RISK & ASSURANCE GROUP

Using Statistics for Precision Assurance Testing: A Worked Example

Release 1.0

January 2017

Eric Priezkalns

Using Statistics for Precision Assurance

This paper is taken from *Revenue Assurance: Expert Opinions for Communications Providers*, a book published by CRC Press and which I helped to write.

1.1. Introduction

Some revenue assurance practitioners know little about statistics. At the same time, when asked to express an opinion on the adequacy of a size of a sample, or whether there is too much risk to rely solely on testing a sample instead of testing the whole population, most revenue assurance practitioners will give an opinion. I want to draw attention to the inconsistency this generates. Statistics is not a matter of opinion. It is a matter of fact. To know the relationship between a sample size and the confidence in the conclusion drawn from testing that sample requires the performance of calculations, not the expression of an 'expert' or 'experienced' opinion. It can be done by a schoolchild if armed with the right textbook and a calculator. Perhaps the saddest aspect of the deficient use of statistics in revenue assurance is that a relatively low level of statistical sophistication could be learned quite easily from standard sources, and would lead to significant improvements in the efficiency of RA practices and certainty with which findings are reached. Given the difficulty of learning about so many other aspects of revenue assurance, mastering basic statistics would be well within the capabilities of most practitioners.

To illustrate, I will present a worked example of how to apply statistics to an assurance challenge where an unusually high degree of precision is required. By so doing, I hope

to show how statistical techniques can and should frame our thinking in the number of tests to perform and the conclusions that can reliably be drawn. Choosing such a demanding example should also illustrate that the more modest expectations of routine RA can easily be satisfied with much smaller samples, begging the question of how the practitioner should justify the cost of investing in systems to check an entire population in real-time.

1.2. The Scenario

Let us suppose I wish to execute a test plan to determine how many of a certain population suffer from error, with the goal being to see if the error rate is more or less than some target ratio. There is no need to test every item in a population in order to reach a reliable inference about that population. It is more efficient to test a suitably selected and representative sample of the right size. This cost efficiency is most obvious when we have no idea what the result will be; if I test a thousand items without finding an error, it is difficult to justify testing a further million items ‘just in case’. Hence sampling is particularly pertinent when first assessing the scale of problems for a population that has not been tested or measured before, and will be useful in deciding how much resources should be put into controlling and monitoring the population – whether it really does need extensive real-time monitoring across the full population, or whether an occasional sample check is sufficient.

1.3. The Challenge

For this challenge, let us suppose we have a revenue share arrangement with a content provider. Every time their content is enjoyed by one of our customers, we owe them a fee, and the size of the fee is based on the duration for which the customer enjoyed the content. The partner wants to know if they are missing revenues and can ask for an annual audit to check they are not. They are interested in errors where there the duration was under-recorded. Under-recording errors mean the content provider received a smaller fee than they should. The two businesses have agreed a contract which says if the number of errors are above a certain level, we must also pay the content provider a penalty fee. The target error ratio is 1 failure per 50,000, so we must pay the fine if the error rate is worse than this. If our own internal sample check shows that there is more than 1 failure per 50,000 then it makes sense to spend money on fixing the issues that cause the errors, in order to avoid paying a fine when the content provider does its audit at the year end. At less than 1 failure per 50,000 the content provider can only claim back the money on specific leakages they find during their audit, with no additional penalty fee.

A failure rate of 1 in 50,000 can be expressed as a probability of 0.00002 (which is the decimal we get when diving 1 by 50,000). In short, we want to devise a test plan that will tell us if the probability of failure is more than 0.00002.

1.4. Confidence

Statistical confidence is as it sounds: a statement about how confident we are that a conclusion drawn is the right conclusion. When using statistics, we ask if a specific condition is met or not; the result is binary, leading to a simple yes/no answer. Hence, when drawing a conclusion, two kinds of mistakes are possible: we conclude a yes when the answer should be have been no (a false positive), or we conclude a no when the answer should have been yes (a false negative). Confidence is a measure of how sure we are that when we conclude 'yes' the answer really is yes, or when we conclude 'no' the answer really is no.

For example, suppose I want to reach a conclusion with at least 90 per cent confidence. This means that after all the tests are performed, the chance that I reached the wrong conclusion will be less than 10 per cent.

I can set the desired confidence level in advance, but when deciding a sample, it is not possible to know what confidence level will actually be achieved in practice. In short, whatever sample size is picked, it may be necessary to increase the sample to get the desired level of confidence, depending on what the test results are in practice. For example, suppose we sample 10 chargeable events and find an insufficient fee was passed to the content provider for all of them. If the results keep coming out the same way, we soon get a high confidence that our target error rate of 0.00002 is exceeded. On the other hand, suppose I test 50,000 events and find 1 error. That equates to exactly 0.00002 probability of failure. If the actual error rate is very close to 0.0002, I would need to do very many tests to get a high level of confidence when comparing my

test results and reaching a conclusion over whether the probability of failure is more or less than 0.00002.

One final thing to note about confidence is that the confidence of not getting a false positive is different to the confidence of not getting a false negative. So when devising my test, I need to be clear about the direction of confidence I am most interested in. In the case of paying contract penalties, this will be costly and have a negative impact on the relationship with the partner, so we want to be sure that our error rate is less than 1 in 50,000. That means we are wary of a false negative – we do not want to conclude that the error rate is below the target when in fact it is above the target. We want high confidence that the rate is less than 1 in 50,000, so when the content provider does its audit, there is little chance of them being upset and demanding a contract penalty.

Had we created a different scenario, we might have wanted a high degree of confidence in the other direction i.e. we would be worried about false positives. For example, suppose a one-off investment in a new order management system becomes cost-effective so long as it eliminates all errors where the current error rate is 1 order goes missing in every 1,000. In that case, we want to be very confident if we decide the error rate is more than 1 in every 1,000, because we do not want to spend money on a system that will not actually pay itself back in terms of added value. If the error rate appears to be over 1 in 1,000, but the confidence is low, it makes more sense to do more tests and see if confidence improves, rather than rushing to make an expensive one-off purchase that may not actually pay for itself in practice.

For the purposes of our challenge, let us suppose there is a very large contract penalty for missing more than 1 transaction in 50,000, so we want to be careful. We set a desired confidence level of 95 per cent, meaning that after we have done our internal tests, there is less than a 5 per cent chance that we decide the error rate is lower than 1 in 50,000 when it is really higher than 1 in 50,000.

1.5. Questions and Answers

So far, we have said we want 95 per cent confidence that errors are less than 1 in 50,000. But we have not talked about how many tests we need to perform yet! The usual next step is to try to come up with an answer to that. That is what we need to do in practice, but bear in mind the comments above about what happens when test results are very close to our target – we need a lot more tests to attain the desired level of confidence. We will not know if we face that problem until we start testing, so in practice we decide the sample size based on a hunch: how many errors we think we will get.

That said, suppose we want to pick the smallest possible sample size that would give 95 per cent confidence (0.05 probability of a mistake) that errors are fewer than 1 in 50,000 (0.00002). In other words, let us suppose that we find no errors at all in our sample (because the more errors we find, the less likely we are to confidently conclude the error rate is less than 1 in 50,000). What is the sample size?

There is a theoretical minimum number of tests that must be performed to fix, at 0.05, the probability of accepting a system which has a proportion of more than 0.00002 in

error. This theoretical minimum is 149,786. So if we were testing in real life, we would not design our test plans with less than this number of tests in mind. If we did plan to do fewer tests, we could never hope to conclude the error rate was less than 0.00002 with 95 per cent confidence. Only doing the minimum number of tests leads to a very high likelihood of rejecting an acceptable system. For example, a system which in reality has an error rate of only 0.00001 would be rejected as 'not good enough' 78 per cent of the time.

This probability, the probability of rejecting an acceptable system, can be controlled. To do so requires specification of (a) the desired level of the probability of incorrectly rejecting an acceptable system, and (b) the proportion of chargeable events that fail to be processed in this acceptable system.

If the probability of rejecting an acceptable system is controlled at a reasonable level, then a sample size much larger than the 149,786 given above will be needed. That is, the 149,786 should be taken very much as a minimal value, and in practice one would expect to use larger values. Only if we expected the test results to be near perfect, i.e. we never find an error in practice, would we execute a test plan based on the minimum number of tests.

The calculations on which the above figures are based assume that the chargeable events chosen for testing are selected at random from the population of events generated over a given period of time.

1.6. Calculations

Let E_m be the measured duration of the event, and E_t the true duration. There is a deemed to be an error in the fees if

(1) $P(E_m < E_t - \delta) < 0.00002$

where

- P(A) means the proportion of chargeable events in the entire population that satisfy condition A
- δ is a tolerance relating to the accuracy of measurement i.e. if a tolerance of 2 seconds is set, the recorded duration is not considered in error if it is 1 second less than the actual duration.

Here, the term 'population' refers to the population of events over a given period of time. The population may be homogeneous or heterogeneous, but the calculations refer to the entire population. Throughout, we assume that the sample of events chosen for testing are selected at random from the entire population of events. This example is solely concerned with the event's duration, though the same statistical principles and arguments apply to any property of events where the aim is to determine whether a certain proportion of events lie in a given interval.

Let us say that $P(E_m < E_t - \delta) = \theta$. Then (1) requires $\theta < 0.00002$. To test this, we will take a sample of *n* events, and estimate θ from θ', the proportion of the *n* sampled events which have $E_m < E_t - \delta$.

Since the aim of this testing is to protect the relationship with a contractual partner, we need to be confident that we are right if and when we assert that $\theta < 0.00002$. When determining a confidence level it seems clear that when, in fact, $\theta \geq 0.00002$ we wish to have a low probability of wrongly concluding that $\theta < 0.00002$. We will aim to restrict to probability to 0.05 or less.

Unfortunately, these conditions alone are not sufficient to determine an appropriate minimum sample size. This is easily seen by considering a scenario in which the true θ is infinitesimally larger than 0.00002. Now, no matter how large a sample of calls is taken, almost 50 per cent of such samples will have the proportion θ' less than 0.00002. That is, almost half the time when we observe a proportion less than 0.00002 and conclude that the system is acceptable, it will not be. No matter how large a sample size we take, we can never reduce to less than 5 per cent the proportion of such samples which have θ' less than 0.00002. In order to define the problem sufficiently precisely that a minimum necessary sample size can be determined, some additional constraint must be applied.

To overcome the difficulty illustrated in the preceding paragraph, we need to fix a threshold, less than 0.00002. Call this threshold *f*. Now, for a given sample size, an elementary calculation can be made to determine the probability of observing $\theta' \leq f$ when $\theta = 0.00002$. (Such calculations are illustrated below.) Call this probability β.

Then, if we observe a value of $\theta' \le f$ either $\theta \ge 0.00002$ and an event of probability less than or equal to β has occurred, or $\theta < 0.00002$. In particular if $\beta = 0.05$ then we will only conclude that $\theta < 0.00002$ when in fact $\theta \ge 0.00002$ less than 5 per cent of the time. This is what is required.

For fixed threshold *f*, the sample size *n* and the value of β are inversely related, so that we can choose *n* to make $\beta = 0.05$.

The distribution of θ' is *binomial* (n,θ) and a standard approach to the calculations in such problems is to use a normal approximation. This yields the probability of observing $\theta' \le f$ when $\theta = 0.00002$ with a sample size of *n* to be

$$\Phi\left(\frac{f - 0.00002}{\sqrt{0.00002 \times 0.99998 / n}}\right)$$

We will set this to 0.05, so any pair of values (n,f) which satisfy

$$\Phi\left(\frac{(f - 0.00002)\sqrt{n}}{0.004472}\right) = 0.05$$

will do. The 0.05 point of a standard normal distribution is -1.6449, so we require the (n,f) pair to satisfy

$$(2) \qquad n = \frac{0.00005411}{(f - 0.00002)^2}$$

The following table gives (n,f) pairs satisfying the 95 per cent confidence level requirements, assuming normal approximations.

f	n
0.000005	240,489
0.000010	541,100
0.000015	2,164,400

However, with such small values for θ we need to be wary of the adequacy of the normal approximation. This is partly because the values of θ' are bounded below by 0. With modern computers, such approximations are unnecessary, since exact binomial calculations can be performed. The calculations that follow use the Splus binom function.

The minimum sample size will be attained by adopting f = 0. This can be seen from the normal approximation in (2) since $f \in [0, 0.00002]$ but also follows from the fact that the f corresponding to the 5 per cent quantile of the binomial distribution will increase as n increases. If f is set to 0, then using the exact binomial calculations, a sample size of 149,786 is required to give $\beta = 0.05$. In other words, with a sample of 149,786 events, if we observe no errors in the sample then the probability of concluding $\theta < 0.00002$ when in fact $\theta \geq 0.00002$ is only 0.05.

Although the minimum sample size is given by adopting f = 0 this is not a good idea because it leads to a very low power; there is a high probability of concluding the error rate exceeds the required 0.00002 when in fact it is lower. Suppose that $\theta = 0.00001$, which is half the error rate needed to trigger a contractual penalty. If tested using a minimum sample size, then the probability of wrongly concluding $\theta \geq 0.00002$ is given by the probability of observing one or more errors in the sample, with a binomial distribution of 149.786 and parameter $\theta = 0.00001$. This probability is 0.7764. That means there is a greater than 77 per cent chance of wrongly concluding that the required error rate is exceeded if using the 'minimum' sample size when the error rate is in fact half of the stipulated limit.

For another example using the minimum sample size, the probability of incorrectly concluding that $\theta \geq 0.00002$ (= $2 \times 10\text{-}5$) when in fact $\theta = 4.6 \times 10\text{-}6$ (a much lower error rate) is still 0.5.

Although f = 0 gives the minimum sample size required to control the probability of incorrectly deciding that the system is acceptable when it is not, this still leaves a high risk of rejecting a system that is well within the required performance limits. To decrease this risk, we need to increase the sample size and increase f. For any given $f \in [0, 0.00002]$, binomial calculations like those above can be made to determine the sample size which will control the probability of accepting an unacceptable system at the required 5 per cent. For example if f = 0.00001, a sample size of 387,670 is required to give $\beta = 0.05$. This means we will conclude that the system is within the required

accuracy bound only if we observe less than $0.00001 \times 387670 \approx 4$ errors in the sample. With this sample size, the probability of rejecting a system where $\theta = 0.00001$ is 0.5, lower than the 0.78 given above.

1.7. The decision threshold

We need some rationale for our choice of f. The probability of incorrectly rejecting an acceptable system, α, should be part of this. For example, we might want to set a limit of $\alpha = 0.2$. Specifying the probability of incorrectly rejecting a perfectly acceptable system is not sufficient to determine f and n. We also need to specify the performance of the 'acceptable' system for which the rejection probability is being controlled at 0.2.

If we believe our system has an error rate of p1 (less than 0.00002), then we can choose a unique value of f such that the following conditions are all satisfied:

(a) $p1 < f < 0.00002$

(b) $P(\theta' < f \mid \theta = 0.00002) \leq \beta$ (set to 0.05 above)

(c) $P(\theta' > f \mid \theta = p1) \leq \alpha$

This requires specification of α as well as of p1. Denoting the cumulative binomial distribution with sample size n and parameter p by pbinom(f; n,p), the relationships

(3) pbinom(f; n, p1) = 1 - α

and

(4) pbinom(f; n, 0.00002) = β

each provide equations that relate f and n. Solving these simultaneous equations gives values for f and n.

About

2.1. About this document

The Risk & Assurance Group has made this document available for the following purpose:

- For RAG members to review the document to determine if it should be approved and recommended for use by other RAG members only.

2.2. Document History

Version Number	Date Modified	Modified by:	Description of changes
1.0	07/01/2017	Eric Priezkalns	First issue for review by RAG members

2.3. About RAG

The Risk & Assurance Group is a not-for-profit limited company incorporated in England. Its purpose is to provide the services of a professional association to risk and assurance professionals working in the communications industry and other sectors.