multinomial distribution $(n, p)$ and the prior distribution for $p$ is Dirichlet $\text{Dir}(K \alpha_0, \ldots, K \alpha_r)$, then the posterior distribution of $p$ is again a Dirichlet distribution with parameters $K \alpha_0 + W_0, \ldots, K \alpha_r + W_r$. This can be rewritten as a Dirichlet distribution with parameters $K' \alpha'_0, \ldots, K' \alpha'_r$, where

$$K' = K + n \quad \text{and} \quad \alpha'_i = \frac{K \alpha_i + W_0}{K'}.$$

Suppose $X_i, i = 0, \ldots, r$ has a gamma distribution with shape parameter $K \alpha_i$ and scale parameter equal to 1. The joint distribution of the proportions

$$p_i = \frac{X_i}{\sum_{j=0}^r X_j}, \quad \text{for} \quad i = 0, \ldots, r$$

is the Dirichlet distribution $\text{Dir}(K \alpha_0, \ldots, K \alpha_r)$. For further results we refer to e.g. Fang and Zhang (1990).
Chapter 3

A Practical Case

3.1 Introduction

The Department of Examservices (ED) of the Informatie Beheer Groep (IB-Groep) at Groningen in the Netherlands provides a large range of services to students. The public critically monitors the IB-Groep, because in the past they had some organisational problems, due to too optimistic governmental policy, that led to a poor quality of services to her customers. In this period the IB-Groep became a scapegoat of the public. Recently the quality of services really has improved and the IB-Groep is trying hard to retain and even to improve upon this quality. Because of the past, the IB-Groep lies under the magnifying glass of the public. Thus, quality is very important to the IB-Groep.

One of the many tasks of the IB-Groep is the organisation of public exams. These tasks are carried out by ED. This chapter designs a quality control procedure for some of the administrative processes of ED. This research covers the years 1999 through 2001. Why did the manager of ED commission this research?

• The first reason was a reorganisation (Hero). Due to Hero the audit department had to hand over her control tasks for ED to the employees of ED.

• The second reason had to do with the disappearance of the annual check by the accountants of the Ministry of Education (AD, abbreviation of the dutch term ‘Accountants Dienst’). AD does not report about the performance of ED separately any more, but reports over a much broader sector.
ED is just a small part of this sector. The AD will suffice with just a global check instead of the more thorough check that AD used to do. For a more detailed description of this check see Talens (1998).

- The third reason, the most important one according to ED, was the increased awareness of quality thinking caused by the quality charter, market thinking, and responsibility at employee level.

By implementing new better quality control procedures, ED decreases the control risk considerably. This risk is very large at the moment, because with the exception of a superficial check by coworkers no quality control measures are taken.

The check AD used to do was mainly focused on financial errors. The new procedure should, in light of quality thinking, also check on non-financial aspects. Another problem related to the check AD used to do, was that too much time expired before executed activities were inspected. This time interval usually took more than three months. If this time could be shortened, problems would be noticed and dealt with earlier.

**Current developments**

This subsection sketches some recent developments at the the IB-Groep after this research took place. The organisation of the IB-Groep is always under pressure, and with every new executive changes are made in the organisation to comply with the ideas of the new executive. The new executive introduced a new management model that should increase the capacity. The departments in their current shape are reorganised to become ‘production’ departments (again). Within these departments there is no place for specialists and the specialised tasks for all production departments are to be performed by a special department (bedrijfsbureau). Notice that this is an opposite course from the course that was brought into action by Hero and one of the reasons why the manager of ED did commission this research is no longer valid. Also auditors are placed within this department and are no longer part of the new production departments and most control tasks are to be performed by this department too.

Other courses of action are taken within the IB-Groep to improve the quality of their services. The IB-Groep has started using the the INK (Instituut Neder-
3.2 The COSO Report

The report Internal Control-Integrated Framework of the Committee of Sponsoring Organizations of the Treadway Commission (COSO-report) appeared in 1992. This report provides guidelines for setting up and judging internal control systems. According to the COSO-report internal control is a process executed by a supervisory organisation, management or remaining personnel which is designed to acquire reasonable certainty to what extent the following objectives are achieved:

- efficiency and effectiveness of company processes;
- reliability of the financial reporting;
- compliance of legislation and other rules.

Especially the third part plays an important role in setting up and judging internal control systems in our research at the IB-Groep. Also in accordance with the COSO-report each internal control system can be divided into five components

- monitoring;
- information and communication;
- control activities;
- risk assessment;
control environment.

To maintain the quality of an internal control system it should be constantly monitored by its supervisors. Changes in the process could mean that some procedures are redundant or other procedures are necessary. The first two reasons in Section 3.1 that gave rise to this research, are a clear example of this. The quality of information should be high, otherwise decisions based on this information could be wrong, and communication should be effective. The control activities are those activities that affect the administrative organisation and the internal control measures that have been taken to reach the objectives set by management. In our research this part plays a very important role. Risk assessment means that management tries to find and analyse internal and external risks that could jeopardize reaching the objectives set by management. The control environment is determined by the attitude of personnel, especially management, towards the risk control systems. The third reason that gave rise to this research (see Section 3.1) is due to a change in this environment.

3.3 Exam Services

The exams ED organises are divided over five different clusters. Each of these clusters organises one exam or several exams. These clusters carry out many processes. Some of these processes are the same for every exam (e.g. travel declarations) other processes are restricted to one exam or a group of exams. An example of the latter was the second correction of school and state exams. Originally this research would develop a new control procedure for the process of checking travel declarations at the SPD-exam. The SPD-exam is an exam for people in the administrative sector. People concerned with the organisation of exams (members of exam committees, correctors, or employees of ED) fill in these travel declarations. ED inspects them and if they are correct, then the amount declared will be paid. Inevitably, errors are made while inspecting these travel declarations. As long as not too many errors are made, this is not considered to be a problem. ED demands that a certain level of errors in the population is not exceeded and a certain quality of the population is maintained. Noticing and correcting errors at an early stage certainly helps maintaining this quality.
3.4 Requirements and solution

The manager of ED had good reasons to implement new quality control procedures. Consultation with the ED-employees led to the following requirements for quality control procedures.

- Due to hero control tasks should largely be carried out by the ED-employees themselves instead of specialized auditors.

- Normal procedures do not account for correcting errors during the checking time. ED would like a procedure that does account for correcting errors during the checking time.

- If errors are made, ED wants to find the cause of these errors quickly. When this cause has been found, ED can remove this cause and prevent new errors.

- ED would like a procedure that checks on non-financial errors as well, such as time expired before certain processes are finished.

- The number of errors made in a process is not allowed to exceed a certain level defined by management.

A procedure that satisfies these requirements is the AOQL (Average Outgoing Quality Level)-procedure or the revised version of AOQL: EOQL (Expected Outgoing Quality Limit). For a more detailed and technical description of the AOQL/EOQL-procedure see Chapter 6. The difference between the two methods is of a technical nature. The sample design does not differ between the two methods (only sample sizes differ). A short description of the design for the AOQL/EOQL-procedure is given below.

- First management defines the population of the items under investigation (e.g. all travel declarations throughout a year).

- Management defines what is considered to be an error in a process. For instance, for a travel declaration a list of rules can be made this declaration has to satisfy. If at least one of the rules is not satisfied, the declaration is in error.
Management also defines the number of errors allowed in the population after inspection. Usually management provides a proportion of items allowed to be in error. This limit is called the Average Outgoing Quality Limit and will be denoted by \( P_l \).

The population has to be divided into a number of subpopulations (e.g. all travel declarations throughout a month or a week) in a sensible way. There are both statistical as well as efficiency reasons not to choose the number of subpopulations too small. Natural divisions of the population into subpopulations are preferred (e.g. monthly reports). It is also wise to separate periods that are likely to contain more errors (e.g. periods with lots of temporary workers) from other periods.

From every subpopulation a random sample of size \( n \) is taken. Chapter 6 describes how sample size and the number of errors allowed in the sample, \( k_0 \), are found. If this number of errors in the sample is exceeded, the subpopulation has to be checked integrally.

All errors, that are detected during the sampling and during possible integral checks, have to be corrected. Flawless inspections are assumed.

### 3.5 General considerations implementing AOQL

This section gives an outline of the environment in which the AOQL-procedure has to be implemented. In this sketch three different aspects are considered: technical considerations, philosophical considerations, and organisational considerations.

#### 3.5.1 Practical considerations

Employees are very busy during exam moments, therefore additional manpower is needed to execute the control tasks as well. If ED would decide not to check during the exam moments, the cause of the errors would remain unnoticed. These errors would continuously be made in the process, which causes a lot of corrections afterwards (many subpopulations would have to be integrally corrected). Otherwise this problem would not arise, because ED could remove the
cause of the errors in time. For processes that are similar in different clusters a solution could be to combine these processes in one AOQL-procedure, because the exams in different clusters are organised on different dates. Thus, if one cluster is very busy another cluster has capacity to do the inspection activities. It is important to let grow a feeling of joined responsibility for the activities of each other. Criticism should be dealt with in a positive manner and should lead to an improvement of the quality. One has to be very careful that severe arguments between clusters will not arise, causing a bad working environment.

The employees of ED will check their own activities, that is the activities of their direct colleagues. Within the framework of separating production activities and control activities, it is important that an employee does not inspect his or her own production work and it is also important to inspect the control activities of the employees in order to prevent cheating. An employee of ED especially concerned with control tasks within ED could do this. To maintain the quality of the quality control tasks these tasks should not always be performed by one single person, but these tasks should be done by various employees. This monitoring of the control activities is also in line with the COSO-report.

3.5.2 Philosophical consideration

As pointed out in the introduction, inspection of the processes was completely focused on financial errors. Thinking in financial terms has become a way of life at ED. They are very persistent in this way of thinking. For instance, when management had to set an error norm for travel declarations they came with the following norm: the total error amount should not exceed 1% of the total amount. Statements like only in error if error > 50,- could be found in the error definition made by management. Defining an error in this way would lead to an underestimation of the total error amount if many little (< 50,-) errors would occur. Moreover, defining the norm as a percentage of items in error of the total amount of items would be more sensible. In this way also non-financial aspects can be taken into account. If overall quality improves, this will lead to a lower unjust amount in the population. This will appear in the global checks of the AD. If AD is convinced that a good quality control procedure is used, AD will inspect less strict and conclude that the financial statements are presented
3.5.3 Organisational considerations

At ED a group of employees is democratically chosen to decide on policy matters within ED. This group is called the steering committee (BG). The manager of ED should be present at meetings of the BG and they have weekly meetings.

Originally this research would develop a new control procedure for travel declarations for the SPD-exam. This process should be a pilot for other processes at ED. If this would work according to the wishes of ED, other processes of ED would follow. For the department concerned with the organisation of the SPD-exam had just undergone a reorganisation, the BG decided to take the process of second corrections at school- and state exams as the pilot. The following meeting they decided not to, because the next exam moment would be in August. This would give personnel problems, because many employees did plan their holiday in this month. Therefore, they did not have the capacity for the extra inspection activities. Again a week later they finally decided to have the pilot at SPD-enrolments. The BG can be criticised for not being very decisive. Because employees of the clusters are in the BG, they probably hesitated to have the pilot at their own cluster, because the pilot required extra work. Moreover, at the meetings where the pilot was on the agenda the manager was not present to make a final decision. Finally an employee of the SPD cluster agreed to have the pilot at a process of her cluster. Unfortunately SPD-enrolment is a process with pure non-financial aspects. It would have been more interesting to have a pilot with combined financial and non-financial aspects. Of course the relative simplicity of the process could be the reason why the employee of SPD agreed to have the pilot, because not many errors are expected and not much extra work is expected by having to perform inspections of the entire subpopulation. Nonetheless implementing the AOQL-procedure and performing the control activities accompanied with this procedure still demands an additional effort of the employees and therefore it is commendable that they agreed to have the pilot.

The above example shows that an effective communication between management and personnel was not present in this case and although the increased awareness of quality thinking was one of the incentives for this research the
attitude towards the internal control procedures is somewhat hesitant and the control environment can be improved. The effectiveness of communication and the hesitant attitude towards internal control need improvement because they are components on which an internal control system is judged according to the COSO-report.

3.6 The first pilot

The pilot took place at SPD-enrolment. People who would like to do the SPD-exam have to enrol. On the enrolment form they have to fill out some characteristics about themselves. Employees of the SPD-cluster put these data into the computer. We use an improvement of the original AOQL-method, the aforementioned EOQL-procedure to inspect this process. This is not a very complicated process, and beforehand not many errors are expected. So the main reason for doing this process with the EOQL procedure is to gain experience about EOQL in practise and make employees familiar with EOQL.

3.6.1 Sample scheme

The population consists of all enrolments that are processed in March, 2000. Subpopulations are defined as the enrolments processed per day, unless the total amount of enrolments is under 100, then subsequent days are combined until the number of enrolments exceed 100. An enrolment is wrong if:

- the time needed to process the enrolment exceeds three weeks,
- the exam subjects are put wrongly into the computer (code 4).

If a mistake is made in the other characteristics the enrolment is not in error, but these errors will be kept in a database as experience data. These errors are coded in accordance with the questions on the SPD-enrolment form, see Table 3.1. Management has set a norm of 1% of the total amount of enrolments that may be in error. The EOQL procedure should be carried out with the following instructions:

- Determine the number of enrolments in a subpopulation.
Table 3.1. Definition of the error codes at SPD-enrolment.

<table>
<thead>
<tr>
<th>Code</th>
<th>Error made in:</th>
<th>Code</th>
<th>Error made in:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>surname</td>
<td>2a</td>
<td>address</td>
</tr>
<tr>
<td>1b</td>
<td>name</td>
<td>2b</td>
<td>mailing address</td>
</tr>
<tr>
<td>1c</td>
<td>gender</td>
<td>1d</td>
<td>date of birth</td>
</tr>
<tr>
<td>1e</td>
<td>place of birth</td>
<td>1f</td>
<td>native country</td>
</tr>
<tr>
<td>1g</td>
<td>phone number by day</td>
<td>1h</td>
<td>giro or bank account</td>
</tr>
<tr>
<td>1i</td>
<td>current college</td>
<td>1j</td>
<td>enroled as: student/extraneus</td>
</tr>
<tr>
<td>1k</td>
<td>did a previous exam</td>
<td>1l</td>
<td>date + signature</td>
</tr>
<tr>
<td>1m</td>
<td>did a previous exam</td>
<td>1n</td>
<td>special</td>
</tr>
<tr>
<td>1o</td>
<td>date + signature</td>
<td>1p</td>
<td>circumstances</td>
</tr>
</tbody>
</table>

- Set the number of errors allowed in the sample. This is zero if the previous subpopulation was not integrally checked, otherwise this is one. The starting value is zero.

- If the number of errors allowed in the sample is zero, then the sample size equals 37. Otherwise determine the sample size from Table 3.2, taken from Simons, Van Batenburg and Kriens (1989).

- Take a random sample, check the enrolments and if the number of errors allowed is exceeded, check the subpopulation integrally. Correct all errors!

The sample sizes used in this first pilot were found using the EOQL-method in which a Poisson approximation is used for the underlying hypergeometric distribution, see Simons et al. (1989).

Simons et al. also show that the optimal sample size if the number of allowed errors in the sample is equal to zero, using this procedure, equals $e^{-1}/P_i$ (so with $P_i = 0.01$ this becomes 37). Notice that this does not depend on the population size. The sample sizes we use if the number of errors allowed in the sample equals one also originate from Simons et al., see Table 3.2.
3.6. The first pilot

Table 3.2. Sample size \( n \) for a given subpopulation size and number of errors allowed \( k_0 = 1 \) for \( P_l = 1\% \).

<table>
<thead>
<tr>
<th>Population size</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 100 and ( \leq 150 )</td>
<td>70</td>
</tr>
<tr>
<td>&gt; 150 and ( \leq 200 )</td>
<td>73</td>
</tr>
<tr>
<td>&gt; 200 and ( \leq 250 )</td>
<td>75</td>
</tr>
<tr>
<td>&gt; 250 and ( \leq 300 )</td>
<td>76</td>
</tr>
<tr>
<td>&gt; 300 and ( \leq 400 )</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 3.3. Results of the first pilot with the EOQL-method at SPD-enrolment.

<table>
<thead>
<tr>
<th>no.</th>
<th>subpopulation size ( (N_s) )</th>
<th>sample size ( (n) )</th>
<th>errors in criteria</th>
<th>other errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97</td>
<td>37</td>
<td>0</td>
<td>lb</td>
</tr>
<tr>
<td>2</td>
<td>380</td>
<td>37</td>
<td>0</td>
<td>lg</td>
</tr>
<tr>
<td>3</td>
<td>123</td>
<td>37</td>
<td>0</td>
<td>lb</td>
</tr>
<tr>
<td>4</td>
<td>132</td>
<td>37</td>
<td>0</td>
<td>1a, 1g, 1h</td>
</tr>
<tr>
<td>5</td>
<td>132</td>
<td>37</td>
<td>0</td>
<td>1b, 1e</td>
</tr>
<tr>
<td>6</td>
<td>171</td>
<td>37</td>
<td>0</td>
<td>none</td>
</tr>
<tr>
<td>7</td>
<td>294</td>
<td>37</td>
<td>0</td>
<td>1a, 1e, 1f(2), 1g</td>
</tr>
<tr>
<td>8</td>
<td>133</td>
<td>37</td>
<td>0</td>
<td>1a, 1b, 1e, 1g, 1h, 2a</td>
</tr>
<tr>
<td>9</td>
<td>93</td>
<td>37</td>
<td>0</td>
<td>1b, 1c, 1g, 1i, 2a</td>
</tr>
<tr>
<td>10</td>
<td>594</td>
<td>37</td>
<td>0</td>
<td>1b, 1e, 1h</td>
</tr>
<tr>
<td>11</td>
<td>191</td>
<td>37</td>
<td>0</td>
<td>1b, 1c(6), 1g(3), 2a</td>
</tr>
<tr>
<td>12</td>
<td>549</td>
<td>37</td>
<td>0</td>
<td>1b(2), 1e, 1g(3)</td>
</tr>
<tr>
<td>13</td>
<td>110</td>
<td>37</td>
<td>0</td>
<td>1b(2), 1g(2), 1h</td>
</tr>
</tbody>
</table>
If the previous subpopulation has to be checked entirely, the number of errors allowed in the sample from the subpopulation under inspection is equal to one instead of zero. This strategy is also described in Simons et al. and by doing so the assumption is made that the total amount of work will be reduced. A more detailed discussion about the choice of the number of errors allowed in the sample can be found in Chapter 6.

A computer program has been written by the author to help the SPD-employee with this procedure, see Figure 3.1.

### 3.6.2 Results

The results of the first pilot can be found in Table 3.3 and Table 3.4. In thirteen different subpopulations the number of errors allowed in the sample was never exceeded; no integral inspection of any subpopulation took place. The conclusion could be drawn that the process of SPD-enrolment is executed in a very precise way.

**Table 3.4.** Totals of the errors not concerning the main criteria.

<table>
<thead>
<tr>
<th>Type</th>
<th>Total</th>
<th>Type</th>
<th>Total</th>
<th>Type</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>4</td>
<td>1g</td>
<td>14</td>
<td>2a</td>
<td>3</td>
</tr>
<tr>
<td>1b</td>
<td>11</td>
<td>1h</td>
<td>4</td>
<td>2b</td>
<td>0</td>
</tr>
<tr>
<td>1c</td>
<td>8</td>
<td>1i</td>
<td>1</td>
<td>3a</td>
<td>0</td>
</tr>
<tr>
<td>1d</td>
<td>0</td>
<td>1j</td>
<td>0</td>
<td>3b</td>
<td>0</td>
</tr>
<tr>
<td>1e</td>
<td>5</td>
<td>1k</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1f</td>
<td>2</td>
<td>1l</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.6.3 Evaluation

SPD-employees involved with the EOQL-procedure commented on the amount of time the quality control tasks took them. Especially during busy times the quality control tasks would sometimes be neglected. Of course this is not a problem which is specific for the EOQL-method. Any other quality control procedure will give the same problem. The exams ED organise take place at certain dates and activities are concentrated around these dates. Therefore, employees
3.6. The first pilot

Figure 3.1. Screen shot of the input of the sampled SPD-enrolments in the computer program used in the second pilot.

are very busy during exam moments. The following solutions could help to solve the problems of lack of time to carry out the quality control tasks.
• Postpone the quality control tasks until after the busy times connected with
the exam moments, but in such a way that the process has been inspected
and improved before the next exam moment. This has the disadvantage of
not noticing causes of errors during the process. It will cost more time and
money to inspect and improve afterwards. Inspections afterwards can also
be to the disadvantage and cause damage to customers of ED.

• Hiring extra temporary personnel during the exam period. Except for this
being an expensive solution, this will probably also lead to more errors.
Unskilled employees are more likely to make more mistakes, because they
lack the experience.

• If processes between different clusters largely coincide, then the control
tasks can be spread among the different clusters. Very often exam mo-
ments are on different dates for the different clusters.

• If processes between different clusters do not coincide, then the control
tasks could still be spread among the different clusters. This has the dis-
advantage that employees have to get acquainted with new processes. But
there are also some advantages. During other busy periods employees
from less busy clusters can assist the busy clusters and it increases the
collective consciousness of the employees.

3.7 The second pilot

After the first pilot ED decided to have another pilot at SPD-enrolment to ac-
quaint more employees with the EOQL-procedure and to become more familiar
with this procedure. They changed the error definition and \( P_l \) was set at 3%. An
enrolment was considered to be in error if a mistake was made in the processing of

• the current college (code 1i),

• the enrolment as student or extraneus (code 1j),

• place of exam (code 3b),
3.7. The second pilot

- the exam subjects (code 4).

Due to organisational problems the inspection was conducted some time after the actual enrolment. Although we regretted the delay of the inspection we still thought it would provide valuable information about the process.

In the previous pilot forms were randomly chosen, by just picking the correct number of forms from the stack of forms in the subpopulation without looking. Of course, this is a very crude way of random sampling. In the new pilot every form was given a unique number with a stamping machine. The computer program was adjusted and the random number generator of the computer program was used to randomly select the forms that should be in the sample.

We also made progress on finding the optimal sample size in the EOQL-method using the hypergeometric distribution. In fact we found a simple, elegant and fast way to find the optimal sample size. This method is described in Chapter 6. We called this method the Exact Expected Outgoing Quality Limit (EEOQL) method. The computer program was also adjusted such that it could calculate the optimal sample sizes using this new algorithm.

### 3.7.1 Results

This EEOQL-procedure inspected 17 subpopulations. Each subpopulation contained between 300 and 400 enrolments. We used a limit of zero errors in the sample. Using the improved method of finding the optimal sample size, this gave a sample size of 12 enrolments for all samples in the subpopulations where we used a limit of zero errors in the sample. If a previous sample was inspected integrally, thus an error was found, a limit of one error in the sample was used. Only one subpopulation was inspected integrally (the ninth). This subpopulation of 369 enrolments showed another two errors. Therefore, in the tenth subpopulation a limit of one error in the sample was used, resulting in a sample size of 27 enrolments. No errors were found in this sample.

A remark has to be made about the way the EEOQL-procedure was performed. We agreed to check on errors as defined above and other errors would also be reported on as experience data. These experience data are used to give an overall view of the precision and quality of the activities carried out by the workers. The inspection of these data is part of the control activities as described
in the COSO-report. The computer program gives the employee the opportunity to do so. In spite of these agreements the employees of SPD-enrolment decided not to report on the other errors after the second subpopulation and also they decided only to take the criteria $1_i$ and $1_j$ into account. Reason to do so was that for all other errors besides $1_i$ and $1_j$, they did not found it logical to improve these errors after the time that had passed since the actual enrolment took place. After the candidates enrol they receive a survey on which the information they supplied is registered. If the information was processed wrongly by the employees of SPD-enrolment the candidates can inform the employees of SPD-enrolment about this. Experience learns that candidates indeed are very keen on improving information that is directly linked with the exam, but are not so precise concerning background information. Criteria $1_i$ and $1_j$ are examples of this kind of background information. Because the actual enrolments are checked with the data in the computer there is some sense in this way of reasoning. The data in the computer do not represent the original input anymore. This original input has already been corrected for errors.

An evaluation showed that the employees who conducted the EEOQL-procedure were enthusiastic about this procedure and found it an excellent way to control their activities. However, they reported on the following drawbacks of the procedure.

- The EEOQL-procedure gives difficulty in planning the control activities. If the subpopulation is rejected, then the subpopulation has to be inspected integrally. Beforehand it is not known if this will happen, which complicates the planning of the EEOQL-procedure.

- Some employees told that they felt tempted to cover up mistakes that would lead to an integral inspection, because this would save them a lot of work.

- For numbering the enrolments an old-fashioned stamping machine was used. The employees found the use of this machine ergonomically irresponsible, and the machine was too noisy (they got complaints from employees from a floor beneath).

The stamping machine was just used in this experimental pilot and if the
EEOQL-procedure would be implemented on a regular basis, we are sure that more modern methods are available. We know that employees from ED take great pride in the quality of work they deliver, but to fight the temptation of covering up errors an examination of the inspection tasks is very important. An employee of ED especially concerned with control tasks within ED could do this. Otherwise, the aforementioned rotation of control tasks over the different clusters could be a solution. We agree that the EEOQL-procedure complicates the planning of future activities, but we are also sure that the organisation of ED is flexible enough to react to last-minute changes of their planning, because this is a phenomenon that ED is not unfamiliar with.

After these two pilots, manuals for the EEOQL-procedure and for using and adapting the computer program were handed over to ED. We thought they had gathered enough knowledge to conduct the EEOQL-procedure independently and plans were initiated to implement the EEOQL-procedure at other processes.

Shortly after these pilots the activities conducted for the SPD-exam were terminated because the IB-Groep unexpectedly lost the contract of organising this exam.

### 3.8 Conclusions

The EEOQL-procedure turned out to be an excellent tool for the internal control toolbox of the IB-Groep. It satisfied the objectives management had set for a new internal control procedure. Also it was practically performable and took the objectives of the COSO-report into account. Components of the COSO-report that needed further improvement were the effectiveness of the communication and the attitude towards internal control procedures. Despite the positive assessment of the EEOQL-procedure by the IB-Groep, it has not yet become a standard method in the internal control toolbox of ED. Notwithstanding that ED recognized the need to improve their internal control procedures in general and notwithstanding the positive assessment of the EEOQL-procedure, they did not improve and innovate the internal control procedures of their processes by implementing the EEOQL-procedure. Partly this could be put down to new developments in the organisation like the implementation of the new management model. At the moment this implementation takes place and the EEOQL-procedure does
not have the highest priority at the moment. The EEOQL-procedure will be reassessed after the new management model is operational.
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

\[
P\{K = k|n, M, N\} = \frac{\binom{M}{k} \frac{N-M-k}{n-k}}{\binom{N}{n}}.
\] (4.1.1)

We know that

\[ E(K) = n \cdot \frac{M}{N} \]

and

\[ \text{Var}(K) = \frac{n \cdot M \cdot (N - M) \cdot (N - n)}{N^2 \cdot (N - 1)} . \]

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:

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

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

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

\[
P\{K = k|n, M, N+1\} = \frac{(N-n+1) \cdot (N-M+1)}{(N-M-n+k+1) \cdot (N+1)} \times
\]
\[
\times P\{K = k|n, M, N\}
\]
4.1. Properties of the hypergeometric distribution

\[
P(K = k | n, M, N) = P(K = n - k | n, N - M, N) \\
= P(K = M - k | N - n, M, N) \\
= P(K = N - M - n + k | N - n, N - M, N)
\]

\[
P(K \leq k | n, M, N) = 1 - P(K \leq n - k - 1 | n, N - M, N) \\
= 1 - P(K \leq M - k - 1 | N - n, M, N) \\
= P(K \leq N - M - n + k | N - n, N - M, N)
\]

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.

\[
P(K = k | n, M, N) = P(K = k | M, n, N).
\]

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

Notice that \( P(K = k | n, M, N) \) 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,

\[
\Lambda(n, M, N) = P(K \leq k_0 | n, M, N) = \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M}{n-k} \binom{n}{k}.
\] (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 \( \Lambda \) as a function of \( n, M, \) and \( N \). We will discuss some properties of \( \Lambda(n, M, N) \) that are especially useful in Chapter 6.
Chapter 4. Hypergeometric Distribution

**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 $M \leq k_0$ or $n \leq k_0$.
3. $\Lambda(n, M, N) = 0$ if and only if $M > N - n + k_0$.
4. Let $M \in \{0, \ldots, N - 1\}$, then
   \[
   \Lambda(n, M + 1, N) = \Lambda(n, M, N) - \frac{M}{k_0} \binom{N-M-1}{n-k_0-1},
   \]
   or, equivalently,
   \[
   \Lambda(n, M + 1, N) = \Lambda(n, M, N) - \frac{n}{N} \cdot P(K = k_0|n-1, M, N-1)
   \]
   for $n \in \{1, \ldots, N\}$.
5. Let $M \in \{0, \ldots, N - 1\}$, then $\Lambda(n, M + 1, N) \leq \Lambda(n, M, N)$. The inequality is strict if and only if $k_0 \leq M \leq N - n + k_0$ and $n > k_0$.
6. Let $n \in \{0, \ldots, N - 1\}$, then $\Lambda(n + 1, M, N) \leq \Lambda(n, M, N)$. The inequality is strict if and only if $k_0 \leq n \leq N - M + k_0$ 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

\[
\binom{N}{n} \Lambda(n, M + 1, N) = \sum_{k=0}^{k_0} \binom{M+1}{k} \binom{N-M-1}{n-k}
\]

\[
= \sum_{k=0}^{k_0} \left( \binom{M}{k} + \binom{M}{k-1} \right) \binom{N-M-1}{n-k}
\]

\[
= \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M-1}{n-k} + \sum_{k=0}^{k_0-1} \binom{M}{k} \binom{N-M-1}{n-k-1}
\]

and hence
\[
\binom{N}{n} A(n, M, N) = \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M}{n-k} \\
= \sum_{k=0}^{k_0} \binom{M}{k} \left( \binom{N-M-1}{n-k} + \binom{N-M-1}{n-k-1} \right) \\
= \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M-1}{n-k} + \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M-1}{n-k-1} \\
= \binom{N}{n} A(n, M+1, N) + \binom{M}{k_0} \binom{N-M-1}{n-k_0-1}.
\]

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. \(\square\)

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

\[
\lambda(n, M, N) = \begin{cases} 
\frac{A(n, M+1, N)}{A(n, M, N)} & \text{if } M < N - n + k_0, \\
0 & \text{if } N - n + k_0 \leq M \leq N - 1.
\end{cases} \tag{4.1.3}
\]

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

2. \(\lambda(n, M, N) = 0\) if and only if \(M \geq N - n + k_0\).
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

\[
g(n, M, N) = \binom{N}{n} \binom{N-M}{M-k_0}^{-1} > 0.
\]

4. Let \( M \in \{0, \ldots, N - 1\} \), 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 \( n, M \in \{0, \ldots, N - 1\} \) 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

\[
g(n, M, N) = \binom{N}{n} \binom{N-M}{M-k_0}^{-1} = \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M}{n-k_0-1} = \sum_{k=0}^{k_0} \binom{k_0}{k} \frac{k_0!}{k!} \cdot (N - M) \cdot \prod_{h=k+1}^{k_0} \frac{N - M - n + h}{M - h + 1} \cdot \prod_{j=k}^{k_0} \frac{1}{n - j}.
\]

(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. \( \square \)

**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. $\left\lfloor K \cdot \frac{N+1}{n} \right\rfloor$. 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 $k \in \{0, \ldots, n\}$, 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 $M \in \{0, \ldots, N\}$, 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,

$$P\{M_C(K) \ni M \mid M\} \geq 1 - \alpha,$$

for every $M \in \{0, \ldots, N\}$. (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. $M_C(0), \ldots, M_C(n)$ in such a way that (4.2.1) is satisfied for every $M \in \{0, \ldots, N\}$. 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.

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

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

$$P\{M_C(K) \ni M \mid M\} = \sum_{k \in R(M)} P\{K = k \mid n, M, N\}. \quad (4.2.2)$$
Remember that $K \sim H(n, M, N)$.

Now, suppose we construct for every $M \in \{0, \ldots, N\}$ sets $R'(M)$ with values of $k$ such that $\sum_{k \in R'(M)} P\{K = k \mid n, M, N\} \geq 1 - \alpha$ 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

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

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 $M_C(K) = \{M \mid K \in R(M)\}$ will always be a confidence interval. If we observe $K = k$, then the lower and upper confidence interval limits are given by

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

and

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

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 $\beta$ and $\gamma$ 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

$$P\{K = k(1) | M\} \geq P\{K = k(2) | M\} \geq \ldots \geq P\{K = k(n) | M\}.$$

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

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

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($M_C(K)$) the lower confidence bound $M_L(K)$ and max($M_C(K)$) 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
4.2. Confidence sets

these gaps is seldom.

Using this method ties can occur. Ties occur when

\[ P\{ K = k_{(n_k(M))} | M \} = P\{ K = k_{(n_k(M)+1)} | M \} . \]

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.

\[ M \]

**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

\[
P[K \geq k | p = p_L] = \frac{\sum_{i=k}^{n} \binom{Np_L}{i} \binom{N(1-p_L)}{n-i}}{\binom{N}{n}} \geq \frac{\alpha}{2},
\]
and the largest integer value of \( N \cdot p_U \) such that

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

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

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

and

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

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

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

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 $F_{1-rac{\alpha}{2}}(\cdot, \cdot)$ 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

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

and

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

The following relationship relates the tail of a Poisson distribution with the tail of a $\chi^2$-distribution.

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

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

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

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 \( \chi^2 \)-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

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

and

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

Solving these equations gives the following confidence interval

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

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

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

### 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 \( P\{K = k_0|n - 1, M, N - 1\} \). 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 \( P\{K = k_0|n - 1, M, N - 1\} \). 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 \( k_0 \leq M \leq N - n + k_0 - 1 \) and \( k_0 + 1 \leq n \leq N - 1 \), then

\[
\Lambda(n, M, N) = \Lambda(n, M + 1, N) + C_1(n, M, N)
\]
with
\[ C_1(n, M, N) = \frac{M - k_0 + 1}{M + 1} \cdot \frac{N - M - 1}{N - n + k_0} \cdot C_1(n, M + 1, N). \]

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
\[ C_1(n, M, N) = \binom{M}{k_0} \binom{N - M - 1}{n - k_0 - 1} \cdot \binom{N}{n}, \]
and
\[ \Lambda(n, M + 1, N) = \Lambda(n, M + 2, N) + C_1(n, M + 1, N) \]
with
\[ C_1(n, M + 1, N) = \binom{M+1}{k_0} \binom{N-M-2}{n-k_0-1}. \]

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
\[
C_1(n, N - n + k_0, N) = \Lambda(n, N - n + k_0, N) - \Lambda(n, N - n + k_0 + 1, N)
= \Lambda(n, N - n + k_0, N)
= \frac{n! (N - n + k_0)!}{k_0! N!}.
\]

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
\[ C_2(n, M, N) = \binom{n-1}{k_0} \binom{N-n}{M-k_0-1}, \]
and
\[ \Lambda(n - 1, M, N) = \Lambda(n - 1, M + 1, N) + C_1(n - 1, M, N) \]
with
\[ C_1(n - 1, M, N) = \frac{M}{k_0} \binom{N-M-1}{n-k_0-2}. \]

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
\[ \Lambda(n, M, N) = \Lambda(n, M + 1, N) + C_1(n, M, N) \]
4.3. Computing the hypergeometric distribution

with

\[ C_1(n, M, N) = \binom{M}{k_0} \binom{N-M-1}{n-k_0-1} \binom{N}{n} \]

and

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

with

\[ C_2(n, M + 1, N) = \binom{n-1}{k_0} \binom{N-n}{M-k_0} \binom{N}{M+1} \]

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

\[ C_2(n + 1, M, N) = \binom{n}{k_0} \binom{N-n-1}{M-k_0} \binom{N}{M} \]

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

\[ C_1(n + 1, M - 1, N) = \binom{M-1}{k_0} \binom{N-M}{n-k_0} \binom{N}{n+1} \]

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

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

\[
= \Lambda(n + 1, M, N) + \frac{M}{n + 1} \cdot C_1(n + 1, M - 1, N)
\]

\[
= \Lambda(n + 1, M - 1, N) - C_1(n + 1, M - 1, N) + \frac{M}{n + 1} \cdot C_1(n + 1, M - 1, N)
\]

\[
= \Lambda(n + 1, M - 1, N) + \frac{M - n - 1}{n + 1} \cdot C_1(n + 1, M - 1, N)
\]

\[
\square
\]

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

<table>
<thead>
<tr>
<th>( M \setminus n )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.9821</td>
<td>0.9286</td>
<td>0.8214</td>
<td>0.6429</td>
<td>0.375</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.9286</td>
<td>0.7571</td>
<td>0.5</td>
<td>0.2143</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.8214</td>
<td>0.5</td>
<td>0.1786</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.6429</td>
<td>0.2143</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.375</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

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

note that \( \Lambda(3, 8, 8) = 0 \). Again by using Property 4.3.1, part 1 we can calculate
4.3. Computing the hypergeometric distribution

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

\[
\Lambda(3, 6, 8) = \Lambda(3, 7, 8) + C_1(3, 6, 8)
\]

\[
= 3/8 + \frac{6 - 2 + 1}{6 + 1} \times \frac{8 - 6 - 1}{8 - 6 - 3 + 2} \times 3/8
\]

\[
= 3/8 + 5/7 \times 3/8 = 3/8 + 15/56 = 9/14 \approx 0.6429.
\]

We can repeat this procedure until we have found \( \Lambda(3, 3, 8) \) and 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:

\[
\Lambda(4, 6, 8) = \Lambda(3, 6, 8) - \frac{8 - 6}{8 - 4 + 1} \times \frac{6 - 2}{4 - 2 - 1} \times C_1(3, 6, 8)
\]

\[
= 9/14 - 2/5 \times 4 \times 15/56 = 9/14 - 3/7 = 3/14 \approx 0.2143.
\]

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 \( \Lambda \) 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

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

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

\[
\Lambda(n, M, N) = \iota' \cdot P(n, M, N).
\] (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 \( j = 0, \ldots, k_0 \)

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

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

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.
Chapter 5

Properties of AOQL

5.1 Introduction

AOQL is identified with a sampling method which guarantees that the quality of the population investigated meets certain minimum quality requirements after an inspection that allows for correction of errors found. Originally AOQL was designed for industrial purposes but nowadays this method is also used in auditing and in the control of administrative processes.

The Average Outgoing Quality Limit (AOQL) sampling system was developed by Dodge and Romig (1959). It was originally developed for industrial purposes. Applications in auditing can be found in for instance Arkin (1974), Cyert and Davidson (1962), and Kriens and Dekkers (1979). Veenstra and Kriens (1982) applied the method to the control of administrative processes. Van Batenburg and Kriens (1988) criticized the statistical derivation by Dodge and Romig and they presented a modification. This modified version of the AOQL-method is sometimes called the EOQL-method (Expected Outgoing Quality Limit). Simons et al. (1989) developed an algorithm for finding the sample size and the critical level of items in error in the sample for the EOQL-method. However, this algorithm uses a Poisson approximation for the hypergeometric distribution involved, which is allowed if the sample size is small compared to the population size. This chapter presents theory that fully exploits the underlying hypergeometric distribution and hence can always be applied. It also exploits the results we found in Chapter 4 on the hypergeometric distribution, which are essential to the effectiveness of our methods.
The following idea underlies the AOQL-method. Consider a population of size \( N \) consisting of good and bad items. The unknown number of bad items will be denoted by \( M \). The fraction of errors before inspection is given by \( p = \frac{M}{N} \).

A random sample of size \( n \) is taken without replacement from this population and the items in this sample are inspected and corrected. If \( K \), the number of bad items in the sample, exceeds a certain critical level \( k_0 \), then all items in the population have to be inspected. Therefore, after inspection the quality of the population will really have improved, unless no errors are found in the sample. The expected fraction of remaining defects will be denoted by \( \pi \). This sampling method requires \( n \) and \( k_0 \) to be determined cost effectively such that \( \pi \) does not exceed a certain predefined level \( P_l \).

We assume that the inspection is perfect. This means that all errors are detected and that no correct items are marked as bad items, see e.g. Moors (2000), Raats and Moors (2000), and Raats and Moors (2004) for imperfect inspections.

Obviously, \( \pi \) depends on four parameters, namely \( k_0, n, M \), and \( N \). This chapter always assumes that \( k_0, n, M \in \{0, 1, \ldots, N\} \), where \( N \geq 1 \). The choice of \( k_0 \) will be an important feature. For the time being, \( k_0 \) will be considered fixed. We shall look at \( \pi \) as a function of \( n, M, \) and \( N \) for fixed \( k_0 \). We denote the expected fraction of errors after inspection as \( \pi(n, M, N) \).

We use the same extended definition of the binomial coefficients that we used in Chapter 4. Using this definition we do not have to incorporate the usual domain for \( k \), the number of errors in the sample, namely \( k = 0, \ldots, n \) with the restriction that \( k \geq n - (N - M) \) and \( k \leq M \).

Since sampling is done without replacement, the number \( K \) of defective items in the sample follows a hypergeometric distribution with parameters \( n, M \) and \( N \), thus \( K \sim \mathcal{H}(n, M, N) \). Thus for non-negative integers \( k \) we have

\[
P(K = k) = \binom{M}{k} \binom{N-M}{n-k} \binom{N}{n}^{-1}.
\]

We will also use \( \Lambda(n, M, N) \) and \( \lambda(n, M, N) \) in this chapter, especially in proving the theorems we will introduce. They were already defined in Chapter 4 in the following way

\[
\Lambda(n, M, N) = P\{K \leq k_0|n, M, N\} = \sum_{k=0}^{k_0} \binom{M}{k} \binom{N-M}{n-k} \binom{N}{n}^{-1},
\]

(5.1.2)
and
\[
\lambda(n, M, N) = \begin{cases} 
\frac{\Lambda(n, M+1, N)}{\Lambda(n, M, N)} & \text{if } M < N - n + k_0, \\
0 & \text{if } N - n + k_0 \leq M \leq N - 1.
\end{cases} 
\] (5.1.3)

The fraction of errors after inspection \( p_a(K) \) is given by
\[
p_a(K) = \begin{cases} 
\frac{M - K}{N} & \text{if } K = 0, \ldots, k_0, \\
0 & \text{otherwise}.
\end{cases} 
\] (5.1.4)

From (4.1.1) and (5.1.4) it follows
\[
\pi(n, M, N) = E(p_a(K)) = E \left( \frac{M - K}{N} \middle| K \leq k_0 \right) \cdot \Lambda(n, M, N)
\] (5.1.5)

Some obvious cases are:

(i) no inspection \((n = 0)\):
\[
\pi(0, M, N) = \frac{M}{N}, \quad \text{for } M \in \{0, \ldots, N\};
\]

(ii) no errors \((M = 0)\):
\[
\pi(n, 0, N) = 0, \quad \text{for } n \in \{0, \ldots, N\};
\]

(iii) complete inspection \((n = N)\):
\[
\pi(N, M, N) = 0, \quad \text{for } M \in \{0, \ldots, N\}.
\]

In less trivial cases, we can rewrite (5.1.5) in the following way.

**Theorem 5.1.1.** For \( M \in \{1, \ldots, N\} \) and \( n \in \{0, \ldots, N - 1\} \) the following alternative expression holds:
\[
\pi(n, M, N) = \frac{M}{N} \left(1 - \frac{n}{N}\right) \cdot \Lambda(n, M - 1, N - 1).
\]

**Proof.** See Appendix 5.A
We have already mentioned that the statistical derivation of Dodge and Romig is disputable. They use
\[
\pi(n, M, N) = \mathbb{E}\left(\frac{M - K}{N}\right) \cdot \Lambda(n, M, N)
\]
instead of
\[
\pi(n, M, N) = \mathbb{E}\left(\frac{M - K}{N} \mid K \leq k_0\right) \cdot \Lambda(n, M, N)
\]
So, they use the unconditional expected value instead of the conditional on \( e \) in the definition of \( \pi \). Notice that the two are almost alike, except that we lower the arguments \( M \) and \( N \) of \( \Lambda(n, M, N) \) by one.

If the original population does not contain many errors, then the sample will not contain many errors either. Only few errors will be corrected and \( \pi \) will increase as a function of \( M \). If \( M \) increases further, then more errors will occur in the sample and therefore more errors will be corrected. The expected fraction of errors will still increase as a function of \( M \), but not as much as for really small \( M \).

An even further increase of \( M \) will increase the probability of exceeding \( k_0 \), and thus integral inspection of the population becomes more probable. A decrease of \( \pi \) as a function of \( M \) will be the result and finally \( \pi \) will become zero \( (M = N) \). Therefore, it can be expected that \( \pi \) is a unimodal function of \( M \). We will call this the unimodality property of \( \pi \). Of course, one would also expect that an increase in the sample size would decrease the expected fraction of errors after inspection. We will call this the monotonicity property of \( \pi \). This monotonicity property also ensures that \( \max_M \pi(n, M, N) \) decreases for increasing \( n \). Using the Poisson approximation for the hypergeometric distribution it is easy to prove the unimodality and monotonicity properties described above. This chapter will establish these properties also when no approximation is used.

Using the AOQL-method, we have to find the smallest sample size for which the expected fraction of errors after inspection does not exceed a predefined value, \( P_l \). Since we assume that we do not have any a priori knowledge of the amount of errors in the population before inspection, this has to hold for all
possible values of $M$. Checking for the least favourable case, $\max_M \pi(n, M, N)$ will suffice, because $\pi$ is a unimodal function of $M$. Since $\max_M \pi(n, M, N)$ is a decreasing function of $n$, we have to find the smallest sample size for which

$$\max_M \pi(n, M, N) \leq P_l.$$ 

This smallest sample size is the optimal size.

![Figure 5.1](image)

**Figure 5.1.** Plots of $\max_M \pi(n, M, N)$ and $\pi(n, M, N)$ for values of $n = 40$, $n = 77$, and $n = 120$ with $N = 500$, $k_0 = 1$, and $P_l = 0.01$. The optimal $n$ equals 77.

An example of finding the optimal sample size can be found in Figure 5.1. For improved clarity the graphs in this paper are drawn continuously, but one should keep in mind the discrete nature of the parameters. To find the least favourable $M$ for a certain $n$, we could start at the largest/smallest possible value of $M$ and decrease/increase $M$ by one until we find the maximum of $\pi$ and use a bisection method to find the optimal sample size. However, this would be a rather
inefficient and time-consuming way to find the optimal sample size. We will find a very efficient and appealing method that fully exploits the underlying hypergeometric distribution. The theory needed for this method will be described in the subsequent sections.

Although the AOQL sample system guarantees that the expected number of bad items after inspection does not exceed a limit chosen beforehand and minimizes the amount of work due to inspection, there still is a positive probability that the fraction of errors after inspection exceeds this limit. This probability can be sizable if the fraction of errors before inspection exceeds the limit chosen beforehand mildly (Kleijnen, Kriens, Lafluer and Pardoel, 1992).

It is sensible to divide the population in subpopulations and to apply the AOQL-method to each of the subpopulations. This offers protection against too much work if more than $k_0$ errors are found in the sample, because then only the subpopulation has to be fully inspected, instead of the entire population. Another advantage of splitting the population into subpopulations is that errors are discovered sooner. This means that measures can be taken to avoid these errors and to reduce the amount of work in the future. Kleijnen et al. (1992) showed by a simulation study that splitting up the population into subpopulations also reduces the probability that the fraction of errors after inspection exceeds the limit chosen beforehand. How the population has to be divided into subpopulations and how the size of the subpopulations has to be chosen depends on statistical as well as non-statistical reasons (Kleijnen et al., 1992; Kriens, 1988). If the expected fraction of errors in each subpopulation is smaller than $P_l$, then this obviously also holds for the whole population. Therefore, this paper only focuses on how we can determine the sample size $n$ in one population. The condition that in all subpopulations $\pi$ does not exceed $P_l$ is a sufficient but not a necessary condition to ensure that $\pi$ does not exceed $P_l$ in the population. Other approaches are possible, for instance Klaassen (2001) describes a credit-based acceptance sampling system with $k_0 = 0$. His system loosens inspection as the total number of accepted items since the last rejection increases.

The following sections\(^1\) study the behaviour of the expected fraction of errors after inspection, $\pi$, which, if $M \in \{1, \ldots, N\}$ and $n \in \{0, \ldots, N-1\}$, can be

---

\(^1\)All proofs of these sections can be found in Appendix 5.A
written as

\[ \pi(n, M, N) = \frac{M}{N} \left(1 - \frac{n}{N}\right) \cdot \Lambda(n, M - 1, N - 1). \] (5.1.6)

Section 5.2 will prove the unimodality property of \( \pi \) with respect to \( M \), the number of errors in the population, and the monotonicity property with respect to \( n \), the sample size. We already mentioned these properties in this section. We will also establish some other properties of \( \pi \).

Section 5.2 considers the expected fraction of errors \( \pi \) as a function of \( M \). We are especially interested in its maximum \( \pi^*(n, N) \). In order to obtain these maxima for various values of \( n \) and \( N \) we have to consider the values \( M^*(n, N) \) at which \( \pi(n, M, N) \) achieves this maximum. In Section 5.3 we derive various properties of \( M^*(n, N) \) and we focus on its behaviour as a function of \( n \) for fixed values of \( N \) (and \( k_0 \)).

**Table 5.1.** The triangular array of the values of \( M^*(n, N) \) for \( N, n \in \{0, \ldots, 10\} \) and \( k_0 = 1 \).

<table>
<thead>
<tr>
<th>( N \setminus n )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

One may imagine that in order to generate tables we have to produce a triangular array of the values of \( M^*(n, N) \) where \( N = k_0 + 1, k_0 + 2, \ldots, \) and \( n = k_0 + 1, \ldots, N - 1 \). Table 5.1 gives an example of such a triangular array. Section 5.3 develops theory that helps us to fill the row \( M^*(k_0 + 1, N) \), \( \ldots \), \( M^*(N - 1, N) \) as efficiently as possible. We use this theory also in the algorithm of finding the optimal sample size in the EOQL-method for given \( N, k_0 \), and \( P_1 \).
see Chapter 6. Section 5.4 develops theory that will help us to fill the column $M^*(n, k_0 + 1), M^*(n, k_0 + 2), \ldots$ of such a triangular array in an efficient way. Filling the triangular array of $M^*(n, N)$ can be done most efficiently by filling alternately columns and rows in a clever way, as we will see in Chapter 6.

### 5.2 Properties of $\pi(n, \cdot, N)$

We first present some properties that are very easy to prove and follow immediately from Theorem 4.1.1 and the expression for $\pi$ in (5.1.6). Especially, we will look at the behaviour of $\pi$ in a couple of special cases.

**Property 5.2.1.** The function $\pi$ has the following properties.

1. $\pi(n, M, N) = 0$ if and only if $M > N - n + k_0$ or $M = 0$ or $n = N$.
2. $\pi(n, M, N) = \frac{M}{N} \left(1 - \frac{n}{N}\right)$ if and only if $n \leq k_0$ or $M \leq k_0 + 1$ or $n = N$.
3. $\pi(n, N, N) = \begin{cases} 1 - \frac{n}{N} & \text{if } n \leq k_0, \\ 0 & \text{if } n > k_0. \end{cases}$
4. $\pi(N - 1, M, N) = \begin{cases} \frac{M}{N^2} & \text{if } M \leq k_0 + 1, \\ 0 & \text{if } M > k_0 + 1. \end{cases}$
5. If $n < N$, then $\pi(n, M, N) > 0$ for $M \in \{1, \ldots, N - n + k_0\}$, or equivalently, $n \in \{1, \ldots, N - M + k_0\}$.

As we discussed in Section 5.1, we are interested in finding the smallest value of $n$ for which the maximum of $\pi$ over $M$ stays under a certain predefined value. Therefore, we are interested in the behaviour of $\pi^*(n, N) = \max_M \pi(n, M, N)$.

The following theorem will show that $\pi(n, M, N)$ is unimodal. The smallest value of $M$ for which $\pi$ achieves its maximum will be denoted by $M^*(n, N)$. An exception is the case $n = N$, then we have $\pi \equiv 0$ and hence we have $M^*(N, N) = 0$. If there is another solution then it is $M^*(n, N) + 1$, see e.g. example 5.2.1 where $\pi(n, M, N)$ achieves its maximum in two successive values of $M$.

In fact we would also like to construct a triangular array with entries for $n$ and $N$ that gives us $\pi^*(n, N)$ for fixed $k_0$. Basically we have to find $M^*(n, N)$,
because \( \pi^*(n, N) = \pi(n, M^*(n, N), N) \). In completing such an array we have to solve a lot of optimization problems. Hence, it pays to find efficient ways to obtain the \( M^*(n, N) \).

**Example 5.2.1.** For \( k_0 = 2 \) the function \( \pi(13, M, 46) \) achieves its maximum for \( M = 7 \) and \( M^*(13, 46) = 7 \).

**Theorem 5.2.1 (Unimodality property).** Let \( n \in \{0, \ldots, N - 1\} \), then the function \( \pi(n, M, N) \) is unimodal in \( M \). Let \( M^*(n, N) \) be the smallest solution to \( \pi^*(n, N) = \max_M \pi(n, M, N) \), then \( M^*(n, N) \) is the unique solution or \( M^*(n, N) + 1 \) the only possible other solution. If \( n < N \), then \( 1 \leq M^*(n, N) \leq N - n + k_0 \).

**Proof.** See Appendix 5.A

Now, we look at the monotonicity property of \( \pi \). The next theorem will show that the function \( \pi \) is non-increasing in \( n \), and in most cases even decreasing.

**Theorem 5.2.2 (Monotonicity property).** Let \( n, M \in \{0, \ldots, N - 1\} \), then

\[
\pi(n, M, N) \geq \pi(n+1, M, N).
\]

If \( n \leq N - M + k_0 \) and \( M \neq 0 \), then

\[
\pi(n, M, N) > \pi(n+1, M, N).
\]

**Proof.** See Appendix 5.A

According to Theorem 5.2.2, the graph of \( \pi(n, \cdot, N) \) lies strictly above the graph of \( \pi(n+1, \cdot, N) \) for \( M \in \{1, \ldots, N - n + k_0\} \), see Figure 5.2. For other values of \( M \), we have \( \pi(n+1, M, N) = \pi(n, M, N) = 0 \). Hence, we have the following result.

**Theorem 5.2.3.** Let \( n \in \{0, \ldots, N-1\} \), then the function \( \pi^*(n, N) \) is decreasing in \( n \), i.e.

\[
\pi^*(n, N) > \pi^*(n+1, N).
\]
5.3 Properties of $M^*(\cdot, N)$

We have proved the unimodality and monotonicity properties of $\pi$ and could use the method for finding the optimal sample size as described in Section 5.1. This method is rather naive in finding $M^*$. To find $M^*$, we either start at $M = 1$ and increase $M$ by one until $\pi(n, M+1, N) < \pi(n, M, N)$ or start at $M = N-n+k_0$ and decrease $M$ by one until $\pi(n, M-1, N) < \pi(n, M, N)$. Of course, we can find $M^*(n, N)$ in a more efficient way. The results also help to complete the rows and columns of the triangular array of $M^*(n, N)$. The remaining part of this section will focus on this issue. Can we tighten the interval for $M$ where we have to search for $M^*(n, N)$? Do we need to calculate $M^*(n, N)$ for every $n$, keeping $N$ fixed? How does $M^*(n, N)$ change if we increase $n$ or $N$? But first we consider some special cases where we have explicit solutions. As the next theorem will show, we can find $M^*(n, N)$ analytically in the case of $k_0 = 0$.

![Figure 5.2. Plot of $\pi(n, M, N)$ for $n = 0, \ldots, 9$, $N = 10$ and $k_0 = 2$.](image)
5.3. Properties of $M^*(\cdot, N)$

**Theorem 5.3.1.** If $k_0 = 0$, then

$$M^*(n, N) = \left\lceil \frac{N - n}{n + 1} \right\rceil.$$  

([x] is the smallest integer $\geq x$).

*Proof.* See Appendix 5.A

For other values of $k_0$, we cannot find the value of $M^*(n, N)$ in such an easy way and we have to look for other ways. However, there are some special cases collected in Property 5.3.1 below.

**Property 5.3.1.** Some special cases for $\pi^+$ are:

1. $\pi^*(n, N) = 0$ if and only if $n = N$, and by definition $M^*(N, N) = 0$.

2. If $n \leq k_0$, then $M^*(n, N) = N$ and $\pi^*(n, N) = 1 - \frac{n}{N}$.

3. If $N \geq k_0 + 1$, then $M^*(N - 1, N) = k_0 + 1$ and $\pi^*(N - 1, N) = \frac{k_0 + 1}{N^2}$.

We have already seen that the graph of $\pi(n, \cdot, N)$ lies above the one of $\pi(n + 1, \cdot, N)$, but the graph of relative changes of $\pi(n, \cdot, N)$ also lies above the graph of relative changes of $\pi(n + 1, \cdot, N)$ as $M$ changes to $M + 1$, see Figure 5.3. We prove this property in the following theorem.

**Theorem 5.3.2.** Let $n \in \{0, \ldots, N - 2\}$. Then for $M = 1, \ldots, N - n + k_0 - 1$

$$\frac{\pi(n, M + 1, N)}{\pi(n, M, N)} \geq \frac{\pi(n + 1, M + 1, N)}{\pi(n + 1, M, N)}.$$  

*Proof.* See Appendix 5.A

Note that in Figure 5.3 the cases $n = 0, 1, 2$ lead to the same graph, because then $n \leq k_0 = 2$ and $\lambda(n, M - 1, N - 1) = 1$ according to Theorem 4.1.2, part 1. From Theorem 5.3.2 we may expect that $\frac{\pi(n, M + 1, N)}{\pi(n, M, N)} - 1$ will not change sign before its counterpart for $n + 1$ changes sign. This can for example be observed in Figure 5.3. Hence, $\pi(n, M, N)$ will not achieve its maximum before $\pi(n + 1, M, N)$ will achieve its maximum. We now prove that indeed $M^*(n, N)$ is non-increasing in $n$. 
Theorem 5.3.3. Let \( n \in \{0, \ldots, N-1\} \), then

\[
M^*(n, N) \geq M^*(n+1, N).
\]

Proof. See Appendix 5.A

Note that Theorem 5.3.3 already helps us to tighten the interval for possible candidates \( M^*(n+1, N) \) once we have \( M^*(n, N) \). As a consequence we have the following lower bound for \( M^*(n, N) \).

Corollary 5.3.1. Let \( n \in \{0, \ldots, N-1\} \). If \( N \geq k_0+1 \), then \( M^*(n, N) \geq k_0+1 \).

Proof. See Appendix 5.A

The intuitive algorithm for finding \( M^* \) can be improved upon if we can reduce the interval of feasible values of \( M \). The following theorem gives a lower and upper bound for \( M^* \).
Theorem 5.3.4. Let \( n, k_0 \) and \( N \) be given, such that \( N \geq k_0 + 1 \) and \( n \in \{k_0 + 1, \ldots, N - 1\} \), then

\[
\max \left\{ k_0 + 1, \left\lceil \frac{N - n}{n + 1} \right\rceil \right\} \leq M^*(n, N) \leq \min \left\{ N - n + k_0, \left\lceil \frac{N + 1}{n + 1} \cdot k_0 + \frac{N - n}{n + 1} \right\rceil \right\}.
\]

Proof. See Appendix 5.A

Remark 5.3.1. Notice that if we insert \( k_0 = 0 \) into Theorem 5.3.4, then Theorem 5.3.1 follows as a special case.

Now we have found lower and upper bounds for \( M^* \). This makes the algorithm of finding the optimal sample size for given \( k_0 \) and \( N \) more efficient. But it appears that we can find some other properties of \( M^* \), which will improve the algorithm even further and these properties will also help us to fill the rows of the triangular array of \( M^*(n, N) \). We will prove that if we know the values of \( M^*(n, N) \) for which \( k_0 + 1 \leq n \leq M^*(n, N) \), then we shall also get to know the values of \( M^*(n, N) \) for which \( n > M^*(n, N) \) in an easy way. In proving this the behaviour of \( M^*(M^*(n, N), N) \) plays a key role. We use an example with \( N = 17 \) and \( k_0 = 2 \) to illustrate the theorems that we deduce.

In this example we suppose that the values of \( M^*(n, N) \) for \( n \leq M^*(n, N) \) are known and we illustrate the effect of the theorems on the values of \( M^*(n, N) \) for which \( n > M^*(n, N) \). We already know that \( M^*(n, N) = N \) for \( n \leq k_0 \), \( M^*(N - 1, N) = k_0 + 1 \) and \( M^*(N, N) = 0 \).

Example 5.3.1. Let \( N = 17 \) and \( k_0 = 2 \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( M^*(n, 17) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
</tr>
</tbody>
</table>

A first observation is that \( 3 \leq M^*(n, 17) \leq 6 \) for \( n = 7, \ldots, 15 \). This follows from Theorem 5.3.3. More strongly, we will prove that if we would know the value of \( M^*(n, N) \), then we would immediately know that the value of \( M^*(M^*(n, N), N) \) does not exceed \( n \).

Example 5.3.2. Let \( N = 17 \) and \( k_0 = 2 \). If we apply the statement above to Example 5.3.1, then we get
We have \( M^*(7, 17) = M^*(M^*(5, 17), 17) \leq 5 \), and so forth. Notice that we actually get \( M^*(11, 17) = M^*(M^*(3, 17), 17) \leq 3 \). But since we know that \( M^*(n, N) \geq k_0 + 1 \), we also know by Theorem 5.3.3 that \( M^*(n, 17) \) equals 3 for \( n \geq 11 \).

**Theorem 5.3.5.** Let \( n \in \{0, \ldots, N\} \), then

\[
M^*(M^*(n, N), N) \leq n.
\]

**Proof.** See Appendix 5.A □

For \( n \geq k_0 \) we will show that if \( M^*(n, N) \) is at least equal to \( n \), then \( M^*(M^*(n, N), N) \) is always larger than or equal to \( n \).

**Example 5.3.3.** Let \( N = 17 \) and \( k_0 = 2 \). If we apply the statement above to Example 5.3.1, then we get

<table>
<thead>
<tr>
<th>( n )</th>
<th>( M^*(n, 17) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 2 )</td>
<td>17 11 8 7 6 ( \leq 5 )</td>
</tr>
<tr>
<td>3 3 3 3 3 3 0</td>
<td></td>
</tr>
</tbody>
</table>

We have \( M^*(7, 17) = M^*(M^*(5, 17), 17) \geq 5 \), and so forth.

**Theorem 5.3.6.** Let \( k_0 + 1 \leq n \leq M^*(n, N) \), then

\[
M^*(M^*(n, N), N) \geq n.
\]

**Proof.** See Appendix 5.A □

**Example 5.3.4.** Let \( N = 17 \) and \( k_0 = 2 \). If we combine Examples 5.3.2 and 5.3.3, then we get

<table>
<thead>
<tr>
<th>( n )</th>
<th>( M^*(n, 17) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq 2 )</td>
<td>17 11 8 7 6 5 4</td>
</tr>
<tr>
<td>3 3 3 3 3 3 0</td>
<td></td>
</tr>
</tbody>
</table>

Looking at Example 5.3.4, we see that \( M^*(M^*(n, N), N) = n \) for \( k_0 + 1 \leq n \leq M^*(n, N) \). Of course this is an immediate consequence of combining Theorem 5.3.5 and Theorem 5.3.6.
5.3. Properties of $M^*(\cdot, N)$

**Corollary 5.3.2.** Let $k_0 + 1 \leq n \leq M^*(n, N)$, then

$$M^*(M^*(n, N), N) = n.$$ 

We shall prove that if we know the value of $M^*(n, N)$, then we also know that the value of $M^*(M^*(n, N) - 1, N)$ is always larger than $n$. If $k_0 + 1 \leq n \leq M^*(n, N)$, then from Theorem 5.3.3 and Theorem 5.3.6 it follows that $M^*(M^*(n, N) - 1, N) \geq n$. In Theorem 5.3.7 we will have a stronger result: the condition on $n$ can be deleted and the inequality is strict.

**Example 5.3.5.** Let $N = 17$ and $k_0 = 2$. If we apply the statement above to Example 5.3.4, then we get

<table>
<thead>
<tr>
<th>$n$</th>
<th>$M^*(n, 17)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
</tr>
</tbody>
</table>

Theorem 5.3.7 tells us that $M^*(M^*(3, 17) - 1, 17) = M^*(10, 17) \geq 4$. Using Theorem 5.3.3 and $M^*(8, 17) = 4$, we know that $M^*(n, 17) = 4$ for $n = 8, 9, 10$. Thus, apparently to fill the entire row it suffices to calculate the values of $M^*(n, 17)$ for $n = 3, \ldots, 6$ and with the help of the theorems above the rest of the values of $M^*(n, 17)$ will follow from these initial values.

**Theorem 5.3.7.** Let $n \in \{0, \ldots, N - 1\}$, then

$$M^*(M^*(n, N) - 1, N) \geq n + 1.$$ 

**Proof.** See Appendix 5.A

Example 5.3.5 suggests that if we have calculated the values of $M^*(n, N)$ with $n \leq M^*(n, N)$, then we will know the value of $M^*(n, N)$ for all possible sample sizes. The following theorem gives the essential ingredients for proving this and will help us in Chapter 6 to develop an algorithm that fills the row $M^*(k_0 + 1, N), \ldots, M^*(N - 1, N)$ as efficiently as possible. Notice that from Property 5.3.1 follows that for values of $N \leq k_0 + 2$ we already can complete the rows entirely. For values of $N \geq k_0 + 3$, we can use Theorem 5.3.8.

**Theorem 5.3.8.** Let $N \geq k_0 + 3$ and let $n_1$ be the largest value of $n$ for which $n \leq M^*(n, N)$, then $M^*(\cdot, N)$ has the following properties.
1. If \( n = k_0 + 1, \ldots, n_1 \), then \( M^*(n - 1, N) > M^*(n, N) \).

2. If \( M^*(n - 1, N) > M^*(n, N) \) for some \( n \in \{k_0 + 1, \ldots, N - 1\} \), then \( M^*(j, N) = n \) for \( j = M^*(n, N), M^*(n, N) + 1, \ldots, M^*(n - 1, N) - 1 \).

3. The largest value of \( n \) for which \( n \leq M^*(n, N) \) equals \( n_1 \), if and only if \( M^*(n_1, N) \) is equal to \( n_1 \) or \( n_1 + 1 \).

Proof. See Appendix 5.A

Table 5.2 can be used to provide extra insight in the properties of Theorem 5.3.8.

**Table 5.2.** Some values of \( M^*(n, N) \) for \( N \in \{15, \ldots, 30\} \), \( n \in \{0, \ldots, 12\} \) and \( k_0 = 1 \).

<table>
<thead>
<tr>
<th>( N ) ( \backslash ) ( n )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>12</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>13</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>15</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>16</td>
<td>9</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>17</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>18</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>19</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>20</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>21</td>
<td>12</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>22</td>
<td>22</td>
<td>13</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>23</td>
<td>13</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>24</td>
<td>14</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>25</td>
<td>15</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Theorem 5.3.8 helps us to efficiently calculate \( M^*(n, N) \) for fixed \( N \). Only values \( N \geq k_0 + 1 \) and \( n = k_0, \ldots, N - 1 \) are sensible. Therefore, we start with \( M^*(k_0, N) = N \), then we calculate \( M^*(k_0 + 1, N) \), and so forth. Hence \( M^*(n, N) \) is calculated for increasing values of \( n \). We stop if we obtained
5.4. Properties of $M^*(n, \cdot)$

$M^*(n, N) = n$ or $n + 1$. This value is called $n_l$ in Theorem 5.3.8. The theorem tells us for $N \geq k_0 + 3$ that $M^*(n, N)$ is strictly decreasing on $\{k_0, \ldots, n_l\}$. The values of $M^*(n, N)$ on $\{n_l + 1, \ldots, N - 1\}$ are obtained as follows. The sequence

$$M^*(n_l, N), M^*(n_l, N) + 1, \ldots, M^*(k_0, N) - 1$$

is strictly increasing with a difference of unit one between the elements and with $M^*(n_l, N) = n_l$ or $n_l + 1$ and $M^*(k_0, N) = N$. So, we have

$$\{n_l + 1, \ldots, N - 1\} \subset \{M^*(n_l, N), M^*(n_l, N) + 1, \ldots, M^*(k_0, N) - 1\}$$

For all elements in the subset $\{M^*(k, N), M^*(k, N) + 1, \ldots, M^*(k - 1, N) - 1\}$ we have $M^*(\cdot, N) = k$. Theorem 5.3.8, part 1 tells us that for $k = k_0 + 1, \ldots, n_l$ the ‘integer interval’ $\{M^*(k, N), M^*(k, N) + 1, \ldots, M^*(k - 1, N) - 1\}$ is well-defined and consists of at least one element. Moreover, these intervals are disjoint. Hence at the right-hand side we have a disjoint union and it gives a partition of the ‘integer interval’ $\{n_l + 1, \ldots, N - 1\}$. From this observation we have the following result:

**Theorem 5.3.9.** If $N \geq k_0 + 3$ and $n = n_l + 1, \ldots, N - 1$, then

$$M^*(n - 1, N) - 1 \leq M^*(n, N) \leq M^*(n - 1, N).$$

**Proof.** See Appendix 5.A

**Remark 5.3.2.** Notice that if $N \geq k_0 + 3$, then we know the values of $M^*(n, N)$ for $n = n_l + 1, \ldots, N - 1$ as soon as we have calculated the values of $M^*(n, N)$ for $n = k_0 + 1, \ldots, n_l$. The converse is also true.

5.4 Properties of $M^*(n, \cdot)$

In order to fill the columns of the triangular array of $M^*(n, N)$ in an efficient way, we have to know the behaviour of $M^*(n, N)$ as $N$ increases. We will establish that an increment of $N$ with one unit leads to the same $M^*$ or to $M^* + 1$. Therefore, it is important to study the relative change of the curve $\pi(n, \cdot, N)$ as $N$ changes to $N + 1$. 


Theorem 5.4.1. Assume that \( M \in \{1, \ldots, N - n + k_0 - 1\} \). Then

\[
\frac{\pi(n, M + 1, N)}{\pi(n, M, N)} > \frac{\pi(n, M + 2, N + 1)}{\pi(n, M + 1, N + 1)}.
\] (5.4.1)

and if \( M \geq k_0 + 1 \) and \( n > k_0 \), then

\[
\frac{\pi(n, M + 1, N)}{\pi(n, M, N)} < \frac{\pi(n, M + 1, N + 1)}{\pi(n, M, N + 1)}.
\] (5.4.2)

Now, we are ready to prove that the optimal value \( M^* \) remains unchanged or increases by one unit if \( N \) is increased by one unit and \( n \) remains unchanged.

Theorem 5.4.2. Let \( n \in \{0, \ldots, N - 1\} \), then

\[
M^*(n, N) \leq M^*(n, N + 1) \leq M^*(n, N) + 1.
\]

Proof. See Appendix 5.A

The elements in the columns of the triangular array \( M^*(n, N) \) can be calculated in an efficient way by using Theorem 5.4.2. According to this theorem,

\[
\pi^*(n, M, N) = \max(\pi(n, M^*(n, N - 1), N), \pi(n, M^*(n, N - 1) + 1, N)).
\]

If we know \( M^*(n, N - 1) \), then \( M^*(n, N - 1) \) and \( M^*(n, N - 1) + 1 \) are the only two possible solutions for \( M^*(n, N) \). It is possible that \( M^*(n, N - 1) \) and \( M^*(n, N - 1) + 1 \) give the same optimal solution of \( \pi^*(n, N) \) (see Example 5.2.1), in this case by the definition of \( M^* \) we have \( M^*(n, N) = M^*(n, N - 1) \).

We can also show that if \( N \) is increased by one, then the largest value of \( n \) which is equal to or exceeded by \( M^* \) remains unchanged or increases by one unit.

Corollary 5.4.1. Let \( n_l(N) \) be the largest value of \( n \) for which \( n \leq M^*(n, N) \), then

\[
n_l(N) \leq n_l(N + 1) \leq n_l(N) + 1.
\]

Moreover, if \( M^*(n_l(N), N) \) equals \( n_l(N) \), then \( n_l(N + 1) \) equals \( n_l(N) \). If \( M^*(n_l(N), N) = n_l(N) + 1 \) and \( M^*(n_l(N), N + 1) = n_l(N) + 1 \), then \( n_l(N + 1) = n_l(N) \), else \( n_l(N + 1) = n_l(N) + 1 \).
The theory we developed in this chapter will enable us to produce a triangular array of $M^*(n, N)$ where $N = k_0 + 1, k_0 + 2, \ldots$ and $n = k_0 + 1, \ldots, N - 1$ in a very efficient way. The theory helps us to fill in the row $M^*(k_0 + 1, N), \ldots, M^*(N - 1, N)$ and will be used in Chapter 6 to construct an algorithm of finding the optimal sample size in the EOQL-method for given $N, k_0$, and $p_l$. We also developed theory that enables us to fill the column $M^*(n, k_0 + 1), M^*(n, k_0 + 2), \ldots$ of such a triangular array. Chapter 6 introduces an very appealing and efficient algorithm to generate triangular arrays in which we will combine the row- and column filling theory we developed in this chapter.

5.A Proofs of chapter 5

Proof of Theorem 5.1.1

From (5.1.5) we obtain

$$
\pi(n, M, N) = \sum_{k=0}^{\min(k_0, M-1)} \frac{M-k}{N} \cdot \binom{M}{k} \binom{N-M}{n-k}
$$

$$
= \sum_{k=0}^{\min(k_0, M-1)} \frac{M(M-1)!}{k!(M-1-k)!} \cdot \frac{n!(N-n)(N-1-n)!}{N^2(N-1)!} \times \left( \binom{N-M}{n-k} \right)
$$

$$
= \frac{M}{N} \cdot \frac{N-n}{N} \cdot \sum_{k=0}^{\min(k_0, M-1)} \binom{M-1}{k} \binom{N-M}{n-k}
$$

$$
= \frac{M}{N} \left( 1 - \frac{n}{N} \right) \cdot A(n, M - 1, N - 1).
$$

Proof of Theorem 5.2.1 (Unimodality property)

According to Property 5.2.1, part 5, for all $M \in \{1, \ldots, N - n + k_0\}$ the function $\pi(n, M, N)$ is positive. If $N - n + k_0 = 1$ or 2, then the result is trivial. Now suppose that $N - n + k_0$ equals at least 3. The function $\pi(n, M, N)$ is a unimodal
function if we can show that for \( M \in \{1, \ldots, N - n + k_0\} \),

\[
\frac{\pi(n, M + 1, N)}{\pi(n, M, N)} = \frac{M + 1}{M} \cdot \frac{\lambda(n, M - 1, N - 1)}{\lambda(n, M - 1, N - 1)}
\]

is decreasing in \( M \). Because we cannot exclude the possibility that this ratio can be equal to 1 for one specific value of \( M \), it is possible that we have two optimal solutions of \( \pi \) for succeeding values of \( M \). Combining Theorem 4.1.2, part 4 with the fact that \( 1 + \frac{1}{M} \) is decreasing in \( M \) concludes the proof.

**Proof of Theorem 5.2.2 (Monotonicity property)**

The case \( M = 0 \) is trivial. Therefore, we may assume that \( M \neq 0 \). We have to prove that

\[
\pi(n, M, N) = \frac{M}{N} \left(1 - \frac{n}{N}\right) \Lambda(n, M - 1, N - 1)
\]

\[
\geq \frac{M}{N} \left(1 - \frac{n + 1}{N}\right) \Lambda(n + 1, M - 1, N - 1) = \pi(n + 1, M, N).
\]

This holds on account of Theorem 4.1.1, part 6. If \( n \leq N - M + k_0 \), then Theorem 4.1.1, part 6 provides the strict inequality \( \Lambda(n, M - 1, N - 1) > \Lambda(n + 1, M - 1, N - 1) \). From this the strict inequality of the theorem immediately follows.

**Proof of Theorem 5.3.1**

We only have to look at \( \pi(n, M, N) - \pi(n, M - 1, N) \) for \( M = 1, \ldots, N - n \). If this difference changes sign from positive to negative only once, then the maximum is achieved for the largest value of \( M \) such that this difference is positive. By Theorem 5.1.1 it is straightforward to see that the sign of \( \pi(n, M, N) - \pi(n, M - 1, N) \) is the same as the sign of

\[
M \cdot \frac{(N - M)!}{(N - M - n)!} - \frac{(N - M - 1)!}{(N - M - n + 1)!} \cdot \left(\frac{N - M + 1}{N - M - n + 1}\right)
\]

\[
= \frac{(N - M)!}{(N - M - n + 1)!} \cdot \left[M(N - M - n + 1) - (M - 1)(N - M + 1)\right]
\]

\[
= \frac{(N - M)!}{(N - M - n + 1)!} \cdot [N + 1 - M(n + 1)]
\]
Indeed this difference changes sign only once and for non-integer values of \(\frac{N+1}{n+1}\) the maximum will be achieved for the largest integer smaller than or equal to \(\frac{N+1}{n+1}\) or, equivalently, the smallest integer larger than \(\frac{N+1}{n+1} - 1\), i.e. \(M^*(n, N) = \lceil \frac{N-n}{n+1} \rceil\).

For integer values of \(\frac{N+1}{n+1}\), we have two successive values of \(M\) for which the maximum is achieved. The smaller of the two values is the optimal solution, i.e.

\[
M^*(n, N) = \frac{N+1}{n+1} - 1 = \frac{N-n}{n+1}.
\]

**Proof of Theorem 5.3.2**

According to Theorem 5.1.1 and (4.1.2) it suffices to prove

\[
\lambda(n, M-1, N-1) \geq \lambda(n+1, M-1, N-1).
\]

This is true on account of Theorem 4.1.2, part 5.

**Proof of Theorem 5.3.3**

*Proof.* For \(n \leq k_0\) the proof immediately follows from Property 5.3.1, part 2. For \(n = N - 1\) the proof follows from Property 5.3.1, part 1. Let \(n < N - 1\). If \(M^*(n, N) = N - n + k_0\), then the result follows because \(M^*(n+1, N) \in \{1, \ldots, N - n + k_0 - 1\}\) according to Theorem 5.2.1. Now let us assume that \(n \in \{0, \ldots, N-2\}\) and \(M^*(n, N) \in \{1, \ldots, N - n + k_0 - 1\}\) and we apply Theorem 5.3.2. The unimodality property and this theorem show us that

\[
1 \geq \frac{\pi(n, M^*(n, N) + 1, N)}{\pi(n, M^*(n, N), N)} \geq \frac{\pi(n+1, M^*(n, N) + 1, N)}{\pi(n+1, M^*(n, N), N)}.
\]

Using the unimodality property again we find \(M^*(n, N) \geq M^*(n+1, N)\). □

**Proof of Corollary 5.3.1**

According to Property 5.3.1, part 3, we have

\[
M^*(n, N) \geq M^*(N - 1, N) = k_0 + 1.
\]
Proof of Theorem 5.3.4

If \( M \leq k_0 + 1 \), then we know from Property 5.2.1, part 2 that \( \pi(n, M, N) \) increases. For \( M > N - n + k_0 \), we also know that \( \pi(n, M, N) = 0 \). Therefore, we know that \( k_0 + 1 \leq M^*(n, N) \leq N - n + k_0 \). The other bounds can be established as follows. For \( M \in \{k_0 + 2, \ldots, N - n + k_0 - 1\} \) and \( k \in \{0, \ldots, k_0\} \) the \( k \)-th term of \( \pi(n, M, N) \) is called \( \pi_k(n, M, N) \). Look at the term-wise difference \( D_k(n, M, N) \) between \( \pi_k(n, M, N) \) and \( \pi_k(n, M - 1, N) \). Using Theorem 5.1.1, we find

\[
D_k(n, M, N) = \pi_k(n, M, N) - \pi_k(n, M - 1, N)
\]

\[
= \frac{M}{N} \left( 1 - \frac{n}{N} \right) \left( \frac{M - 1}{n - k} \right) \left( \frac{N - M}{n - k} \right) + \frac{M - 1}{N} \left( 1 - \frac{n}{N} \right) \left( \frac{M - 2}{n - k} \right) \left( \frac{N - M + 1}{n - k} \right)
\]

\[
= C_k(n, N, M) \left( \frac{M}{M - k - 1} - \frac{N - M + 1}{N - M - n + k + 1} \right)
\]

with

\[
C_k(n, N, M) = \frac{(N - n)(M - 1)! (N - M)! n! (N - n - 1)!}{N \cdot k! (n - k)! (M - k - 2)! N! (N - M - n + k)!}.
\]

Therefore, \( D_k(n, M, N) \) is positive for integers \( M \) for which

\[
\frac{M}{M - k - 1} - \frac{N - M + 1}{N - M - n + k + 1} > 0.
\]

This holds if \( M < \frac{(N + 1)(k + 1)}{n + 1} \) or, equivalently, if \( M \in \{0, \ldots, M^*_k(n, N)\} \), where

\[
M^*_k(n, N) = \left\lfloor \frac{(N + 1)(k + 1)}{n + 1} - 1 \right\rfloor = \left\lfloor \frac{N + 1}{n + 1} \cdot k + \frac{N - n}{n + 1} \right\rfloor.
\]

Note that \( M^*_k(n, N) \) is increasing in \( k \). Since \( \pi \) is a unimodal function of \( M \), we have \( M^*(n, N) \geq M^*_0(n, N) \), because for \( M \leq M^*_0(n, N) \) all terms \( D_k(n, M, N) \) are positive and then \( \pi \) increases. From this it follows that \( M^*_0(n, N) \) is a lower bound for \( M^*(n, N) \). For \( M > M^*_0(n, N) \) all terms \( D_k(n, M, N) \) are non-positive and then \( \pi \) is non-increasing. Therefore, \( M^*_k_0(n, N) \) is an upper bound for \( M^*(n, N) \).
Proof of Theorem 5.3.5

In case $n = 0$ we have

$$M^*(M^*(0, N), N) = M^*(N, N) = 0,$$

and in case $n = N$ we have

$$M^*(M^*(N, N), N) = M^*(0, N) = N.$$

Now, we assume $n \in \{1, \ldots, N - 1\}$. In order to guarantee later on in this proof that $\pi(M^*(n, N), n + 1, N) > 0$, we have to check the case $M^*(n, N) = N - n + k_0$ separately. Suppose $M^*(n, N) = N - n + k_0$, then

$$M^*(M^*(n, N), N) = M^*(N - n + k_0, N) \leq N - (N - n + k_0) + k_0 = n.$$  

Hence, we are now allowed to assume $M^*(n, N) < N - n + k_0$ and hence we have $\pi(M^*(n, N), n + 1, N) > 0$. Due to the unimodality property of $\pi$ it suffices to prove

$$1 \leq \frac{\pi(M^*(n, N), n, N)}{\pi(M^*(n, N), n + 1, N)} = \frac{n}{n+1} \cdot \frac{A(M^*(n, N), n - 1, N - 1)}{A(M^*(n, N), n, N - 1)} .$$

By using Theorem 4.1.1, part 4 we can write this as

$$\frac{n}{n+1} \cdot \left( 1 + \frac{M^*(n, N)}{N - 1} \cdot \frac{P(K = k_0|M^*(n, N) - 1, n - 1, N - 2)}{A(M^*(n, N), n, N - 1)} \right) \geq 1 ,$$

or, equivalently,

$$\frac{P(K = k_0|M^*(n, N) - 1, n - 1, N - 2)}{A(M^*(n, N), n, N - 1)} \geq \frac{N - 1}{n \cdot M^*(n, N)} .$$  \hspace{1cm} (5.A.1)

We know that

$$1 \leq \frac{\pi(n, M^*(n, N), N)}{\pi(n, M^*(n, N) + 1, N)} = \frac{M^*(n, N)}{M^*(n, N) + 1} \cdot \frac{A(n, M^*(n, N) - 1, N - 1)}{A(n, M^*(n, N), N - 1)} .$$

By using Theorem 4.1.1, part 4 we can write this as

$$\frac{M^*(n, N)}{M^*(n, N) + 1} \cdot \left( 1 + \frac{n}{N - 1} \cdot \frac{P(K = k_0 | n - 1, M^*(n, N) - 1, N - 2)}{A(n, M^*(n, N), N - 1)} \right) \geq 1 ,$$

or, equivalently,

$$\frac{P(K = k_0 | n - 1, M^*(n, N) - 1, N - 2)}{A(n, M^*(n, N), N - 1)} \geq \frac{N - 1}{n \cdot M^*(n, N)} ,$$

which shows that (5.A.1) holds true. This completes the proof.
Proof of Theorem 5.3.6

If \( n = k_0 + 1 \), then we know from Theorem 5.3.5 that \( M^*(k_0 + 1, N) \leq k_0 + 1 \), but from Theorem 5.3.4 we know that \( M^*(k_0 + 1, N) \geq k_0 + 1 \), because \( M^*(k_0 + 1, N) \geq k_0 + 1 \). Hence \( M^*(k_0 + 1, N) = k_0 + 1 \). This establishes the theorem for \( n = k_0 + 1 \). Therefore we assume \( n > k_0 + 1 \). For notational purposes we will write \( M^* \) instead of \( M^*(n, N) \) in the remainder of this proof. If \( M^* = n \), then the result is obvious. Hence, we assume \( M^* > n \) in the sequel. By the unimodality property we know that

\[
\pi(n, M^* - 1, N) < \pi(n, M^*, N),
\]

which can be simplified to

\[
(M^* - 1) \cdot A(n, M^* - 2, N - 1) < M^* \cdot A(n, M^* - 1, N - 1).
\]

Using Theorem 4.1.1, part 4 we can rewrite this as

\[
(M^* - 1) \cdot \left( A(n, M^* - 1, N - 1) + \frac{(M^* - 2) \binom{N-M^*}{k_0} \binom{n-M^*}{n-k_0-1}}{\binom{N-1}{n}} \right) < M^* \cdot A(n, M^* - 1, N - 1),
\]

or, equivalently,

\[
A(n, M^* - 1, N - 1) > \frac{(M^* - 1) \binom{M^* - 2}{k_0} \binom{N-M^*}{n-k_0-1}}{\binom{N-1}{n}}.
\]

Applying Theorem 4.1.1, part 4 again, this gives

\[
A(n, M^*, N - 1) > \frac{(M^* - 1) \binom{M^* - 2}{k_0} \binom{N-M^*}{n-k_0-1} - (M^* - 1) \binom{M^* - 1}{k_0} \binom{N-M^* - 1}{n-k_0-1}}{\binom{N-1}{n}}. \tag{5.A.2}
\]

The unimodality property ensures that it is sufficient to prove that

\[
\pi(M^*, n - 1, N) < \pi(M^*, n, N).
\]

Using the same reasoning as above and by changing the role of \( M^* \) and \( n \) we can rewrite this to

\[
A(M^*, n, N - 1) > \frac{(n - 1) \binom{n-2}{k_0} \binom{N-n}{M^*-k_0-1} - (n-1) \binom{n-1}{k_0} \binom{N-n-1}{M^*-k_0-1}}{\binom{N-1}{M^*}}. \tag{5.A.3}
\]
By Theorem 4.1.1, part 1, (5.A.2) and (5.A.3) it is sufficient to prove that
\[
(M^* - 1) \cdot \frac{(M^*-2)(N-M^*)}{k_0(n-k_0-1)} - \frac{(M^*-1)(N-M^*-1)}{k_0(n-k_0-1)} \geq \frac{(n-2)(N-n)(M^*-k_0-1)}{k_0(M^*-k_0-1)}.
\]

By writing out and cancelling common factors we get
\[
n \cdot \frac{N-M^*}{(n-k_0-1)(N-n-M^*+k_0+1)} - \frac{1}{(M^*-k_0-1)(n-k_0-1)} \geq M^* \cdot \frac{N-n}{(M^*-k_0-1)(N-n-M^*+k_0+1)}.
\]

Since all terms in the denominator are positive this is equivalent to
\[
(M^* - n) \cdot \frac{N-n-M^*+k_0+1-M^*n+N(k_0+1)}{n+1} \geq 0.
\]

Since we assumed $M^* > n$, it remains to show that
\[
M^* \leq \frac{N+1}{n+1} \cdot k_0 + \frac{N-n}{n+1} + \frac{N+1}{n+1}.
\]

But Theorem 5.3.4 tells us that
\[
M^* \leq \left\lceil \frac{N+1}{n+1} \cdot k_0 + \frac{N-n}{n+1} \right\rceil,
\]
and since $\frac{N+1}{n+1} > 1$ this completes the proof.

**Proof of Theorem 5.3.7**

Suppose $M^*(n, N) = 1$. Since $M^*(0, N) = N$, we have that
\[
M^*(M^*(n, N) - 1, N) = N \geq n + 1.
\]

Let us now consider the case that $M^*(n, N) \geq 2$. We know that
\[
1 > \frac{\pi(n, M^*(n, N) - 1, N)}{\pi(n, M^*(n, N), N)} = \frac{M^*(n, N) - 1}{M^*(n, N)} \cdot \frac{A(n, M^*(n, N) - 2, N - 1)}{A(n, M^*(n, N) - 1, N - 1)}.
\]
Using Theorem 4.1.1, part 4 we can write this as
\[
\frac{M^*(n, N) - 1}{M^*(n, N)} \cdot \left( 1 + \frac{n}{N - 1} \cdot \frac{P[K = k_0|n - 1, M^*(n, N) - 2, N - 2]}{\Lambda(n, M^*(n, N) - 1, N - 1)} \right) < 1,
\]

or, equivalently,
\[
P[K = k_0|n - 1, M^*(n, N) - 2, N - 2] < \frac{N - 1}{n \cdot (M^*(n, N) - 1)}. \tag{5.A.4}
\]

The unimodality property ensures that we only have to prove that
\[
1 > \frac{\pi(M^*(n, N) - 1, n, N)}{\pi(M^*(n, N) - 1, n + 1, N)} = \frac{n + 1}{n} \cdot \frac{\Lambda(M^*(n, N) - 1, n - 1, N - 1)}{\Lambda(M^*(n, N) - 1, n, N - 1)}.
\]

holds. By using Theorem 4.1.1, part 4 we can write this as
\[
\frac{n}{n + 1} \cdot \left( 1 + \frac{M^*(n, N) - 1}{N - 1} \cdot \frac{P[K = k_0|M^*(n, N) - 2, n - 1, N - 2]}{\Lambda(M^*(n, N) - 1, n, N - 1)} \right) < 1,
\]

or, equivalently,
\[
P[K = k_0|M^*(n, N) - 2, n - 1, N - 2] < \frac{N - 1}{n \cdot (M^*(n, N) - 1)},
\]

which holds because of (5.A.4).

**Proof of Theorem 5.3.8**

Since \( n \geq k_0 + 1 \), we have by Corollary 5.3.2 and Theorem 5.3.5
\[
M^*(M^*(n, N), N) = n > n - 1 \geq M^*(M^*(n - 1, N), N),
\]

which implies \( M^*(n, N) \neq M^*(n - 1, N) \). Since \( M^*(n, N) \) is non-increasing in \( n \), it follows \( M^*(n - 1, N) > M^*(n, N) \), which concludes the proof of part 1.

Now we prove part 2. From the assumption it follows that \( M^*(n, N) \leq M^*(n - 1, N) - 1 \) for some \( n \in \{k_0 + 1, \ldots, N - 1\} \). From Theorem 5.3.5 we know that \( M^*(M^*(n, N), N) \leq n \). Moreover, we have \( M^*(M^*(n - 1, N) - 1, N) \geq n \) on account of Theorem 5.3.7. Because \( M^*(j, N) \) is non-increasing in \( j \), we have \( M^*(j, N) = n \) for \( j = M^*(n, N), \ldots, M^*(n - 1, N) - 1 \).

To prove part 3 first note that \( M^*(k_0, N) = N \) and \( M^*(N - 1, N) = k_0 + 1 \). Since \( M^*(n, N) \) is non-increasing in \( n \) and \( k_0 + 1 \leq M^*(k_0 + 1, N) \), there is a
largest integer \( n_l \in \{k_0 + 1, \ldots, N - 1\} \) such that \( M^*(n_l, N) \geq n_l \). According to Theorem 5.3.7 we have

\[
M^*(M^*(n_l, N) - 1, N) \geq n_l + 1
\]  

(5.A.5)

By definition of \( n_l \) we have

\[
M^*(n_l + 1, N) < n_l + 1.
\]  

(5.A.6)

Since \( M^* \) is non-increasing in \( n \) and because of (5.A.5) and (5.A.6), we get

\[
M^*(n_l, N) - 1 < n_l + 1
\]  

(5.A.7)

If we rewrite (5.A.7) in \( M^*(n_l, N) \leq n_l + 1 \), then, because we know that \( M^*(n_l, N) \geq n_l \), we may conclude that \( M^*(n_l, N) \) is equal to \( n_l \) or \( n_l + 1 \).

Conversely, if \( M^*(n_l, N) = n_l \) or \( n_l + 1 \), then

\[
M^*(n, N) \geq M^*(n_l, N) \geq n_l \geq n, \text{ for } n = k_0 + 1, \ldots, n_l.
\]

If \( M^*(n_l, N) = n_l \), then

\[
M^*(n, N) \leq M^*(n_l, N) = n_l < n, \text{ for } n = n_l + 1, \ldots, N - 1.
\]

If \( M^*(n_l, N) = n_l + 1 \), then

\[
M^*(n, N) \leq M^*(n_l + 1, N) = M^*(M^*(n_l, N), N) = n_l < n,
\]

for \( n = n_l + 1, \ldots, N - 1 \).

**Proof of Theorem 5.3.9**

For \( k = \{k_0 + 1, \ldots, n_l\} \) define

\[
I_k = \{M^*(k, N), M^*(k, N) + 1, \ldots, M^*(k - 1, N) - 1\}.
\]

This is well-defined as we remarked after the formulation of Theorem 5.3.8. Moreover, if \( j \in I_k \), then \( M^*(j, N) = k \). Let \( n \in \{n_l + 1, \ldots, N - 1\} \), then for some unique \( k \in \{k_0 + 1, \ldots, n_l\} \) we have \( n \in I_k \) and hence \( M^*(n, N) = k \). If we look at \( n - 1 \), then there are two possibilities:

- \( n - 1 \in I_k \) which implies \( M^*(n - 1, N) = k \);
• $n - 1 \in I_{k+1}$ and this implies $M^*(n - 1, N) = k + 1$.

Special attention has to be paid to the situation $k = n_l$, because then the possibility of $n - 1 \in I_{k+1}$ has to be excluded, since the properties of the $I_k$ can only be used for $k \in \{k_0 + 1, \ldots, n_l\}$. Suppose that $n \geq n_l + 1$ and that

$$M^*(n_l, N) \leq n \leq M^*(n_l - 1, N) - 1.$$  

If $M^*(n_l, N) = n_l$, then $n \in I_{n_l}$ also implies $n - 1 \in I_{n_l}$. If $M^*(n_l, N) = n_l + 1$ and $n \geq n_l + 2$, then $n \in I_{n_l}$ also implies $n - 1 \in I_{n_l}$. We have to check separately the case $n = n_l + 1$ and $M^*(n_l, N) = n_l + 1$. By Corollary 5.3.2 we have

$$M^*(n_l + 1, N) = M^*(M^*(n_l, N), N) = n_l,$$

and

$$M^*(n_l, N) - 1 = M^*(n_l + 1, N) < M^*(n_l, N).$$

**Proof of Theorem 5.4.1**

First we prove (5.4.2). According to Theorem 5.1.1 and (4.1.2) it suffices to prove that

$$\lambda(n, M - 1, N - 1) < \lambda(n, M - 1, N).$$

From Theorem 4.1.2, part 6 it follows that this holds for $M \geq k_0 + 1, n > k_0$ and $N \geq n + M - k_0$.

Secondly we prove (5.4.1). We observe that

$$\frac{\pi(n, M + 1, N)}{\pi(n, M, N)} > \frac{\pi(n, M + 2, N + 1)}{\pi(n, M + 1, N + 1)}$$

if and only if

$$\frac{M + 1}{M} \cdot \lambda(n, M - 1, N - 1) > \frac{M + 2}{M + 1} \cdot \lambda(n, M, N).$$

Since $\frac{M+1}{M} > \frac{M+2}{M+1}$ it suffices to prove that

$$\lambda(n, M - 1, N - 1) \geq \lambda(n, M, N).$$  (5.4.8)

For $M \leq k_0$ or $n \leq k_0$ this follows from Theorem 4.1.2, part 1, because then $\lambda(n, M - 1, N - 1) = 1$. Now, we assume $M > k_0$ and $n > k_0$. We use the
following approach. From Theorem 4.1.2, part 3 we can deduce that if \( N \geq n + M - k_0 \), then (5.A.8) is equivalent to

\[
1 - \frac{1}{g(n, M - 1, N - 1)} \geq 1 - \frac{1}{g(n, M, N)}.
\]

By writing out \( g \) this gives

\[
\frac{(N-1)}{n} \Lambda(n, M - 1, N - 1) \frac{1}{\binom{M-1}{k_0} \binom{N-M-1}{n-k_0-1}} \geq \frac{\binom{N}{n}}{\binom{M}{k_0} \binom{N-M-1}{n-k_0-1}},
\]

or, equivalently,

\[
\sum_{k=0}^{k_0} \binom{M-1}{k} \frac{1}{\binom{M-1}{k_0}} \frac{(N-M)}{n-k} \geq \sum_{k=0}^{k_0} \frac{\binom{M}{k}}{\binom{M}{k_0}} \frac{(N-M)}{n-k}.
\]

Finally, we rewrite this to

\[
\sum_{k=0}^{k_0} \frac{(M-1)}{k} \frac{(N-M)}{n-k} \geq \sum_{k=0}^{k_0} \frac{M - k_0}{M - k} \frac{(M-1)}{k_0} \frac{(N-M)}{n-k}.
\]

This inequality holds, since \( \frac{M-k_0}{M-k} \leq 1 \).

**Proof of Theorem 5.4.2**

For \( n \leq k_0 \) the proof immediately follows from Property 5.3.1, part 2. In the following we assume \( n > k_0 \). Hence \( N \geq k_0 + 1 \). By Corollary 5.3.1, \( M^*(n, N) \geq k_0 + 1 \). If \( M^*(n, N) = k_0 + 1 \), then we have \( M^*(n, N) = k_0 + 1 \leq M^*(n, N + 1) \). Suppose now that \( M^*(n, N) - 1 \geq k_0 + 1 \). Moreover, \( M^*(n, N) - 1 \leq N - n + k_0 - 1 \). By using Theorem 5.2.1 and (5.4.2) of Theorem 5.4.1, we derive that

\[
1 < \frac{\pi(n, M^*(n, N), N)}{\pi(n, M^*(n, N) - 1, N)} < \frac{\pi(n, M^*(n, N), N + 1)}{\pi(n, M^*(n, N) - 1, N + 1)}.
\]

Because of the unimodality property, we have \( M^*(n, N) \leq M^*(n, N + 1) \), since \( M^*(n, N) \) improves \( M^*(n, N) - 1 \).

Finally we consider the inequality on the right-hand side. Suppose that \( M^*(n, N + 1) = k_0 + 1 \), then \( M^*(n, N + 1) \leq M^*(n, N) + 1 \). Therefore,
we may assume that \( M^*(n, N + 1) - 1 \geq k_0 + 1 \geq 1 \). Suppose we have \( M^*(n, N) = N - n + k_0 \). Since \( M \in \{1, \ldots, N - n + k_0 + 1\} \) is the support of \( \pi(n, M, N + 1) \), we have \( M^*(n, N + 1) \leq N - n + k_0 + 1 = M^*(n, N) + 1 \). Therefore, in this case the right-hand side holds. Now we consider the situation that \( M^*(n, N) \leq N - n + k_0 - 1 \). Note that \( M^*(n, N + 1) - 1 \geq M^*(n, N) - 1 \geq k_0 + 1 \).

By using Theorem 5.2.1 and (5.4.1) of Theorem 5.4.1, we find that

\[
1 \geq \frac{\pi(n, M^*(n, N) + 1, N)}{\pi(n, M^*(n, N), N)} > \frac{\pi(n, M^*(n, N) + 2, N + 1)}{\pi(n, M^*(n, N) + 1, N + 1)}.
\]

Now, we observe that \( M^*(n, N) + 2 \) is worse than \( M^*(n, N) + 1 \). Hence, \( M^*(n, N) + 1 \) should be at least equal to or larger than \( M^*(n, N + 1) \) because of the unimodality property.

**Proof of Corollary 5.4.1**

Let \( n_l(N) \) be the largest value of \( n \) for which \( n \leq M^*(n, N) \), then Theorem 5.4.2 tells us that \( M^*(n_l(N), N + 1) \) either equals \( M^*(n_l(N), N) \) or \( M^*(n_l(N), N) + 1 \). We know by Theorem 5.3.8, part 3 that \( M^*(n_l(N), N) \) either equals \( n_l(N) \) or \( n_l(N) + 1 \).

Suppose that \( M^*(n_l(N), N) = n_l(N) \), then \( M^*(n_l(N), N + 1) = n_l(N) \) or \( n_l(N) + 1 \). From Theorem 5.3.8, part 3 it is clear that \( n_l(N + 1) = n_l(N) \).

Now suppose that \( M^*(n_l(N), N) = n_l(N) + 1 \), then \( M^*(n_l(N), N + 1) = n_l(N) + 1 \) or \( n_l(N) + 2 \). In case of \( M^*(n_l(N), N + 1) = n_l(N) + 1 \) it is clear from Theorem 5.3.8, part 3 that \( n_l(N + 1) = n_l(N) \). In case \( M^*(n_l(N), N + 1) = n_l(N) + 2 \) it follows from Theorem 5.3.8, part 3 that \( n_l(N + 1) \geq n_l(N) + 1 \).

By Theorem 5.3.8, part 1 it then follows that

\[
M^*(n_l(N) + 1, N + 1) < M^*(n_l(N), N + 1) = n_l(N) + 2.
\]

This implies

\[
M^*(n_l(N) + 1, N + 1) \leq n_l(N) + 1 \leq n_l(N + 1).
\]

This is only possible if \( n_l(N) + 1 \geq n_l(N + 1) \). Therefore we may conclude \( n_l(N + 1) = n_l(N) + 1 \) in this case.
Chapter 6

EEOQL: The Exact Version of AOQL

6.1 Introduction

This chapter will apply the theory of the previous chapter to develop an algorithm to find the optimal sample size in the EOQL-procedure for fixed values of $N$, $k_0$, and $P_l$. Because the exact underlying distribution (hypergeometric) is used to find the optimal sample size, we have called this method the Exact Expected Outgoing Quality Limit (EEOQL) method. This algorithm will be described in Section 6.2. We will also use it in Section 6.3 to generate tables (triangular arrays) with values of $M^*(n, N)$, and thus $\pi^*(n, N)$, for all possible combinations of $n$ and $N$.

Section 6.4 compares the sample sizes acquired by our EEOQL-method, the AOQL-method of Dodge and Romig, and the EOQL-method that uses an approximation for the underlying hypergeometric distribution.

Previously, we kept the value of $k_0$ fixed. Section 6.5 allows this value to vary in order to decide which value of $k_0$ we have to choose to minimize the expected value of items to be inspected.

The conclusions are discussed in Section 6.6.
6.2 Finding the optimal sample size

This section describes an algorithm that finds the optimal sample size, \(n^*\), in the EEOQL-procedure for fixed values of \(N\), \(k_0\), and \(P_l\). The following definition shows what we mean by the optimal sample size.

**Definition 6.2.1.** The optimal sample size \(n^*\), is the smallest value of \(n\) such that

\[
\pi^*(n, N) \leq P_l.
\] (6.2.1)

The algorithm heavily depends on the theory we developed in the previous chapter. The algorithm we use has two main advantages:

- It uses an efficient way of finding \(M^*(n, N)\).
- It allows us to use recursive formulas to calculate the hypergeometric distribution function.

The first advantage is very helpful because to calculate \(\pi^*(n, N)\) we need to know \(M^*(n, N)\). The second advantage is important because to calculate the expected fraction of errors after inspection we need to calculate the hypergeometric distribution function. Remember that if \(M \in \{1, \ldots, N\}\) and \(n \in \{0, \ldots, N-1\}\), then

\[
\pi(n, M, N) = \frac{M}{N} \left(1 - \frac{n}{N}\right) \cdot A(n, M - 1, N - 1).
\]

The efficient way of finding \(M^*\) and thereby \(\pi^*\), and an outline of the algorithm, are given in Subsection 6.2.1. Subsection 6.2.2 describes how we can use recursive formulas to calculate the hypergeometric distribution.

### 6.2.1 Description of the algorithm

We will assume that \(0 \leq P_l \leq 1\), \(n \in \{0, 1, \ldots, N\}\), \(N \in \{1, 2, \ldots\}\), and \(k_0 \in \{0, 1, \ldots\}\). There indeed exists an optimal sample size, because \(\pi^*(\cdot, N)\) is non-increasing, \(\pi^*(0, N) = 1\), and \(\pi^*(N, N) = 0\). As the examples below show, sometimes a simple expression for \(\pi^*\) can be found. These cases can be used in the algorithm to find the optimal sample size in an efficient way.

**Example 6.2.1.** (See Property 5.3.1, part 2). If \(n \in \{0, \ldots, k_0\}\), then we find \(M^*(n, N) = N\) and \(\pi^*(n, N) = 1 - \frac{n}{N}\). From this it immediately follows that
if \( P_l \geq 1 - \frac{k_0}{N} \), then this coincides with \( n^* \in \{0, \ldots, k_0\} \), and \( n^* \) is the smallest value of \( n \) for which \( 1 - \frac{n}{N} \leq P_l \) holds, i.e. \( n^* = \lceil (1 - P_l) \cdot N \rceil \). Notice that if \( N \in \{1, \ldots, k_0\} \), then \( n^* = \lceil (1 - P_l) \cdot N \rceil \).

**Example 6.2.2.** (See Property 5.3.1, part 3). If \( N \geq k_0 + 1 \), then we know that \( M^*(N - 1, N) = k_0 + 1 \) and \( \pi^*(N - 1, N) = \frac{k_0 + 1}{N^2} \).

**Example 6.2.3.** If \( N = k_0 + 2 \), then

a) \( M^*(n, N) = N \) and \( \pi^*(n, N) = 1 - \frac{n}{N} \) for \( 0 \leq n \leq k_0 \).

b) \( M^*(k_0 + 1, k_0 + 2) = k_0 + 1 \) and \( \pi^*(k_0 + 1, k_0 + 2) = \frac{k_0 + 1}{(k_0 + 2)^2} \).

c) \( M^*(k_0 + 2, k_0 + 3) = k_0 + 1 \) and \( \pi^*(k_0 + 2, k_0 + 3) = \frac{k_0 + 1}{(k_0 + 3)^2} \).

**Example 6.2.4.** If \( N = k_0 + 3 \), then

a) \( M^*(n, N) = N \) and \( \pi^*(n, N) = 1 - \frac{n}{N} \) for \( 0 \leq n \leq k_0 \).

b) \( M^*(k_0 + 1, k_0 + 3) = k_0 + 1 \) and \( \pi^*(k_0 + 1, k_0 + 3) = \frac{k_0 + 1}{(k_0 + 3)^2} \).

c) \( M^*(k_0 + 2, k_0 + 3) = k_0 + 1 \) and \( \pi^*(k_0 + 2, k_0 + 3) = \frac{k_0 + 1}{(k_0 + 3)^2} \).

We can use the examples above to initialize the algorithm. Diagram 1 describes this initialization. It deals with the following simple situations:

- The case \( P_l \geq 1 - \frac{k_0}{N} \) gives \( n^* = \lceil (1 - P_l) \cdot N \rceil \in \{0, \ldots, k_0\} \). This also covers the cases \( N \in \{1, \ldots, k_0\} \).
- The case \( P_l \leq \frac{k_0 + 1}{N^2} \) leads to \( n^* = N - 1 \) or \( n^* = N \).
- Explicit solutions are given for \( n^* \) in case \( N \in \{k_0 + 1, k_0 + 2, k_0 + 3\} \).

Therefore if \( N \leq k_0 + 3 \) the optimal sample size will be found by this initialization and if it does not find the optimal sample size, then we know that \( N > k_0 + 3 \) and

\[
\pi^*(N - 1, N) < P_l < \pi^*(k_0, N),
\]

(6.2.2)
since we know \( \pi^*(n, N) \) for \( n \in \{0, \ldots, k_0, N - 1, N\} \) after this initialization step.

Now, we could find the optimal sample size by calculating \( \pi^*(k_0 + 1, N) \), \( \pi^*(k_0 + 2, N) \), \ldots, until this value does not exceed \( P_l \) anymore. This means we decrease the right-hand side of (6.2.2) by increasing the sample size by one
Begin with $N$, $k_0$, and $P_l$

$P_l \geq 1 - \frac{k_0}{N}$?

YES $n^* = \lceil (1 - P_l) \cdot N \rceil$

NO $(N \geq k_0 + 1)$

$P_l \leq \frac{k_0 + 1}{N^2}$?

YES $P_l = \frac{k_0 + 1}{N^2}$?

YES $n^* = N - 1$

NO $n^* = N$

NO $(N \geq k_0 + 2)$

$N \leq k_0 + 3$?

YES $N = k_0 + 3$?

NO $(N = k_0 + 2)$ $n^* = k_0 + 1$

$N := k_0 + 1$ and $M := N - 2$

START

$P_l \leq \frac{k_0 + 1}{(k_0 + 3)^2}$?

YES $n^* = k_0 + 2$

NO $n^* = k_0 + 1$

EXIT

Diagram 1. Start of the algorithm.

unit starting at $k_0 + 1$ until we find that value for which $\pi^*$ does not exceed $P_l$ anymore. According to Definition 6.2.1 the value of $n$ for which this happens is the optimal sample size. Since for each new value of $n$ we have to find $M^*$, this may not be the most efficient method. Similarly we could try to increase the left-hand side of (6.2.2) by decreasing the sample size by one unit starting at $N - 2$ until we find that value for which $P_l$ does not exceed $\pi^*$ anymore. According
to Definition 6.2.1 the value of \( n \) for which this happens is the optimal sample size if \( \pi^* = P_l \) and else the optimal sample size is this value plus one. Still the algorithm is inefficient. This chapter will construct a kind of mixture of these methods that fully exploits the theory developed in Chapter 5.

We will show that while we try to decrease the right-hand side of (6.2.2) by increasing the sample size by one unit, we can simultaneously increase the left-hand side. Theorem 5.3.8 is the key theorem in constructing this algorithm.

The first step in the kernel of the algorithm is to decrease the right-hand side of (6.2.2) by finding \( \pi^*(k_0 + 1, N) \). Theorem 5.3.8, parts 1 and 3 tell us that \( M^*(k_0 + 1, N) < M^*(k_0, N) \). To find \( \pi^*(k_0 + 1, N) \), we can compute \( \pi(k_0 + 1, M, N) \) starting at \( M = M^*(k_0, N) - 1 = N - 1 \) and decrease \( M \) by one unit until we have found \( M^*(k_0 + 1, N) \). At the same time Theorem 5.3.8, part 2 shows that as long as \( N - 1 \geq M \geq M^*(k_0 + 1, N) \) we know that \( M^*(M, N) = \pi^*(k_0 + 1, N) \). Since \( N - 1 \geq M^*(k_0 + 1, N) \) we know that \( M^*(N - 1, N) = k_0 + 1 \) and \( \pi^*(N - 1, N) = \pi(N - 1, k_0 + 1, N) \). However, we already processed this during the initialization of the algorithm. Now, suppose that \( M = N - 2 \geq M^*(k_0 + 1, N) \), which is equivalent to \( \pi^*(k_0 + 1, N - 2, N) \geq \pi(k_0 + 1, N - 1, N) \), then we know that \( M^*(N - 2, N) = k_0 + 1 \) and \( \pi^*(N - 2, N) = \pi(N - 2, k_0 + 1, N) \). Now there are two possibilities:

- If \( \pi^*(N - 2, N) \geq P_l \), then according to the definition of \( n^* \) we obtain \( n^* \) as follows: if equality holds, then \( n^* = N - 2 \) and else \( n^* = N - 1 \).

- If \( \pi^*(N - 2, N) < P_l \), we will continue our search for \( M^*(k_0 + 1, N) \) by checking if \( N - 3 \geq M^*(k_0 + 1, N) \), which is equivalent to \( \pi^*(k_0 + 1, N - 3, N) \geq \pi(k_0 + 1, N - 2, N) \).

Thus while we are in the process of decreasing the right-hand side of (6.2.2) by increasing \( k_0 \) to \( k_0 + 1 \) we succeeded in increasing the left-hand side of (6.2.2) by decreasing \( N - 1 \) to \( N - 2 \). Suppose we have finally found \( M^*(k_0 + 1, N) \), then we have also found \( M^*(M, N) \) and thus \( \pi^*(M, N) \) for \( M = M^*(k_0 + 1, N), \ldots, N - 1 \). Note that since we have found \( M^*(k_0 + 1, N) \) the algorithm has not yet stopped and thus \( P_l > \pi^*(M^*(k_0 + 1, N), N) \). Therefore, by decreasing the right-hand side of (6.2.2) by increasing the sample size from \( k_0 \) to \( k_0 + 1 \), we have increased the left-hand side of (6.2.2) by decreasing the sample size.
from \(N - 1\) to \(M^*(k_0 + 1, N)\). If \(\pi^*(k_0 + 1, N) \leq P_l\) then the algorithm stops and \(n^* = k_0 + 1\), else we try to find \(\pi^*(k_0 + 2, N)\). Again by Theorem 5.3.8, parts 1 and 3 we know that \(M^*(k_0 + 1, N) - 1 \geq M^*(k_0 + 2, N)\), and by Theorem 5.3.8, part 2 we then know that \(M^*(M^*(k_0 + 1, N) - 1, N) = k_0 + 2\). If \(\pi^*(M^*(k_0 + 1, N) - 1, N)\) exceeds or equals \(P_l\), then the algorithm stops, else we continue our search for \(\pi^*(k_0 + 2, N)\) by checking if \(M^*(k_0 + 1, N) - 2 \geq M^*(k_0 + 2)\) and repeat the procedure above.

More generally we can say that we start our algorithm at \(n = k_0 + 1\) and \(M = N - 2\) and the algorithm stops if the following equation does no longer hold:

\[
\pi^*(M, N) < P_l < \pi^*(n, N). \tag{6.2.3}
\]

While we try to decrease the right-hand side of (6.2.3) by increasing \(n\) to \(n + 1\) we are also able to increase the right-hand side by decreasing \(M\) from \(M^*(n, N) - 1\) to \(M^*(n + 1, N)\). Notice that by this algorithm the \(P_l\) gets ‘ambushed’ from two sides at once.

Theorem 5.3.8 tells us that we are only allowed to use the algorithm above as long as \(n \leq n_l\). Remember that \(n = n_l\) if and only if \(M^*(n, N) = n\) or \(n + 1\). However, according to Remark 5.3.2 we then have found \(\pi^*(n, N)\) for all possible sample sizes. If the algorithm has not yet finished, this means that \(\pi^*(n_l + 1, N) < P_l < \pi^*(n_l, N)\). From this it is immediately clear that \(n^* = n_l + 1\). A schematic overview of the kernel of the algorithm can be found in Diagram 2. The part of the algorithm in which the question ‘\(M = n\)?’ is posed prevents double calculations in case of \(\pi^*(n_l - 1, N) < P_l < \pi^*(n_l, N)\) and \(M^*(n_l, N) = n_l\).

The algorithm poses the question ‘\(M \geq M^*(n, N)\)?’. We can check whether this inequality holds by comparing the value of \(\pi(n, M, N)\) with the value of \(\pi(n, M + 1, N)\). As long as this difference is non-negative, then we know that \(M \geq M^*(n, N)\) and if \(\pi^*(M, N) < P_l\), and the algorithm continues with computing \(\pi(n, M - 1, N)\). If this difference equals zero, then we know \(M = M^*(n, N)\), and if \(\pi^*(M, N) < P_l\), we reduce \(M\) by one and know that this value of \(M\) is smaller than \(M^*(n, N)\), therefore if \(\pi^*(n, N) = \pi(n, M + 1, N) > P_l\) and \(M > n\) we continue with increasing \(n\) by one and calculating \(\pi(n, M, N) = \pi(n, M^*(n - 1, N) - 1, N)\). If this difference is smaller than zero, then we
6.2. Finding the optimal sample size

Start with $n = k_0 + 1; M = N - 2$

Diagram 2. Schematic description of the kernel of the algorithm.

know that $M = M^*(n, N) - 1$ and if $\pi^*(n, N) = \pi(n, M + 1, N) > P_l$ and $M > n$ we continue with increasing $n$ by one and and calculating $\pi(n, M, N) = \pi(n, M^*(n - 1, N) - 1, N)$.

If we have to check whether $\pi(n, M, N) = \pi(n, M + 1, N)$, then, due to computational errors, equality will not always hold exactly when using a computer. Research did show that in cases where equality should hold the computer program gave an absolute error less than $10^{-14}$, and in cases where equality should not hold the difference was far less than $10^{-10}$. To be save we assumed
in the computer program that equality holds when the computer program gave an absolute difference less than $10^{-12}$. We used the same accuracy in checking whether $\pi^*(n, N) = P_l$.

![Graphical illustration of the algorithm](image)

**Figure 6.1.** Graphical illustration of the algorithm for $N = 10$, $k_0 = 2$, and $\pi^*(6, 10) < P_l < \pi^*(3, 10)$.

A graphical illustration is given in Figure 6.1 for $N = 10$, $k_0 = 2$ under the assumption that $\pi^*(6, 10) < P_l < \pi^*(3, 10)$. The dotted arrows symbolise the process of trying to decrease the right-hand side of (6.2.3) and the bold arrows the ‘bonus’ of increasing the left-hand side while doing so. According to Example 6.2.1 we know $M^*(0, 10) = M^*(1, 10) = M^*(2, 10) = 10$. By Example 6.2.2 we know $M^*(9, 10) = 3$ and we also know that $M^*(10, 10) = 0$. Since we know that $\pi^*(6, 10) < P_l < \pi^*(3, 10)$, the algorithm will not finish during the initialization stage of the algorithm (see Diagram 1). Since we know that $\pi(3, 9, 10) > \pi(3, 10, 10) = 0$, we know that $9 \geq M^*(3, 10)$, but this implies that $M^*(9, 10) = 3$. The points $(3, 9)$ and $(9, 3)$ are a dual pair.
Figure 6.2. Three dimensional graphical illustration of the algorithm for \(N = 10, k_0 = 2,\) and \(\pi^*(6, 10) < P_l < \pi^*(3, 10).\)

We already knew that \(M^*(9, 10) = 3\) by Example 6.2.2. This information is processed in Diagram 1. Therefore the algorithm effectively starts in Diagram 2 with \(n = k_0 + 1 = 3\) and \(M = N - 2 = 8\) (instead of \(M = 9\)). Since \(\pi(3, 8, 10) > \pi(3, 9, 10)\) we know that \(8 \geq M^*(3, 10).\) This implies \(M^*(8, 10) = 3.\) The pairs \((3, 8)\) and \((8, 3)\) are dual. Now we have obtained a new \(M^*\) and can calculate \(\pi^*(8, 10) = \pi(8, 3, 10).\) This is smaller than \(P_l.\) We note that \(8 = M \neq n = 3\) and the algorithm continues by decreasing \(M\) by one unit and it becomes 7. We find that \(7 \geq M^*(3, 10),\) because \(\pi(3, 7, 10) > \pi(3, 8, 10).\) This gives \(M^*(7, 10) = 3,\) the pairs \((3, 7)\) and \((7, 3)\) are dual, and since \(\pi^*(7, 10) < P_l\) and \(7 = M \neq n = 3\) the algorithm continues with \(M = 6.\) Because \(\pi(3, 6, 10) < \pi(3, 7, 10)\) we know that \(6 < M^*(3, 10)\) and therefore \(M^*(3, 10) = 7.\) Hence we can calculate \(\pi^*(3, 10) = \pi(3, 7, 10)\) and obtain that it is larger than \(P_l.\) We still have \(6 = M > n = 3.\) We increase \(n\) by one and it becomes 4. We know that \(M^*(3, 10) > M^*(4, 10).\) This implies
6 ≥ M*(4, 10) and therefore M*(6, 10) = 4. The pairs (4, 6) and (6, 4) are dual. We have again obtained a new M* and calculate π*(6, 10) = π(6, 4, 10) which is smaller than P*. Since 6 = M ≠ n = 4 we can continue our algorithm by decreasing M with one unit and it becomes 5.

Figure 6.2 gives a representation of the case we described above in three dimensions, i.e. n, M, and π. Notice that the projection on the plane π = 0 gives the same result as Figure 6.1. We also projected π*(n, N) = π(n, M*(n, N), N) on the plane M = 0. From this projection we can see that (6.2.3) is not violated yet, therefore the search for n* continues.

The algorithm can stop because the right-hand side inequality of (6.2.3) is violated, or because the left-hand side inequality of (6.2.3) is violated, or finally because P* lies between π*(n_l, N) and π*(n_l + 1, N) and then n* = n_l + 1. Figure 6.3 illustrates these three examples. This figure treats the same case as

![Figure 6.3](image_url)

**Figure 6.3.** The values of π*(n, 10) for all possible values of n for N = 10, k_0 = 2, with the different stop-areas A, B, C, D, and E. The values of M*(n, 10) are also given.
before with $k_0 = 2$ and $N = 10$. This figure plots the values of $\pi^*(n, 10)$ with the corresponding values of $M^*(n, 10)$ for all possible values of $n$. If $P_l$ lies in area $A$ or $E$ with inclusion of the boundaries, then the algorithm will stop during the initialization. If $P_l$ lies in area $B$ with inclusion of the lower boundary only, then the algorithm will stop because of a violation of the right-hand side inequality of (6.2.3). The algorithm will stop due to a violation of the left-hand side inequality of (6.2.3), if $P_l$ lies in area $D$ with inclusion of the upper boundary only. Finally, if $P_l$ is located within area $C$ without inclusion of the boundaries, then the algorithm will stop because we have calculated $\pi^*(n, N)$ for all possible values of $n$ by then and therefore $n^* = 5$, because $n_l = 4$.

### 6.2.2 Calculating the hypergeometric distribution

Chapter 4 deduced some recursive properties to calculate the hypergeometric distribution. Now, we will show how we use these formulas in the algorithm and show that the algorithm is indeed very efficient and accurate in calculating the hypergeometric distribution.

We know that in order to calculate $\pi(n, M, N)$ we have to calculate the value of $\Lambda(n, M - 1, N - 1)$, because if $M \in \{1, \ldots, N\}$ and $n \in \{0, \ldots, N - 1\}$, then

$$\pi(n, M, N) = \frac{M}{N} (1 - \frac{n}{N}) \cdot \Lambda(n, M - 1, N - 1).$$

The values of $\pi^*$ in the cases we treated in the initialization of the algorithm follow immediately from the Examples 6.2.1-6.2.4 and need no further explanation.

The kernel of the algorithm consists of three parts where a more detailed discussion about the calculation of $\Lambda$ is needed. These three parts are:

- **A**: The part that enables us to answer the question $M \geq M^*(n, N)$?

- **B**: The part that enables us to continue our algorithm by increasing the sample size by one unit after the question $\pi^*(n, N) > P_l$? is answered affirmatively and $M$ still exceeds $n$.

- **C**: The part that enables us to answer the question $\pi^*(M, N) < P_l$?
Part A

The kernel of the algorithm starts at \( n = k_0 + 1 \) and \( M = N - 2 \). We have to check whether \( M = N - 2 \geq M^*(k_0 + 1, N) \). To answer this question we need to compute and compare the values of \( \pi(k_0 + 1, N - 1, N) \) and \( \pi(k_0 + 1, N - 2, N) \). We start with computing \( \pi(k_0 + 1, N - 1, N) \) and therefore we need to calculate \( A(k_0 + 1, N - 2, N - 1) \). Note that \( \pi(k_0 + 1, N, N) = 0 \), because \( A(k_0 + 1, N - 1, N - 1) = 0 \). From Property 4.3.1, part 1 it is immediately clear that

\[
A(k_0 + 1, N - 2, N - 1) = 0 + \frac{(k_0 + 1)!(N - 2)!}{k_0!(N - 1)!} = \frac{k_0 + 1}{N - 1}.
\]

We continue with calculating \( \pi(k_0 + 1, N - 2, N - 1) \) and therefore we have to calculate \( A(k_0 + 1, N - 3, N - 1) \). Now, we are allowed to use Property 4.3.1, part 1 to calculate this quantity. If we apply this property to \( A(k_0 + 1, N - 3, N - 1) \) we get

\[
A(k_0 + 1, N - 3, N - 1) = A(k_0 + 1, N - 2, N - 1) + \frac{N - 3 - k_0 + 1}{N - 3 + 1} \times \frac{N - 1 - (N - 3) - 1}{N - 1 - (N - 3) - (k_0 + 1) + k_0} \times (A(k_0 + 1, N - 2, N - 1) - A(k_0 + 1, N - 1, N - 1))
\]

More generally, suppose in our algorithm \( M < M^*(n - 1, N) - 1 \), then in order to calculate \( \pi(n, M, N) \) for the purpose of checking whether \( M \geq M^*(n, N) \), we need to calculate \( A(n, M - 1, N - 1) \). If \( k_0 + 1 \leq n \leq N - 2 \) and \( k_0 + 1 \leq M \leq N - n + k_0 - 1 \), then this quantity can be calculated in a simple way by using the recursive property of Property 4.3.1, part 1. This gives

\[
A(n, M - 1, N - 1) = A(n, M, N - 1) + \frac{M - k_0}{M} \times \frac{N - M - 1}{N - M - n + k_0} \times (A(n, M, N - 1) - A(n, M + 1, N - 1)).
\]

From this property it is clear that we can calculate \( A(n, M - 1, N - 1) \) with the help of \( A(n, M, N - 1) \) and \( A(n, M + 1, N - 1) \). Since \( M < M^*(n, N) - 1 \), we have already calculated these values of \( A \) while we calculated \( \pi(n, M + \)
1, \( N \) and \( \pi(n, M + 2, N) \). Our algorithm is constructed in such a way that the conditions on \( n \) and \( M \) are always satisfied. For \( n \) this is immediately clear. The conditions on \( M \) are also satisfied. The value of \( M \) always exceeds \( k_0 + 1 \) in our algorithm because \( M^*(n, N) \geq k_0 + 1 \) and the algorithm stops if \( M = n \). The value of \( M \) does not exceed \( N - n + k_0 - 1 \) because \( M^*(n, N) \leq N - n + k_0 \) and we will only apply this property if \( M < M^*(n - 1, N) - 1 \). We are always able to apply this property if \( M < M^*(n - 1, N) - 1 \), which will become clear further on. A graphical display of the above can be found in Figure 6.4, graph A.

**Figure 6.4.** Schematic overview of the different recursive properties we use to calculate \( A \) in our algorithm.
Part B

Suppose we have found \( M^*(n-1, N) \) and \( \pi^*(n-1, N) > P_l \), then we increase the sample size by one and we have to calculate \( \pi(n, M^*(n-1, N) - 1, N) \) and therefore \( \Lambda(n, M^*(n-1, N) - 2, N-1) \) as the first step to find \( M^*(n, N) \). We will do this in two steps:

**B1**: We will compute \( \Lambda(n, M^*(n-1, N) - 1, N-1) \) from \( \Lambda(n-1, M^*(n-1, N) - 1, N-1) \) and \( \Lambda(n-1, M^*(n-1, N), N-1) \). These values of \( \Lambda \) are known to us because we computed these values to find \( \pi(n-1, M^*(n-1, N), N) \) and \( \pi(n-1, M^*(n-1, N) + 1, N) \).

**B2**: We will compute \( \Lambda(n, M^*(n-1) - 2, N-1) \) from \( \Lambda(n-1, M^*(n-1, N) - 1, N-1) \), which is already known to us, and from \( \Lambda(n, M^*(n-1, N) - 1, N-1) \), which we computed in the previous step.

We concentrate on the first step, \( B_1 \), first. Because the construction of the algorithm ensures that \( k_0 + 2 \leq n \leq N - 1 \), because we have just increased \( n \) by one and we started the kernel of the algorithm at \( n = k_0 + 1 \). It also ensures that \( k_0 + 2 \leq M^*(n-1, N) \leq N - n + k_0 + 1 \), because \( M^*(n-1, N) \geq M^*(n, N) \geq k_0 + 1 \) and \( M^*(n, N) \leq N - n + k_0 \). Therefore, we are allowed to use Property 4.3.1, part 2 to calculate \( \Lambda(n, M^*(n-1, N) - 1, N-1) \) from the values of \( \Lambda(n-1, M^*(n-1, N) - 1, N-1) \) and \( \Lambda(n-1, M^*(n-1, N), N-1) \). This gives

\[
\Lambda(n, M^*(n-1, N) - 1, N-1) \\
= \Lambda(n-1, M^*(n-1, N) - 1, N-1) - \frac{N - M^*(n-1, N)}{N-n} \times \\
\times \frac{M^*(n-1, N) - k_0 - 1}{n - k_0 - 1} \times (\Lambda(n-1, M^*(n-1, N) - 1, N-1)+ \\
- \Lambda(n-1, M^*(n-1, N), N-1)).
\]

A graphical display can be found in Figure 6.4, graph \( B_1 \).

At the second step, \( B_2 \), we know the values of \( \Lambda(n-1, M^*(n-1, N) - 1, N-1) \) and \( \Lambda(n, M^*(n-1, N) - 1, N-1) \). The construction of the algorithm ensures that \( k_0 + 2 \leq M^*(n-1, N) \leq N - n + k_0 + 1 \), which we already proved at part \( B_1 \), and \( k_0 + 1 \leq n \leq N - 1 \), which is obvious. Hence, we are allowed to use Property 4.3.1, part 3 to calculate \( \Lambda(n, M^*(n-1, N) - 2, N-1) \) from
6.2. Finding the optimal sample size

\[ A(n - 1, M^*(n - 1, N) - 1, N - 1) \text{ and } A(n, M^*(n - 1, N) - 1, N - 1). \]

This gives

\[
A(n, M^*(n - 1, N) - 2, N - 1) \\
= A(n, M^*(n - 1, N) - 1, N - 1) + \frac{n}{M^*(n - 1, N) - 1} \times \\
\times (A(n - 1, M^*(n - 1, N) - 1, N - 1) + \\
- A(n, M^*(n - 1, N) - 1, N - 1)).
\]

A graphical display can be found in Figure 6.4, graph B2. Note that, if necessary, we are now allowed to use Property 4.3.1, part 1 again to calculate \( A(n, M^*(n - 1, N) - 3, N - 1) \) to find \( \pi(n, M^*(n - 1, N) - 2, N) \) in our search for \( M^*(n, N) \).

**Part C**

In our algorithm, if \( M \geq M^*(n, N) \), then we know that \( M^*(M, N) = n \) and we have to check if \( \pi^*(M, N) \geq P_j \). To do so we have to calculate \( \pi(M, n, N) \) and thus \( A(M, n - 1, N - 1) \). If \( n = k_0 + 1 \), then \( A(M, n - 1, N - 1) = A(M, k_0, N - 1) = 1 \) for all values of \( M \) and we have no problems calculating \( \pi \). If \( n > k_0 + 1 \), then we use Theorem 4.1.1, part 1, which shows that \( A(M, n - 1, N - 1) \) equals \( A(n - 1, M, N - 1) \). To compute \( A(n - 1, M, N - 1) \) we can use Property 4.3.1, part 4, since we know the values of \( A(n, M - 1, N - 1) \) and \( A(n, M, N - 1) \). We computed the values of \( A(n, M - 1, N - 1) \) and \( A(n, M, N - 1) \) while we computed \( \pi(n, M, N) \) and \( \pi(n, M + 1, N) \). Because \( k_0 + 1 \leq M \leq N - n + k_0 \) and \( k_0 + 1 \leq n \leq N - 1 \), we are allowed to use Property 4.3.1, part 4 to calculate \( A(M, n - 1, N - 1) \). This gives

\[
A(M, n - 1, N - 1) \\
= A(n - 1, M, N - 1) \\
= A(n, M - 1, N - 1) + \frac{M - n}{n} \times (A(n, M - 1, N - 1) + \\
- A(n, M, N - 1)).
\]

A graphical display can be found in Figure 6.4, graph C. For \( M = M^*(n - 1, N) - 1 \), we have to calculate \( \pi^*(M^*(n - 1, N) - 1, N) = \pi(M^*(n - 1, N) - 1, n, N) \), and therefore \( A(M^*(n - 1, N) - 1, n - 1, N - 1) \), which corresponds with \( A(n - 1, M^*(n - 1, N) - 1, N - 1) \). But this value is already known
to us, because we computed this value when we computed $\pi^*(n - 1, N) = \pi(n - 1, M^*(n - 1, N), N)$. Notice that if $M = M^*(n - 1, N) - 2 \geq M^*(n, N)$, then we have to calculate $\pi^*(M^*(n - 1, N) - 2, N)$. We already calculated the corresponding value of $A$ when we calculated $\pi(n - 1, M^*(n - 1, N) - 1, N)$, unless $\pi(n - 1, M^*(n - 1, N), N)$ equals $\pi(n - 1, M^*(n - 1, N) + 1, N)$, because then we did not compute $\pi(n - 1, M^*(n - 1, N) - 1, N)$.

We have shown that our algorithm calculates the hypergeometric distribution function in an efficient way. But we do not know yet how accurate our computation of the hypergeometric distribution function is. To check this we compared the values of the hypergeometric distribution function we calculated with the values that are computed according to Wu (1993). Wu showed that his method using prime number factorization to rewrite the factorials, and then using cancelation of common factors to simplify the computation, compares favourably to the method described in the IMSL-libraries (IMSL libraries, 1984). In the least favourable cases the differences between our computations and those according to Wu were smaller than $10^{-13}$ and often much smaller than this. Therefore, we can state that our approach is not only efficient but also very accurate.

### 6.3 Generating tables

This section describes how we will use the results from the previous chapter to generate tables (triangular arrays) for $M^*$ and $\pi^*$ simultaneously for all pairs $(N, n)$, with $N = 0, \ldots, N_l$ and $n = 0, \ldots, N_l$, for a certain value of $k_0$. These tables can be generated up to arbitrary $N = N_l$. Notice that these tables are the same triangular arrays we mentioned in Chapter 5. We will denote these tables by $\{M_{N,n}\}$ and $\{\pi_{N,n}\}$. Their elements are denoted by $M_{N,n}$ and $\pi_{N,n}$. Of course the elements of these tables are of no interest to us if $n > N$. For $n = N$, the elements of both tables are filled with zeros because of Property 5.3.1, part 1. The elements of the first $k_0 + 1$ columns of $\{M_{N,n}\}$ and $\{\pi_{N,n}\}$ for which $n \leq k_0$ can be filled in the following way. According to Property 5.3.1, part 2, we know that $M_{N,n} = N$ and $\pi_{N,n} = 1 - \frac{n}{N}$ for $n \leq k_0$ and $N > n$. It is sensible to fill $\{M_{N,n}\}$ and $\{\pi_{N,n}\}$ column-wise. We start filling each column at $(n + 1, n)$. We increase $N$ by one until we reach $N_l$, then we go to the next column until
Table 6.1. The values of $M_{N,n}$ at $N, n \in \{0, \ldots, 7\}$ and $k_0 = 2$.

<table>
<thead>
<tr>
<th>$N \setminus n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

$n = N_l$. According to Theorem 5.4.2, we have

$$\pi_{N,n} = \max \left( \pi(n, M_{N-1,n}, N), \pi(n, M_{N-1,n} + 1, N) \right).$$

(6.3.1)

If we know the value of $M_{N-1,n}$, then either this value or this value plus one are the only two possible values that $M_{N,n}$ can take. We fill $M_{N,n}$ with the value for which $\pi_{N,n}$ takes on the larger value. It is possible that the two feasible values of $M_{N,n}$ give the same value of $\pi_{N,n}$ (see Example 5.2.1), in this case $M_{N,n} = M_{N-1,n}$ as defined in Theorem 5.4.2.

Chapter 4 introduces $P(n, M, N)$ as a $(k_0 + 1)$-vector, with elements

$$P_j(n, M, N) = P(K = j|n, M, N) = \binom{M}{j} \binom{N-M}{n-j} \binom{N}{n},$$

for $j = 0, \ldots, k_0$, and also $\iota' = (1, \ldots, 1)$, a $(k_0 + 1)$-vector. Combining this with (6.3.1), we get

$$\pi_{N,n} = \max \left( \frac{M_{N-1,n}}{N} \left(1 - \frac{n}{N} \right) \cdot \iota' \cdot P(n, M_{N-1,n} - 1, N - 1), \frac{M_{N-1,n} + 1}{N} \left(1 - \frac{n}{N} \right) \cdot \iota' \cdot P(n, M_{N-1,n}, N - 1) \right).$$

To calculate $\pi_{N,n}$, we have to calculate the probability vectors $P(n, M_{N-1,n} - 1, N - 1)$ and $P(n, M_{N-1,n}, N - 1)$. We can reduce the time needed to compute these values of the hypergeometric distribution function by using Properties 4.3.2, and 4.3.3. Since we have already computed $P(n, M_{N-1,n} - 1, N - 2)$
Table 6.2. The values of $\pi_{N,n}$ at $N, n \in \{0, \ldots, 7\}$ and $k_0 = 2$.

<table>
<thead>
<tr>
<th>$N \backslash n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.5000</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.6667</td>
<td>0.3333</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.7500</td>
<td>0.5000</td>
<td>0.1875</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.8000</td>
<td>0.6000</td>
<td>0.2400</td>
<td>0.1200</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.8333</td>
<td>0.6667</td>
<td>0.3000</td>
<td>0.1667</td>
<td>0.0833</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.8571</td>
<td>0.7143</td>
<td>0.3265</td>
<td>0.1959</td>
<td>0.1224</td>
<td>0.0612</td>
<td>0</td>
</tr>
</tbody>
</table>

When we computed $\pi_{N-1,n}$, we can use Property 4.3.2 to compute $P(n, M_{N-1,n} - 1, N - 1)$. To compute $P(n, M_{N-1,n}, N - 1)$, we can use Property 4.3.3 to compute this value from $P(n, M_{N-1,n} - 1, N - 2)$ again.

We start computing a column at $(n+1, n)$ and then we know that $M_{n+1,n} = k_0 + 1$ and we also know that $P_j(n, k_0, n) = 0$ for $j = 0, \ldots, k_0 - 1$ and $P_{k_0}(n, k_0, n) = 1$. We use $P(n, k_0, n)$ to calculate $\pi(n, k_0 + 1, n + 1)$. From this it is clear that the conditions needed to use Properties 4.3.2 and 4.3.3 are always satisfied when calculating the values of $P$ for subsequent elements of this column of $\{\pi_{N,n}\}$.

We fill the array column-wise, each time we have filled a column we increase $n$ by one. This enables us to use Theorem 5.3.8 when we look at the array row-wise. Remember that $n_l$ is the largest value of $n$ for which $n \leq M_{N,n}$.

Theorem 5.3.8 tells us that after we have found $M_{N,n}$ and it proves that $n \leq n_l$, then we immediately find the values of $\{M_{N,n}\}$ for other columns in the same row. To be more precise, then we know that $M_{N,n} = n$ for $j = M_{N,n}, \ldots, M_{N,n-1} - 1$.

Notice that Remark 5.3.2 shows that once we have found $M_{N,n}$ for all values of $n \in \{k_0 + 1, \ldots, n_l\}$, then we have found the values of $M_{N,n}$ for all the other cells of this row, i.e. for $n \in \{n_l + 1, \ldots, N - 1\}$. This property we can use in our array-filling algorithm, because we know that when we arrive at a particular cell in our algorithm and the value of $M_{N,n}$ is not yet known to us, then we know that $n \leq n_l$ and we are allowed to apply Theorem 5.3.8. Conversely, if this value is known to us when we arrive at the cell, then we know that $n > n_l$ and there is no need to apply Theorem 5.3.8, because all values of this row are already known.
6.3. Generating tables

A more schematic overview of the procedure we use to generate tables is given below.

1. Fill the elements of \( \{M_{N,n}\} \) and \( \{\pi_{N,n}\} \) with \(-1\) if \( n > N \) and the other elements with zeros.

2. For \( n \leq \min(k_0, N_l) \) and \( N > n \) calculate the elements of \( \{M_{N,n}\} \) and \( \{\pi_{N,n}\} \) by \( M_{N,n} = N \) and \( \pi_{N,n} = 1 - \frac{n}{N} \).

3. If \( N_l \leq k_0 + 1 \), then stop, else set \( n = k_0 \).

4. Set \( n = n + 1 \) and \( N = n + 1 \). Fill \( M_{N,n} \) with \( k_0 + 1 \) and \( \pi_{N,n} \) with \( \frac{k_0 + 1}{N^2} \). Set \( P(n, k_0, n) = (0, \ldots, 0, 1) \).

5. Set \( N = N + 1 \).

6. Make the distinction between the following two cases.

   (a) If \( M_{N,n} = 0 \), use Properties 4.3.2 and 4.3.3 to calculate the corresponding values of \( P \) needed to fill \( \pi_{N,n} \) and \( M_{N,n} \) with the help of (6.3.1).

   Use Theorem 5.3.8 to fill the elements \( M_{N,j} \) with the value \( n \) for \( j = M_{N,n}, \ldots, M_{N,n-1} - 1 \).

   (b) If \( M_{N,n} > 0 \), then we use the following procedure. In case \( M_{N,n} = M_{N-1,n} \), we use Property 4.3.2 to compute the value of \( P \) needed to compute \( \pi_{N,n} \). In case \( M_{N,n} = M_{N-1,n} + 1 \), we use Property 4.3.3 to compute the value of \( P \) needed to compute \( \pi_{N,n} \).

7. Return to step 5 until \( N = N_l \).

8. Return to step 4 until \( n = N_l - 1 \).

This procedure could be extended by using Theorem 5.3.1, which tells us that if \( k_0 = 0 \), then \( M_{N,n} = \lceil \frac{N-n}{n+1} \rceil \).

We have shown that our algorithm calculates the hypergeometric distribution function in an efficient way. But we do not know yet how accurate our computation of the hypergeometric distribution function is. To check this we compared
Table 6.3. Pieces of the tables \( \{M_{N,n}\} \) and \( \{\pi_{N,n}\} \) at \( N \in \{22, 23, 24\}, n \in \{20, \ldots, 24\} \) and \( k_0 = 2 \).

<table>
<thead>
<tr>
<th>( N \setminus n )</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_{N,n} )</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>( \pi_{N,n} )</td>
<td>0.0124</td>
<td>0.0062</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.0170</td>
<td>0.0113</td>
<td>0.0057</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.0208</td>
<td>0.0156</td>
<td>0.0104</td>
<td>0.0052</td>
<td>0</td>
</tr>
</tbody>
</table>

the values of the hypergeometric distribution function we calculated with the values that are computed according to Wu (1993) as we did in Section 6.2. In the least favourable cases the difference was smaller than \( 10^{-12} \) and often much smaller than this. Therefore, we can state that our approach is not only efficient but also very accurate.

Our algorithm also has to check whether

\[
\pi(n, M_{N-1,n}, N) = \pi(n, M_{N-1,n} + 1, N)
\]

Of course due to computational errors this will not always hold exactly when using a computer. Research shows that we may assume that equality holds if the computer gives a difference less than \( 10^{-12} \).

If we would use these tables to find the optimal sample size in an EOQL-sampling plan for a certain predefined value of \( P_l \) with fixed \( k_0 \) and \( N \), then we have to find the smallest value of \( n \) such that

\[
\pi_{N,n} \leq P_l
\]

Tables 6.1, 6.2, and 6.3 give some examples of pieces of tables generated for \( k_0 = 2 \). Notice that in Table 6.2 the optimal solution for \( P_l = 0.01 \) for all tabulated population sizes is to check the entire population. In fact, it turns out that \( N = 18 \) is the first value of \( N \) for which we can find an optimal solution \( (n^* = 17, M^*(17, 18) = 3 \) and \( \pi^*(17, 18) = 0.0093) \) without checking the
entire population. Suppose we want to find the optimal \( n \) at \( N = 24 \), \( k_0 = 2 \) and \( P_L = 0.01 \). Using Table 6.3, we would find \( n = 23 \) as the optimal solution, because \( n = 23 \) is the smallest value of \( n \) for which the function \( \pi^*(n, N) \leq 0.01 \). But \( \pi(23, 3, 24) = 0.0052 \), which is considerably smaller than \( P_L \). In this case it could well be that \( n = 22 \) would be preferred as optimal solution, because \( \pi(22, 3, 24) = 0.0104 \) lies considerably closer to \( P_L \). This shows a clear advantage of using a table; it provides more insight.

Before using the theory we have developed in this chapter, it took us more than a day to generate \( \{M_{N,n}\} \) and \( \{\pi_{N,n}\} \) for \( N_l = 2000 \) and \( k_0 = 2 \) on a Pentium 4 computer. Using the algorithm we presented in this section it takes under two minutes to accomplish this task. This shows how efficient the algorithm actually is.

## 6.4 Comparison between the various methods

The previous sections described an algorithm how we can find the optimal sample size of the EOQL-method, which is an improvement of the AOQL-method, using the true underlying hypergeometric distribution instead of an approximation by the Poisson distribution. This section will compare the AOQL-method used by Dodge and Romig, the EOQL-method we discussed in Chapter 5 proposed by Simons et al. (1989), and the method developed by us. The first two methods use the approximation by the Poisson distribution and our method uses the true underlying hypergeometric distribution. We will call our method the EEOQL-method (Exact-EOQL).

Table 6.4, 6.5 and 6.6 give the result of our comparison of the optimal sample sizes under varying conditions of the three different approaches. From the tables we can see that the EOQL-method tends to be too conservative and the sample sizes of the AOQL-method tend to be too tight. Hence, the sampling costs are too high for the EOQL-method and the AOQL-method does not guarantee that the expected fraction of errors after inspection does not exceed \( P_I \). Simons et al. already showed that for large \( N \), the AOQL-method and the EOQL-method give the same optimal sample size. From the tables we can see that for large \( N \), the EEOQL-method gives the same optimal sample size. This is exactly what we would expect, because for large \( N \), the Poisson-approximation becomes better.
We can also see that for larger values of $k_0$ and smaller values of $P_L$, the differences between the methods become larger. These differences can be quite substantial.
6.4. Comparison between the various methods

Table 6.4. Comparison between the optimal sample sizes, using the EEOQL-method, the AOQL-method, and the EOQL-method at $P_L = 0.005$ for $k = 0, 1, 2$.

<table>
<thead>
<tr>
<th>$P_L = 0.005$</th>
<th>$N$</th>
<th>EEOQL</th>
<th>AOQL</th>
<th>EOQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_0 = 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>38</td>
<td>29</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>47</td>
<td>38</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>43</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>58</td>
<td>50</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>64</td>
<td>57</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>68</td>
<td>64</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>70</td>
<td>67</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>71</td>
<td>69</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>73</td>
<td>74</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>$k_0 = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>44</td>
<td>39</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>61</td>
<td>52</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>75</td>
<td>63</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>94</td>
<td>80</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>117</td>
<td>101</td>
<td>134</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>139</td>
<td>126</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>148</td>
<td>138</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>153</td>
<td>144</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>167</td>
<td>166</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>$k_0 = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>46</td>
<td>43</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>66</td>
<td>59</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>84</td>
<td>74</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>113</td>
<td>97</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>151</td>
<td>131</td>
<td>174</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>198</td>
<td>177</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>220</td>
<td>201</td>
<td>229</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>232</td>
<td>215</td>
<td>239</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>270</td>
<td>267</td>
<td>271</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.5. Comparison between the optimal sample sizes, using the EEOQL-method, the AOQL-method and the EOQL-method at $P_L = 0.01$ for $k_0 = 0, 1, 2$.

<table>
<thead>
<tr>
<th>$P_L = 0.01$</th>
<th>[ N ]</th>
<th>[ EEOQL ]</th>
<th>[ AOQL ]</th>
<th>[ EOQL ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ k_0 = 0 ]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>25</td>
<td>22</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>29</td>
<td>25</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>31</td>
<td>27</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>33</td>
<td>30</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>34</td>
<td>32</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>36</td>
<td>35</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>36</td>
<td>36</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>36</td>
<td>36</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>[ k_0 = 1 ]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>38</td>
<td>32</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>47</td>
<td>40</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>54</td>
<td>46</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>62</td>
<td>54</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>70</td>
<td>63</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>77</td>
<td>72</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>79</td>
<td>76</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>80</td>
<td>78</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>[ k_0 = 2 ]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>42</td>
<td>37</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>57</td>
<td>49</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>67</td>
<td>58</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>83</td>
<td>72</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>99</td>
<td>89</td>
<td>106</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>116</td>
<td>108</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>123</td>
<td>116</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>126</td>
<td>121</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>136</td>
<td>136</td>
<td>137</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.6. Comparison between the optimal sample sizes, using the EEQQL-method, the AOQL-method and the EOQL-method at $P_L = 0.05$ for $k_0 = 0, 1, 2$.

<table>
<thead>
<tr>
<th>$k_0 = 0$</th>
<th>$N$</th>
<th>EEQQL</th>
<th>AOQL</th>
<th>EOQL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>7</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>$k_0 = 1$</td>
<td>50</td>
<td>14</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>15</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>16</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>16</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>16</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>$k_0 = 2$</td>
<td>50</td>
<td>20</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>22</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>24</td>
<td>22</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>25</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>26</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>27</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>27</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>28</td>
<td>28</td>
<td>28</td>
</tr>
</tbody>
</table>
The EOQL-method with the Poisson-approximations, uses the following definition for the expected fraction of errors after inspection

\[
\pi(n, M, N) = \min(k_0, M) \sum_{k=0}^{M-k} \frac{M-k}{N} \cdot P(K = k | n, M, N).
\]

This method approximates the hypergeometric probability

\[
P(K = k | n, M, N) = \binom{M}{k} \binom{N-M}{n-k} \binom{n}{k},
\]

by the Poisson probability

\[
P(K = k | n, M, M) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad \text{with} \quad \lambda = \frac{nM}{N}.
\]

But we can also write \(\pi\) as

\[
\pi(n, M, N) = \frac{M}{N} (1 - \frac{n}{N}) \cdot \sum_{k=0}^{k_0} P(K = k | n, M - 1, N - 1),
\]

and approximate the hypergeometric probability

\[
P(K = k | n, M - 1, N - 1) = \binom{M-1}{k} \binom{N-M}{n-k} \binom{n-1}{k},
\]

by the Poisson probability

\[
P(K = k | n, M - 1, N - 1) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad \text{with} \quad \lambda = \frac{n(M-1)}{N-1}.
\]

We will call this approach the modified EOQL-method. Table 6.7 compares this modified EOQL-method with our exact approach for different values of \(k_0\) and \(P_l\). The values of the optimal sample sizes of the modified EOQL-approach lie closer to the values of the exact approach than the values of the AOQL- and EOQL-approach. However, the modified EOQL-method still differs from the exact method. Especially for small values of \(k_0\) and \(P_l\), and small population sizes the modified EOQL approach gives optimal sample sizes which are too small, and for larger population sizes it gives optimal sample sizes which are too large. For very large population sizes the modified EOQL-method and the exact method give the same optimal sample size again.
Table 6.7. Comparison between the optimal sample sizes, using the EEOQL-method (E) and the modified EOQL-method (P) at different values of $P_L$ for $k_0 = 0, 1, 2$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$P_L = 0.5%$</th>
<th>$P_L = 1%$</th>
<th>$P_L = 5%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_0 = 0$</td>
<td>E</td>
<td>P</td>
<td>E</td>
</tr>
<tr>
<td>50</td>
<td>38</td>
<td>34</td>
<td>25</td>
</tr>
<tr>
<td>75</td>
<td>47</td>
<td>45</td>
<td>29</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>53</td>
<td>31</td>
</tr>
<tr>
<td>150</td>
<td>58</td>
<td>63</td>
<td>33</td>
</tr>
<tr>
<td>250</td>
<td>64</td>
<td>70</td>
<td>34</td>
</tr>
<tr>
<td>500</td>
<td>68</td>
<td>73</td>
<td>36</td>
</tr>
<tr>
<td>750</td>
<td>70</td>
<td>74</td>
<td>36</td>
</tr>
<tr>
<td>1000</td>
<td>71</td>
<td>74</td>
<td>36</td>
</tr>
<tr>
<td>10000</td>
<td>73</td>
<td>74</td>
<td>37</td>
</tr>
<tr>
<td>$k_0 = 1$</td>
<td>E</td>
<td>P</td>
<td>E</td>
</tr>
<tr>
<td>50</td>
<td>44</td>
<td>42</td>
<td>38</td>
</tr>
<tr>
<td>75</td>
<td>61</td>
<td>58</td>
<td>47</td>
</tr>
<tr>
<td>100</td>
<td>75</td>
<td>72</td>
<td>54</td>
</tr>
<tr>
<td>150</td>
<td>94</td>
<td>93</td>
<td>62</td>
</tr>
<tr>
<td>250</td>
<td>117</td>
<td>118</td>
<td>70</td>
</tr>
<tr>
<td>500</td>
<td>139</td>
<td>143</td>
<td>77</td>
</tr>
<tr>
<td>750</td>
<td>148</td>
<td>152</td>
<td>79</td>
</tr>
<tr>
<td>1000</td>
<td>153</td>
<td>156</td>
<td>80</td>
</tr>
<tr>
<td>10000</td>
<td>167</td>
<td>167</td>
<td>84</td>
</tr>
<tr>
<td>$k_0 = 2$</td>
<td>E</td>
<td>P</td>
<td>E</td>
</tr>
<tr>
<td>50</td>
<td>46</td>
<td>45</td>
<td>42</td>
</tr>
<tr>
<td>75</td>
<td>66</td>
<td>63</td>
<td>57</td>
</tr>
<tr>
<td>100</td>
<td>84</td>
<td>79</td>
<td>67</td>
</tr>
<tr>
<td>150</td>
<td>113</td>
<td>107</td>
<td>83</td>
</tr>
<tr>
<td>250</td>
<td>151</td>
<td>146</td>
<td>99</td>
</tr>
<tr>
<td>500</td>
<td>198</td>
<td>197</td>
<td>116</td>
</tr>
<tr>
<td>750</td>
<td>220</td>
<td>220</td>
<td>123</td>
</tr>
<tr>
<td>1000</td>
<td>232</td>
<td>233</td>
<td>126</td>
</tr>
<tr>
<td>10000</td>
<td>270</td>
<td>270</td>
<td>136</td>
</tr>
</tbody>
</table>
6.5 Choice of $k_0$

Previously, we considered $k_0$ to be fixed. If we allow $k_0$ to vary, we can minimize the expected number of items to be inspected. We denote the number of items to be inspected by $I$, and we denote $W = E(I)$. If we find less or equal than $k_0$ defects in the sample we have to inspect $n$ items, and if we find more than $k_0$ defects we have to inspect the entire population. This gives the following expression for $E(I)$,

$$W(k_0, n, M, N) = E(I(k_0, n, M, N)) = n \cdot P\{K \leq k_0|n, M, N\} + N \cdot P\{K > k_0|n, M, N\} = N - (N - n) \cdot P\{K \leq k_0|n, M, N\}.$$  

Using the EEOQL-procedure, we can find the optimal sample size $n^*$, which depends on $N$, $k_0$ and $P_l$. If we look at $W$ for various values of $k_0$ keeping $N$ and $P_l$ fixed, then the values of $k_0$ are accompanied by the corresponding values of $n^*(N, k_0, P_l)$. The expected number of items to be inspected also depends on $M$, and this is an unknown quantity to us. To make a sensible choice out of the possible values of $k_0$, we first have to observe the behaviour of $W$ for different values of $k_0$ while we vary $M$. Therefore we study the curves $W_{k_0}(M)$ for varying combinations of $N$ and $P_l$. These curves are defined by

$$W_{k_0}(M|N, P_l) = N - (N - n^*(N, k_0, P_l)) \cdot P\{K \leq k_0|n^*(N, k_0, P_l), M, N\}.$$  

Notice that $W_{k_0}(M)$ is equal to $n^*(N, k_0, P_l)$ if $M \leq k_0$, and $W_{k_0}(M) = N$ if $M > N - n^*(N, k_0, P_l) + k_0$. For other values of $M$ the function $W_{k_0}(M)$ is a monotone increasing function of $M$. Figure 6.5 shows $W_{k_0}(M)$ for different values of $k_0$ while we vary $P_l$ keeping $N$ fixed. Figure 6.6 shows $W_{k_0}(M)$ for different values of $k_0$ while we vary $N$ keeping $P_l$ fixed. Of course, we are interested in $\min_{k_0} W_{k_0}(M)$. We will denote the value of $k_0$ for which $W_{k_0}(M)$ takes on its minimum by $k_0^*(M)$. The figures show that $\min_{k_0} W_{k_0}(M)$ is very sensitive in $M$. By sensitive we mean that a small change in $M$ will change the value of $k_0$ for which $W_{k_0}(M)$ will take on its minimum. In fact, the smaller $N$ and $P_l$ are, the more sensitive $\min_{k_0} W_{k_0}(M)$ becomes. However, the differences between the different values of $W_{k_0}(M)$ become smaller, choosing the wrong value of $k_0$ has less consequences for the expected amount of work to be done.
Figure 6.5. The function $W(k_0(M))$ for values of $k_0 \in \{0, 1, 2, 3, 5, 10, 15\}$ ($k_0 = 0$ is striped), at $N = 250$ and for values of $P_l$ equal to 0.5%, 1%, 2.5%, and 5%. The vertical line equals $\lceil N \cdot P_l \rceil$. 
Figure 6.6. The function $W_{k_0}(M)$ for values of $k_0 \in \{0, 1, 2, 3, 5, 10, 15\}$ ($k_0 = 0$ is striped), at $P_l = 0.5\%$ and for values of $N$ equal to 50, 100, 500, and 1000. The vertical line equals $\lceil N \cdot P_l \rceil$. 
Another observation we make from observing these curves is that $k^*_0(M)$ seems to increase with $M$.

Since we do not have any knowledge of the true value of $M$, it is impossible to choose a value of $k_0$ that minimizes the expected amount of items to be inspected. Hence, finding a minimizing value of $k_0$ is an illusion. If we have some information (confidence interval) or make some assumptions on what the value of $M$ might be, finding an minimizing $k_0$ will still be very difficult, if not impossible, due to the sensitivity of $k^*_0$ in $M$.

Looking at Figure 6.5 and 6.6 a strategy could be constructed which could minimize the amount of work in the long run. If we would apply the EEQL-procedure several times in a row, it can be expected that $M$ lies in the neighbourhood (perhaps a bit less) of the number of errors allowed in the population, i.e. $N \cdot P_l$. Because $N \cdot P_l$ can take a non-integer value, in contrast to $M$, we will use $\lceil N \cdot P_l \rceil$ instead of $N \cdot P_l$. If our assumption holds, it is not difficult to see that choosing $k_0$ equal to zero (striped line in the figures) would not be very cost effective in most cases, although this is often done in practice. Choosing $k_0 = k^*_0(\lceil N \cdot P_l \rceil)$ would be minimizing if the true value of $M$ equals $\lceil N \cdot P_l \rceil$.

For values of $M$ not much larger than this value this choice of $k_0$ might not be minimizing anymore, probably it will be minimizing for a value of $k_0$ larger than the one we chose, but if we look at Figure 6.5 and 6.6 the difference with the minimizing solution seems to be relatively small. For values of $M$ not much smaller than $\lceil N \cdot P_l \rceil$ this choice of $k_0$ might not be minimizing anymore too, most likely it will be minimizing for some value of $k_0$ smaller than the one we chose, but looking at Figure 6.5 and 6.6 the difference with the minimizing solution can be quite substantial. By choosing $k_0 = k^*_0(\lceil N \cdot P_l \rceil) - 1$, it seems we solve this problem largely and still the difference with the minimizing solution for values of $M$ not much larger than $\lceil N \cdot P_l \rceil$ will not become too large. For very small values of $M$ and for relatively large values of $M$ the difference with the minimizing solution still can be very substantial. But for very large values of $M$ one would expect an inspection of the entire population, and probably this will give insight in what causes the high number of defects. Measures will be taken to solve this problem and probably the value of $M$ will decrease rapidly. Therefore, the method will not stay inefficient for a long time.

We will also study the behaviour of $W_M(k_0)$. This means we vary $k_0$ while we
fix $M$. We are especially interested in values of $M$ equal to $M^*(n^*(N, k_0, P_l), N)$. It is possible that for different values of $k_0$ we find the same value of $M^*$.

Now, we denote by $K(M)$ the set of all possible values of $k_0$ for which $M = M^*(n^*(N, k_0, P_l), N)$. This could also be an empty set.

**Conjecture 6.5.1.** Let $K(M)$ be a non-empty set. If $M > 0$, then $\min_{k_0} W_{M-1}(k_0)$ takes on its minimum for a value of $k \in K(M)$.

Note that if $K(M)$ has only one element, we exactly know for which value of $k_0$ the function $\min_{k_0} W_{M-1}(k_0)$ takes on its minimum. The difference between the minimal solution and the solution for the other elements of $K(M)$ is not very large. This conjecture shows that if we find a certain optimal sample size $n^*(N, k_0, P_l)$ with corresponding $M^*(n^*(N, k_0, P_l), N)$, then in the neighbourhood of this $M^*$ the amount of work to be done when we use this value of $k_0$ will not differ too much from the minimal solution.

Further research should be done to determine the way in which $k_0$ could be chosen. Here, we only gave some heuristic solutions.

### 6.6 Conclusions

The Exact Expected Outgoing Quality Limit (EEOQL) method is a very useful control instrument, which is easy to implement with the help of general software. It does not make the mistake Dodge and Romig made in their AOQL-method, which leads to sample sizes that are too low. Hence, the AOQL-method gives a false feeling of reassurance, because this method does not guarantee that the expected fraction of errors after inspection does not exceed the predefined level set by management. The EEOQL-method uses the underlying hypergeometric distribution instead of the Poisson approximation. This Poisson approximation leads to sample sizes that are too large and therefore it leads to sampling costs that are too high. The EEOQL-method is both efficient as well as accurate. Especially for relatively small values of $N$ and $P_l$, our EEOQL-method compares favourably to the AOQL-method of Dodge and Romig and the EOQL-method of Simons et al. (1989).
Bibliography


Clopper, C.J. and E.S. Pearson (1934), ‘The use of confidence or fiducial limits illustrated in the case of the binomial’, Biometrika 26, 404–413.


Bibliography


IMSL libraries (1984), Houston, Texas: IMSL.


Bibliography


Van Manen, J.A. (1990), Harde en zachte cijfers, Rede aanvaarding ambt van hoogleraar, University of Groningen.


Samenvatting (Summary in Dutch)

De accountant moet beoordelen of de jaarrekening die hij controleert een getrouw beeld geeft, opdat gebruikers van de jaarrekening met een ‘gerust’ hart hun economische beslissingen kunnen nemen op basis van de informatie die de jaarrekening hen verschafte. Om tot dit oordeel te komen probeert de accountant betrouwbare informatie te verzamelen. De gegevens die de accountant moet beoordelen zijn niet altijd onomstreden. De beoordeling wordt mede gebaseerd op schattingen en voorspellingen. De accountant en de verantwoordelijke voor de jaarrekening verschillen nogal eens van mening over deze zogenaamde ‘zachte’ cijfers en vaak wordt dan een compromis tussen beide gesloten. De gebruiker is niet op de hoogte van de gevoerde discussie en zou misschien andere beslissingen nemen wanneer dit wel het geval was. Feit blijft dat de accountant op basis van de door hem verzamelde en beschikbare informatie een zo getrouw mogelijk beeld van de werkelijkheid probeert te creëren.

Hier kunnen we parallellen trekken met de statisticus. Allereerst probeert een statisticus ook om een zo getrouw mogelijk beeld te schetsen op basis van informatie die onzeker is. Zowel de statisticus als de accountant hebben te maken met ‘zachte’cijfers. Ten tweede is het verzamelen van betrouwbare informatie tegen lage kosten essentieel voor het economisch en efficiënt controleren van de jaarrekening. Het nemen van steekproeven is dan vaak een uitkomst. Statistici zijn natuurlijk experts op dit gebied, rekening houdend met de rol die het toeval speelt kunnen zij onderzoeken of afwijkingen te wijten zijn aan het toeval of dat een systematische fout hier de oorzaak van is. De statisticus kan de accountant prima van dienst zijn.

Dit proefschrift richt zich op twee soorten controles waar de statistiek, en
Samenvatting

dan met name de steekproeftheorie, de accountant kan helpen. Om te bepa-
len welke en hoeveel gegevensgerichte controles de accountant moet uitvoeren
zal hij onder meer willen weten hoe de interne beheersingsmaatregelen fun-
tioneren, die van belang zijn voor de controle van de jaarrekening. Wanneer de
accountant het idee heeft dat deze niet goed werken dan zal de accountant meer
gegevensgerichte controles moeten uitvoeren om het controllerisico te beperken.
Deze gegevensgerichte controles kunnen bestaan uit o.a. cijferbeoordelingen,
verbandsc controles en/of detailcontroles.

De gegevensgerichte detailcontroles kunnen onder meer toegepast worden
om na te gaan of de balansen op de jaarrekening geen materiële fout bevat-
ten. Van een steekproef uit de posten van de te cont rolleren populatie worden
de geboekte waarden vergeleken met de waarden die de posten uit de populatie
volgens de accountant zouden moeten hebben. Op basis van de informatie uit de-
ze steekproef wordt dan een schatting gemaakt voor het totale verschil tussen de
geboekte en ‘echte’ waarden van de posten. Aange zien deze schatting gebaseerd
is op een steekproef gaat deze gepaard met onzekerheid. Om deze onzekerheid
tot uiting te brengen wordt niet de schatting van de totale afwijking maar de
betrouwbaarheidsbovengrens gebruikt om te bepalen of de totale afwijking mo-
gelijk de materialiteit overschrijdt rekening houdend met de onzekerheid. Een
populatie die gecontroleerd wordt door de accountant bezit over het algemeen
weinig posten waarvoor de accountant een ander bedrag zou vinden dan het be-
drag dat in de boeken staat. Dat wil zeggen dat een steekproef dan ook weinig in-
fomatie oplevert over de verdeling van de hoogte van het afwijkende bedrag als
een post inderdaad afwijk t. Deze eigenschap zorgt ervoor dat de statistische me-
thoden die normaal gesproken vaak gebruikt worden, de zogenaamde klassieke
methoden, niet goed werken wanneer ze hier worden toegepast. In de afgelopen
decennia is er een scala aan statistische methoden ontwikkeld om dit probleem
to ondervangen. Een overzicht van deze methoden tot aan het einde van de jaren
tachtig wordt gegeven in het rapport van de National Research Council on Sta-

tistical Models and Analysis in Auditing uit 1989. Hoofdstuk 2 van dit proefschrift
geeft een actueel overzicht. Vooral de zogenaamde ‘Bayesiaanse’ methoden, die
de accountant in staat stellen om informatie gebaseerd op ervaring en expertise te
gebruiken om in combinatie met de informatie uit de steekproef tot een schatting
van de betrouwbaarheidsbovengrens te komen, en ‘resample’ methoden, die met
behulp van de getrokken steekproef heel veel nieuwe steekproeven genereren,
om zo ‘extra’ informatie te verkrijgen en met behulp van die informatie een betrouwbaarheidsbovengrens bepalen, zijn de laatste twee decennia zeer populair. In de praktijk wordt nog altijd veel gebruik gemaakt van de ‘Stringer bound’. Het succes van deze methode moet haast wel liggen aan het feit dat deze relatief makkelijk is uit te rekenen, want een bevredigende verklaring voor deze heuristische methode is tot op heden nooit gegeven, zelfs geen intuïtieve. Bovendien blijkt uit diverse onderzoeken dat deze methode zeker niet superieur is aan andere alternatieven en vaak erg conservatief is, dat wil zeggen dat de werkelijke betrouwbaarheid de gewenste betrouwbaarheid overschrijdt.

Het rapport van de National Research Council on Statistical Models and Analysis in Auditing refereerde reeds aan “… the generally scattered and ad hoc nature of the existing methodology”. Hedentendage valt nog steeds het gebrek aan coherente tussen de methoden en de overvloed aan gelegenheidsmethoden op. Het gebrek aan structuur is wellicht inherent aan het feit dat iedere beroepsgroep, dus ook de accountancy, experimenteert met allerlei methoden, gemotiveerd of niet. Zo groeit er dan een systeem dat weinig gestructureerd is, maar door de terugkoppeling aan ervaringen toch in de meeste situaties lijkt te voldoen. Bovendien zijn accountants terughoudend met het uitwisselen van gegevens omdat het hier vaak ‘gevoelige’ informatie betreft. Dit komt de structuur natuurlijk ook niet ten goede. Maar ook voor gelegenheidsoplossingen blijft natuurlijk van belang dat een methode daadwerkelijk voldoet en niet alleen lijkt te voldoen. Een goede dialoog tussen statisticus en accountant is hiervoor erg belangrijk. Meer algemeen onderzoek blijft van evident belang omdat vele onderzoeks vragen nog onbeantwoord zijn. Het rapport van de National Research Council on Statistical Models and Analysis in Auditing gaf reeds een opsomming van deze vragen, en vele van die vragen zijn nog steeds niet beantwoord. Bovendien blijft het natuurlijk een uitdaging voor onderzoekers om die methode te vinden die alle andere overtreft.

Frauduleuze rapportering door enkele grote Amerikaanse en Europese beurs genoteerde ondernemingen heeft het vertrouwen in kapitaalmarkten, bestuurders van ondernemingen, toezichthouders en in de overige schakels van de verslag gevingsketen behoorlijk doen afnemen. Om dit vertrouwen te laten toenemen hebben verscheidene landen aanvullende wetgeving ontworpen, of zijn deze aan het ontwerpen, omtrent de rapportering van de interne beheersing door beurs genoteerde bedrijven. Zo is in de Verenigde Staten de door de senatoren Sarbannes
en Oxley ontworpen Sarbannes-Oxley wet (SOx) opgelegd aan beursgenoteerde bedrijven. Om frauduleuze rapportering tegen te gaan was eerder al het COSO (Committee of Sponsoring Organizations of the treadway Commision) rapport opgesteld. Dit rapport geeft een gemeenschappelijke definitie van het begrip interne beheersing en presenteert een raamwerk waarmee het mogelijk is om de interne beheersingsmaatregelen te beoordelen en te verbeteren. Blijkbaar voldoet dit rapport niet voldoende en is strengere regelgeving nodig. De nieuwe regelgeving richt zich met name op het gedeelte van de interne beheersing dat betrekking heeft op de financiële rapportage (ICFR). SOx stelt dat de organisatie dient te verklaren middels kwartaal- en jaarrapportages dat de betrouwbaarheid van de informatieverzorging is gewaarborgd en het totaalbeeld dat daaruit ontstaat met betrekking tot ICFR niet misleidend is. De stappen die hiertoe ondernomen moeten worden zijn nu niet meer exclusief weggelegd voor de externe accountant, maar moeten ook door de organisatie zelf worden uitgevoerd.

De externe accountant zal naast de verklaring over de jaarrekening ook een verklaring over ICFR moeten afgeven. De PCAOB (Public Company Accounting Oversight Board) die de regels voor controle van de externe accountant opstelt, heeft een standaard uitgebracht die niet alleen een controle vereist van het evaluatieproces dat door het management is uitgevoerd, maar ook een zelfstandig oordeel vereist over de effectiviteit van ICFR. Deze controle van ICFR moet door dezelfde accountant gedaan worden als die de jaarrekening controleert.

Deze nieuwe ontwikkelingen omtrent interne beheersing, zorgen er onder meer voor dat er behoefte is aan goede en betrouwbare controleprocedures, zowel voor gebruik door de organisatie zelf als voor gebruik door de externe accountant. Een procedure die in de praktijk reeds jaren gebruikt wordt om te bepalen of een proces voldoet aan de regels die door de interne beheersing worden opgelegd is de AOQL-procedure (AOQL = Average Outgoing Quality Limit). Deze, oorspronkelijk door Dodge and Romig (1959) ontwikkelde, methode garandeert, wanneer we dit proces over een bepaalde periode beschouwen, dat het verwachte percentage eenheden dat na de controle afwijkt van de regels opgelegd door de interne beheersing een van te voren bepaalde grens niet overschrijdt. Deze methode was oorspronkelijk ontwikkeld voor industriële doeleinden, maar werd later ook toegepast in administratieve processen en gebruikt voor accountancy doeleinden. Het proces dat beschouwd wordt, produceert dus een populatie die juiste en onjuiste elementen kan bevatten. De AOQL-methode wordt dan op
de volgende manier toegepast.

- Het maximale acceptabele te verwachten aantal onjuiste elementen dat na inspectie in de populatie achter mag blijven, dient door het management te worden vastgesteld.

- De populatie onder controle dient te worden opgesplitst in een aantal sub-populaties (bijvoorbeeld op basis van weken, maanden of kwartalen). Deze keuze wordt zo gemaakt dat een optimale situatie ontstaat met betrekking tot het aantal steekproefelementen dat moet worden onderzocht.

- Voordat uit elke subpopulatie een aselecte steekproef wordt getrokken zonder teruglegging, wordt de acceptatiegrens bepaald, dat wil zeggen dat de populatie wordt afgekeurd als het aantal elementen in de steekproef dat afwijkt deze acceptatiegrens overschrijdt. Een afgekeurde subpopulatie wordt vervolgens integraal gecontroleerd.

- Alle onjuiste elementen in de steekproef en in eventuele integrale controles moeten worden gecorrigeerd.

De keuze van de optimale steekproefgrootte kan worden afgelezen met behulp van speciale tabellen. De optimale steekproefgrootte is de minimale steekproefgrootte waarvoor in het meest ongunstige geval nog gegarandeerd kan worden dat de verwachte fractie onjuiste elementen dat na inspectie in de populatie mag achterblijven een van tevoren gedefinieerde grens niet overschrijdt. Helaas bevat de wiskundige formulering van Dodge en Romig een fout zodat deze tabellen niet de juiste gegevens bevatten. Dit is reeds opgemerkt door Simons et al. (1989). Zij presenteren een nieuwe methode, de zogenaamde EOQL-methode (EOQL= Expected Outgoing Quality Limit). Behalve dat deze methode wiskundig gezien enige bezwaren heeft, gebruikt deze net als de AOQL-methode ook nog eens een wiskundige benadering om de optimale steekproefgrootte te berekenen. De Poisson verdeling wordt gebruikt als benadering voor de werkelijke onderliggende hypergeometrische verdeling. Vooral als de populatieomvang beperkt is, zorgt deze benadering voor afwijkingen ten opzichte van de optimale steekproefgrootte als er geen benadering wordt gebruikt. Hoofdstuk 3 van dit proefschrift beschrijft een onderzoek dat bij de IB-Groep op de afdeling Examendiensten is uitgevoerd, waarin populaties van beperkte omvang inderdaad
een rol speelden tijdens de implementatie van de EOQL-methode op de controle
van administratieve processen.

Vandaar dat in hoofdstuk 4, 5 en 6 een methode wordt ontwikkeld, die gebruik
maakt van de werkelijke onderliggende hypergeometrische verdeling in plaats
van de benadering met de Poisson verdeling. In hoofdstuk 4 worden nieuwe
eigenschappen voor de hypergeometrische verdeling afgeleid, waarmee vervol-
gens in hoofdstuk 5 essentiële eigenschappen worden afgeleid voor de grootte-
den die in de EOQL-methode een belangrijke rol spelen. Deze eigenschappen in
combinatie met bepaalde recursieve kenmerken van de hypergeometrische ver-
deling hebben geleid tot de ontwikkeling van een nieuw algoritme voor het be-
palen van de optimale steekproefgrootte in de EOQL-methode. Dit algoritme
wordt in hoofdstuk 6 beschreven. Omdat deze methode gebruik maakt van de
‘exacte’ onderliggende (hypergeometrische) verdeling hebben wij deze metho-
de gedoopt tot de EEOQL-methode (EEOQL= Exact Expected Outgoing Quality
Limit). Het nieuwe algoritme blijkt zeer efficiënt en accuraat te zijn en bovendien
is de methode eenvoudig te programmeren in gangbare software. Tevens wordt
in hoofdstuk 6 een algoritme ontwikkeld dat tabellen genereert voor de verwach-
tte fractie fouten in de populatie na controle voor alle mogelijke combinaties van
steekproef- en populatiegrootte. Het gebruik van deze tabellen vergemakkelijkt
de steekproefopzet van de EEOQL-methode.

Uit een vergelijking, die in hoofdstuk 6 wordt gedaan, tussen de AOQL-
methode, de EOQL-methode en de EEOQL-methode blijkt dat de AOQL-metho-
de optimale steekproefgroottes geeft die te laag zijn. Dit leidt tot een onterecht
gevoel van geruststelling. De kosten van de EOQL-methode zijn in vergelijking
tot de EEOQL-methode te hoog omdat de EOQL-methode optimale steekproef-
groottes genereert die te hoog zijn.

Dit proefschrift rechtvaardigt de bewering dat de EEOQL-methode in de ge-
reedschapskist met controle maatregelen die deel uit maken van de interne be-
heersing niet mag ontbreken.