# Item-by-item sampling for promotional purposes

**Neil H. Spencer & Lindsey Kevan de Lopez**

Statistical Services and Consultancy Unit

University of Hertfordshire Business School Working Paper (2012)

# Item-by-item sampling for promotional purposes

Neil H. Spencer & Lindsey Kevan de Lopez

Statistical Services and Consultancy Unit, Business School, University of Hertfordshire, Hatfield, Hertfordshire, AL10 9AB

E-mail for correspondence: N.H.Spencer@herts.ac.uk

Abstract

In this paper we present a method for sampling items that are checked on a pass/fail basis, with a view to a claim being made about the success/failure rate for the purposes of promoting a company's product/service. Attention is paid to the appropriate use of statistical phrases for the claims and this leads to the development of Bayesian credible intervals. The hypergeometric distribution is used to calculate successive stopping rules so that the costs of sampling can be minimised. Extensions to the sampling procedure are considered so as to allow the potential for stronger and weaker claims to be made as sampling progresses. The relationship between the true error rate and the probabilities of making correct claims is discussed.

## 1. Background

There are many situations where a set of items are examined to see if they are, in some way, incorrect. A typical example could be in a manufacturing context where, say, electronic components are tested to see if they work within required parameters. Each item is tested (perhaps to destruction but not necessarily so) and either "passes" or "fails". Another example might be where a weight loss programme is keeping records of the weights registered by its participants and wishes to examine the figures recorded on their computer database to see if they have been transferred correctly from paper

records. Each weight has either been recorded correctly or it has not. It is the authors' involvement in work relating to this latter example which is at the root of the work described in this paper and hence will be the scenario used henceforward.

The research presented is a development of work undertaken for the company Weight Wins in connection with its "Pounds for Pounds" programme. Part of this sample audit was in connection with the NHS (National Health Service) trial of the programme in east Kent. An analysis of the results of this trial was carried out by Relton et al (2011).

The company running the weight loss programme wishes to be able to demonstrate the accuracy of their recording systems and hence add credibility to analyses of the recorded weights that show that the weight loss programme is being successful. It may be possible to conduct a complete examination of all the weight records and discover the proportion that have been correctly entered on to the database. However when there are many records, a sampling approach is required.

In section 2, we discuss the statistical measure that is to be used to meet the requirements of the company and in section 3 we address the sample size issues for the preferred measure. Similarities between the sampling proposed here and acceptance sampling is discussed in section 4. The issue of prior probabilities is discussed in section 5 and extensions to the sampling procedure are discussed in section 6. The issue of there being multiple stopping points (and hence multiple opportunities to make a false claim) is dealt with in section 7 and conclusions are given in section 8.

## 2. Choice of statistical measure

The company wishes to make a statement about the accuracy of the recording system and, thus, a way of measuring this must be found.

One option is to carry out a formal test of hypotheses along the lines of $H_0$: the true proportion of incorrect weights is 5%; $H_1$: the true proportion of incorrect weights is less than 5%. A sample size calculation can be carried out and the appropriate number of records sampled at random from the population of records. A conclusion that the company would hope to draw could be worded "We have sufficient evidence to claim that the true proportion of incorrect weights is less than 5% when we use the 5% level of significance". A less statistically scrupulous company might say "We have shown that our records are over 95% correct" but for the purposes of this paper (and in our actual experience) we will assume that the company wishes to be statistically sounds in its claims.

The statistically correct form of words for the result of a hypothesis test is not of the form which lends itself to the purposes of promoting a company's procedures. A confidence interval approach offers an alternative. A two-sided interval could be reported as "A 95% confidence interval for the proportion of incorrect weights is from 0.33% to 6.33%" or a one-sided interval could be reported as "A 95% confidence interval has an upper limit of 5.33%". Neither of these statements is much of an improvement on the wording used for the hypothesis test result and an expanded "95% of intervals such as this would contain the true proportion of incorrect weights" is not helpful.

What the company actually wants to say is "There is a 95% probability that the true proportion of incorrect weights is less than 5%". This is equivalent to creating a one-sided 95% Bayesian credible interval. It is this statistical methodology that we employ in this paper.

## 3. Sample size for Bayesian credible interval

We define the size of the population of weight certificates being examined as being $N$ records, the number of incorrect weight records in the population as being $R$, the size of the sample examined as being $n$ records and the number of incorrect records in the sample as being $r$.

For a desired maximum error rate of $R^*/N$ and wishing to make a statement that there is at least a $100\times(1-\alpha)\%$ chance of the true error rate being no more than this, we require $\Pr(R \le R^* | r,n,N) = \sum_{R=0}^{R^*} \Pr(R|r,n,N) \ge (1-\alpha)$.

As $\Pr(r|R,n,N) = \frac{\Pr(R,r,n,N)}{\Pr(R,n,N)}$, we get the following:

$$\Pr(R|r,n,N) = \frac{\Pr(R,r,n,N)}{\Pr(r,n,N)} = \frac{\Pr(r|R,n,N)\Pr(R,n,N)}{\sum_{R'=r}^{N-(n-r)} \Pr(r|R',n,N)\Pr(R',n,N)}.$$

Under an equal prior probabilities assumption (which we examine further below), we have $\Pr(R',n,N) = \Pr(R,n,N) = p^*$ for all $R$, $R'$. This gives

$$\Pr(R|r,n,N) = \frac{\Pr(r|R,n,N)p^*}{p^* \sum_{R'=r}^{N-(n-r)} \Pr(r|R',n,N)} = \frac{\Pr(r|R,n,N)}{\sum_{R'=r}^{N-(n-r)} \Pr(r|R',n,N)}.$$

We are operating under a system governed by the hypergeometric distribution. For this distribution, $\sum_{R'=r}^{N-(n-r)} \Pr(r|R',n,N) = \frac{N+1}{n+1}$, hence $\Pr(R|r,n,N) = \frac{n+1}{N+1}\Pr(r|R,n,N)$.

Thus, $\Pr(R \le R^*|r,n,N) = \sum_{R=0}^{R^*} \frac{n+1}{N+1}\Pr(r|R,n,N) = \frac{n+1}{N+1}\sum_{R=0}^{R^*}\Pr(r|R,n,N) = \frac{n+1}{N+1}\sum_{R=r}^{R^*}\Pr(r|R,n,N)$ as $\Pr(r|R,n,N) = 0$ for $R < r$.

Thus, for $r$ = 0, 1, 2, ..., we want to identify $n_0$, $n_1$, $n_2$, … which are the smallest $n$ such that $\frac{n+1}{N+1}\sum_{R=r}^{R^*}\Pr(r|R,n,N) \ge (1-\alpha)$ or equivalently $\sum_{R=r}^{R^*}\Pr(r|R,n,N) \ge (1-\alpha)\frac{N+1}{n+1}$.

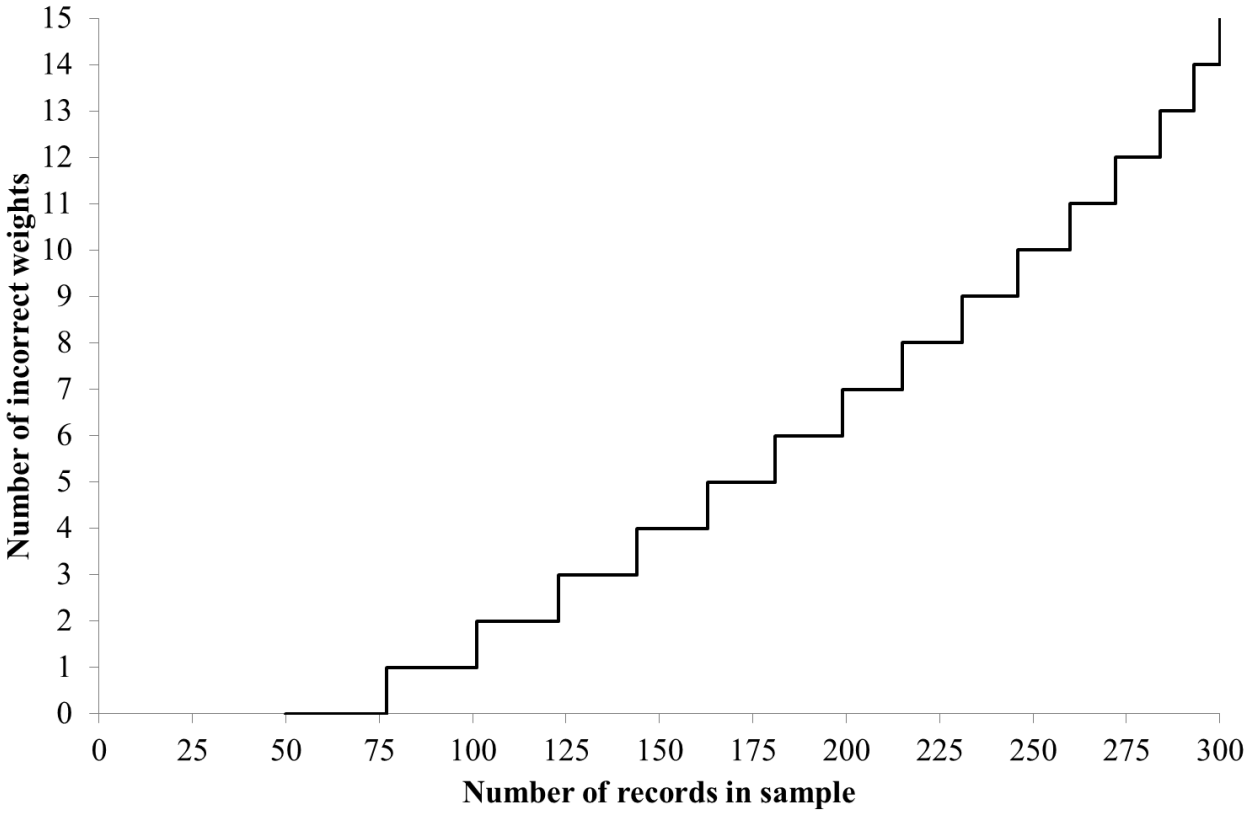
Thus, the required sample size depends on the number of incorrect records that are identified as the random sampling takes place and, as a result, the sampling strategy needs to account for this. From the start of the sampling, if no incorrect records are found by the time $n_0$ records have been inspected, the sampling can stop. If one incorrect record has been found after $n_0$ records have been inspected then the sampling must continue until a total of $n_1$ records have been inspected and, if no further incorrect records have been found, it can stop there. This pattern could continue until all $N$ records in the population have been examined. However, in practice, it is more likely that the stopping strategy would be adapted, as discussed in section 6, so that not all records were examined.

For example, with a company wanting to say "There is a 95% probability that the true proportion of incorrect weights is less than 5%" we have $\alpha = 0.05$ and $R^*/N = 0.05$. If $N = 300$ weight records then $R^* = 15$. Applying the above sample size calculations yields Table 1. It shows the minimum number of weight records that need to be checked and if no incorrect weights are discovered, the sampling can stop and the company make the claim that they wish. If, say, one incorrect weight has been identified in the initial sample of 50, the table shows that the sample size required has increased to 77. If this point is reached with no more incorrect weights being discovered then the sampling can stop and the company can make their claim. The procedure can also be shown graphically in Figure 1. In this figure, if the combination of sample size and number of incorrect weights reaches the line (including the part of the line running along the horizontal axis) then the sampling stops and the company can make their claim.

Table 1: Sample size requirements for varying numbers of incorrect weights in sample (5% claim)

| Number of incorrect weights | Minimum sample size for number of incorrect weights for 5% claim | Number of incorrect weights | Minimum sample size for number of incorrect weights for 5% claim |
|---|---|---|---|
| 0 | 50 | 8 | 215 |
| 1 | 77 | 9 | 231 |
| 2 | 101 | 10 | 246 |
| 3 | 123 | 11 | 260 |
| 4 | 144 | 12 | 272 |
| 5 | 163 | 13 | 284 |
| 6 | 181 | 14 | 293 |
| 7 | 199 | 15 | 300 |

Figure 1: Stopping criteria for sampling scheme (5% claim)




4. Relationship to acceptance sampling

Acceptance sampling is typically used in a manufacturing context when a supplier or customer checks a batch to assess whether the proportion of faculty items is low enough for the batch to be accepted or high enough for the batch to be rejected. There are various forms of acceptance sampling (see, for example, Montgomery, 2009) but the one that most closely mirrors the situation faced here is item-by-item sequential sampling.

In this system for sampling, items from the batch are chosen randomly one at a time and either judged to be faulty or not. If no faulty items are identified then the sampling will stop after a prescribed number of items have been investigated, as is the case described above for the weight records, and the batch will be declared to be of acceptable quality. If one or more faulty items are identified then the

sampling continues. Eventually the number of faulty items relative to the number of items inspected will be small enough or large enough for the procedure to trigger a stopping criteria and the batch accepted or rejected respectively. Alternatively the number of samples inspected will reach a predetermined maximum and a default decision made to accept or reject the batch.

The stopping rules for this type of acceptance sampling come from the Sequential Probability Ratio Test developed by Wald (1947). It is based on the binomial distribution which is a good approximation to the hypergeometric distribution when *N* is large relative to *n*. In the scenario described in this paper, it cannot be assumed that *N* will be large relative to *n,* and thus we do not use Wald's method.

It is also the case that acceptance sampling methods are designed not to provide a probability statement about the proportion of incorrect records, but to decide whether or not a batch of items is of sufficient quality to be accepted. As such, although there are clear parallels between the sampling undertaken in this paper and acceptance sampling, one cannot be a substitute for the other.

## 5. Prior probabilities

In section 3, the calculations shown make an assumption of equal prior probabilities such that $\Pr(R',n,N)=\Pr(R,n,N)=p^*$ for all $R$, $R'$ where $R$, $R'$ are alternative numbers of incorrect weight records in the population, *N* is the size of the population and *n* is the sample size.

At first glance this may appear to be an unrealistic assumption. Surely it is more likely that a relatively small number of weight records will have been entered incorrectly than a large number? However, it is not unreasonable to suppose that there might be some systematic error in the record entry, due perhaps to an incorrect application of a transformation required to convert weight measured in stones and pounds into kilogrammes. Again, if there is more than one person undertaking the entry of the weight records then it is possible for some people to have systematic errors and others not to have such errors. In these circumstances, any value of *R* is now feasible. Even if investigations reveal than only one

person undertook the data entry, it is likely that this would have been done over a period of many days (as the weight records arrive) and thus consistency in the data entry process is hard to guarantee.

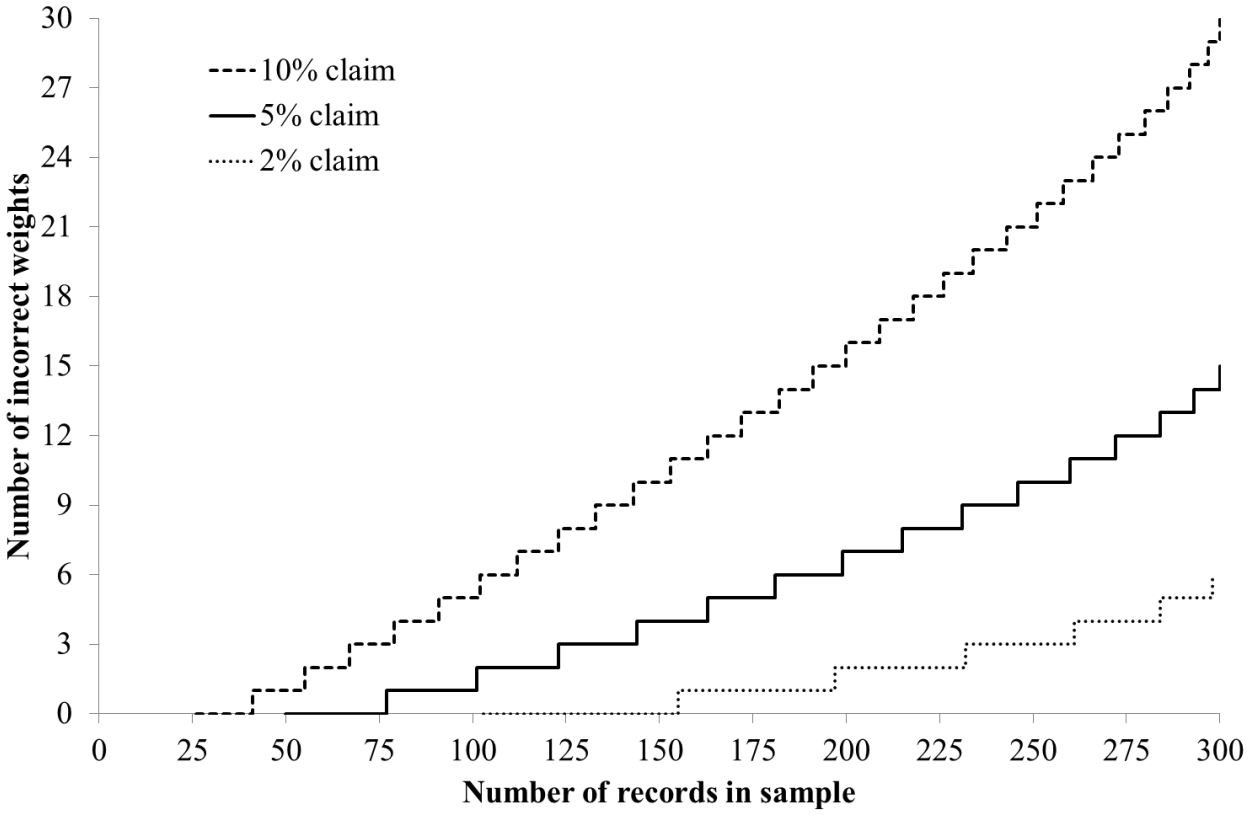
As a result of these issues, it would be difficult to justify any particular assumption being made about the prior probabilities. However, the equal prior probabilities assumption can be seen as adopting a neutral position and is thus appropriate in this work.

## 6. Extension to sampling procedure

While the company might start out with the aim of saying "There is a 95% probability that the true proportion of incorrect weights is less than 5%", clearly they would prefer to be saying "There is a 95% probability that the true proportion of incorrect weights is less than 2%" (or some other smaller figure). Also, if the level of incorrect weights is such that is not possible for them to make the 5% claim, they would nevertheless like (if possible) to make the claim "There is a 95% probability that the true proportion of incorrect weights is less than 10%".

For each of the 2% and 10% claims, tables such as Table 1 can be constructed using the calculations of section 3. Figure 1 can be adapted to include additional lines for each alternative claim and this is shown in Figure 2. Whilst 50 is the minimum sample size needed to be able to make the 5% claim, if this position is reached with no incorrect weights identified, the sampling could continue. If a sample size of 103 is reached with still no incorrect weights then the 2% claim can be made. However, if after proceeding beyond 50 an incorrect weight is discovered before 77 records have been checked then the trace on the graph will go above the 5% claim line and sampling would have to continue until the trace reaches it again. Alternatively, if several incorrect weights are found at a relatively early stage of the sampling, it may be decided that the 10% claim is all that will be aimed at or possibly no claim at all will be made.

Figure 2: Stopping criteria for sampling scheme (10%, 5% and 2% claims)




7. Multiple stopping points

Because we have a series of potential stopping points as the sample size increases, this means that there are multiple opportunities where sampling will stop and a claim about the accuracy of the recording made. This is akin to the multiple testing problem in hypothesis testing when, because there are multiple opportunities for a significant result to be identified, the probability of a Type II error occurring becomes inflated.

In practice, it is likely that before any sampling takes place, a decision will have been made as to the maximum number of records that are to be sampled. For the procedure illustrated in Figure 2, this may well be a maximum of 103 records. This will enable the 2% claim to be made if there are no incorrect weights identified but also allow the 5% claim to be made if there are 1 or 2 incorrect weights. If there

are between 3 and 6 incorrect weights identified, the 10% claim can be made. However, the sampling may stop before 103 records have been investigated. The possible stopping points are outlined in Table 2. The bracketed phrases show repetitions of previous stopping criteria.

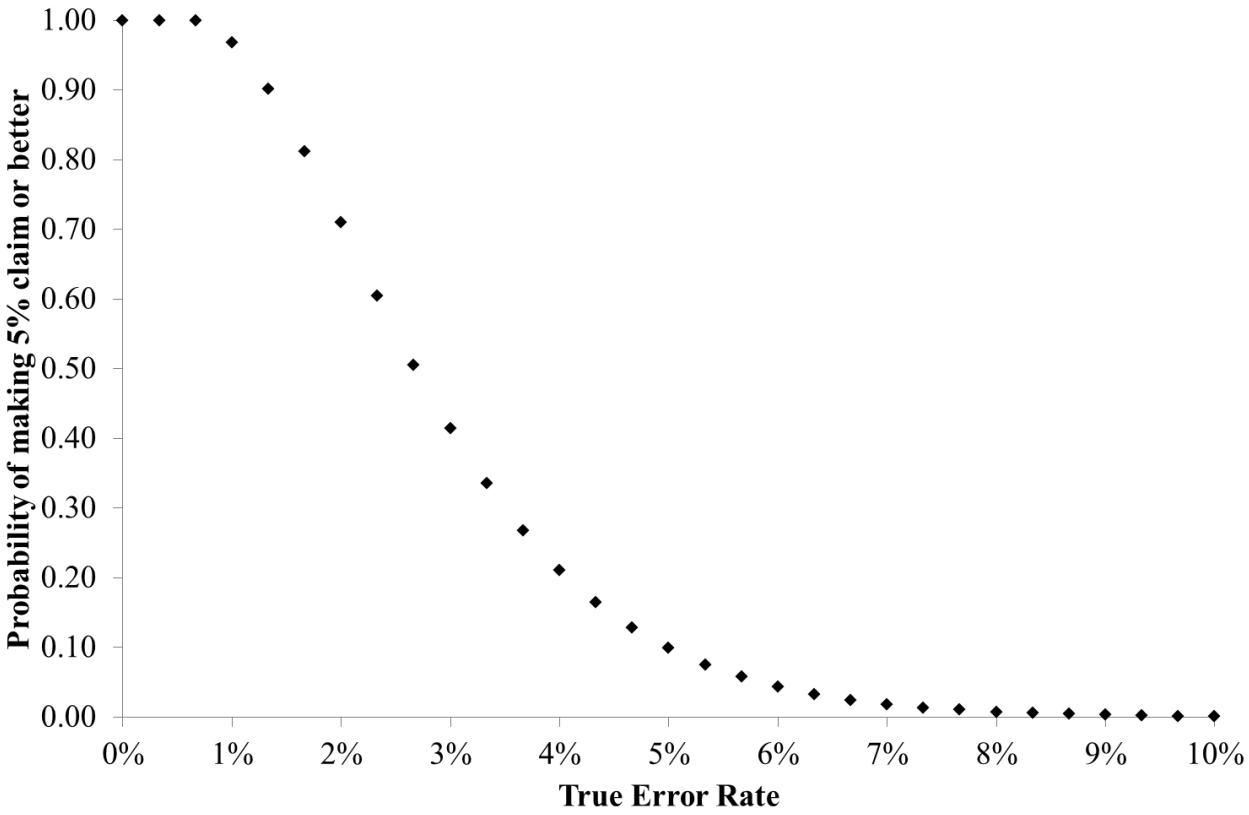
Table 2: Stopping points with maximum sample size of 103

| Sample size | Stop and make 2% claim if: | Stop and make 5% claim if: | Stop and make 10% claim if: | Stop and make no claim if: |
|---|---|---|---|---|
| 0 to 66 | | | | more than 6 incorrect |
| 67 | | | 3 incorrect | |
| 68 to 76 | | | 3 incorrect | |
| 77 | | 1 incorrect | 3 incorrect | |
| 78 | | (1 incorrect) | 3 incorrect | |
| 79 | | (1 incorrect) | 3 or 4 incorrect | |
| 80 to 90 | | (1 incorrect) | (3 or 4 incorrect) | |
| 91 | | (1 incorrect) | 3, 4 or 5 incorrect | |
| 92 to 100 | | (1 incorrect) | (3, 4 or 5 incorrect) | |
| 101 | | 1 or 2 incorrect | (3, 4 or 5 incorrect) | |
| 102 | | 1 incorrect* | 3, 4, 5 or 6 incorrect | |
| 103 | none incorrect | 1 incorrect* | | |

* Would also stop here if 2 were incorrect but if this were the case then would have already stopped at sample size defined in previous line.

The probability of making a claim can be calculated for different possible values of the true error rate, assuming the stopping points in Table 2. This is shown in graphical form in Figure 3 for the chances of making a claim at the 5% level or better. A similar graph could be constructed for the chances of making a claim at the 10% level or better.

Figure 3: Probability of making 5% claim (or better) for different true error rates




The chances of making an incorrect claim for the circumstances described in Table 2 are shown in Figure 3 to be not great. For those true error rates greater than 5%, the probabilities of making a 5% claim are less than 10%. The penalties for incorrectly making a claim in the scenario described are not devastating and mainly consist of "looking stupid" if it subsequently emerges that the claim was incorrect. Indeed, here, the greater risk to the company is that of not making a claim when the truth is

that it could be made. For true error rates near to 5%, the probability of correctly deciding to make a claim are small. This is inevitably the case with these relatively small population and sample sizes.

In other circumstances, the chance of correctly making a claim when the true error rate is less than 5% may be higher. However, at the same time, if the true error rate is above 5%, the probability of incorrectly making a claim will be higher. If this is judged to be a problem then the number of stopping points can be reduced. This can be done by removing the bracketed stopping points in Table 2. Then, for instance, rather than stopping if one incorrect weight is identified at any of the sample sizes 78 to 100 inclusive, the sampling would always continue until sample size 101. This would give the opportunity for more than two incorrect weights to be spotted in the interval 78 to 101 and thus the 5% claim could not be made.

## 8. Conclusions

In this paper we have described a sampling process for companies wishing to make claims about the accuracy of recording weights in a database. There are numerous other situations in which the same methodology could be used, for example from making claims about the acceptability of a manufactured part to the accuracy of automatic word recognition technology. Emphasis has been upon the validity of the wording of the statistical claims made and the defensibility of assumptions made. The issue of multiple stopping points meaning increased chances of incorrect claims has been addressed.

## References

Montgomery, D.C. (2009) *Statistical Quality Control: A Modern Introduction*, 6th edition. John Wiley & Sons, Inc.: Hoboken, New Jersey.

Relton, C., Strong, M. & Li, J. (2011) “The 'Pounds for Pounds' Weight Loss Financial Incentive Scheme: An Evaluation of a Pilot in NHS Eastern and Coastal Kent”. *Journal of Public Health* **33**, 4, pp536-542.

Wald, A. (1947) *Sequential Analysis*. John Wiley & Sons: New York.