Sampling Schemes for Validating Service Level Agreements

Tanja Zseby*, Sebastian Zander+
Centre for Advanced Internet Architectures. Technical Report 040706A
zseby@fokus.fraunhofer.de, szander@swin.edu.au

***Abstract*-Service Level Agreements (SLAs) specify the network Quality of Service (QoS) that providers are expected to deliver. Providers have to verify if the actual quality of the network traffic complies with the SLA. Ideally this should be done without introducing significant additional network load and the operational costs should be small. In this paper we propose a novel approach for the passive validation of SLAs based on direct samples of the customer traffic. The SLA contains pre-defined thresholds for QoS metrics, the maximum violation proportion and accuracy boundaries. We model the validation problem as proportion estimation of non-conformant traffic. Then we compare different sampling schemes according to their sampling errors and present a novel solution for estimating the error prior to the sampling. Finally we derive a solution for finding the optimum sample rate based on the SLA parameters.**

Keywords- Non-intrusive Network Measurements, Sampling, SLA Validation, Multiplayer Online Gaming

I. INTRODUCTION

QoS provisioning always brings about the need for QoS auditing to validate the SLA. We consider highly interactive applications as the most likely driver for QoS provisioning and QoS auditing. Among those multiplayer online games have become very popular in the Internet and the increasing number of users makes them important from a business perspective [7]. Especially first person shooters (FPSs) like the well known Quake or Half-Life are QoS-sensitive [1]. Not only does the QoS affect all the players but also differences between players can lead to unfairness.

For SLA validation it is necessary to measure the real customer traffic, as active measurements based on test traffic may not reflect the true performance observed by the user. Passive (non-intrusive) measurements allow the direct measurement of the customer traffic [1]. They also have the advantage that no additional traffic is inserted into the network. The problem is that in high-speed networks it is impossible to observe and measure each packet because of the limited resources of the network elements (routers). Even with dedicated measurement boxes it is hard and costly to keep up with the number of packets transmitted.

In this paper we propose a novel solution for passive SLA validation based on direct samples of the customer traffic. First we formulate a statistical SLA, which defines a performance metric threshold, the allowed proportion of violators and accuracy boundaries for an estimation-based validation. Then we compare different sampling schemes according to their sampling errors. Unfortunately it is not possible to exactly compute the sampling error without knowing the real violator proportion. Therefore we discuss different approaches to estimate the sampling error prior to the sampling process. We present a novel solution that slightly overestimates the sample rate but always guaranties that the accuracy boundaries in the SLA are met. Furthermore it uses a fixed sample rate so that no adaptations during runtime are necessary. Finally we derive a solution for computing the minimum sampling rate needed based on the SLA parameters.

The rest of the paper is structured as follows. Section II proposes a novel type of SLA. Section III discusses the different sampling techniques and derives a mathematical model for the SLA validation. Section IV concludes and outlines future work.

II. SERVICE LEVEL AGREEMENTS

Most applications do not require exact values for quality parameters. It is rather more important that the parameters are within a specific range and do not exceed pre-defined thresholds. Furthermore, it is often tolerable to have some outliers (e.g. a few packets that exceed the threshold), as long as there are not too many.

We propose a statistical SLA that is based on an estimation of the quality metric, instead of exact measurement. It is important to include the expected estimation accuracy in the SLA to allow customers to assess the estimation. Therefore error boundaries and confidence levels for the estimation are part of the SLA.

A. *Proportion vs. Percentile Estimation*

The goal is to validate if the packets in a flow are conformant to the delay guarantees given in an SLA. An estimation of the whole delay distribution is difficult and contains much more information than needed. The estimation of mean and standard deviation of the delay values gives first insights about the quality situation, but is inadequate to validate the SLA conformance.

The percentiles of the delay distribution reveal the delay value below which we can assume the majority (e.g. 95%) of the observed packet delays [4]. It provides a valuable parameter for assessing the general network

* Fraunhofer FOKUS, Berlin, Germany
+ Swinburne University of Technology, Melbourne, Australia

situation but is unsuitable to quantify non-conformance. If the percentile lies above the defined threshold, the approach does not provide information about what percentage of packets really violated the contract.

Instead of estimating percentiles, we propose to estimate the percentage of packets above the given threshold. Then we can formulate the SLA validation as estimation of the proportion of packets that exceed the delay limit d_{max}. A packet with delay $d>d_{max}$ is considered a violator (hit,1), packets with delay $d \leq d_{max}$ are considered conformant (no-hit, 0).

B.*User-friendly Confidence Limits*

For the SLA validation it is required to check whether the real proportion of violators in the measurement interval stays below the proportion threshold:

$$P \overset{?}{\leq} P_{thres} \tag{1}$$

The proportion of violators m in the sample is an unbiased, consistent and efficient estimator for the proportion of violators in the parent population:

$$\hat{P} = p = \frac{m}{n} \tag{2}$$

With (2) we just get an estimate, which usually differs from the real value. We also have to consider the estimation error that can occur. It is possible to define error bounds with a certain probability.

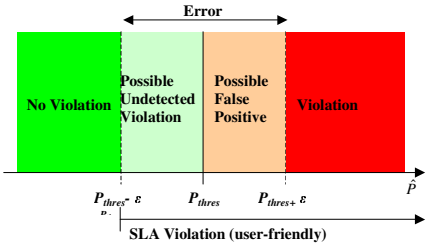

Figure 1: Violation Decision

From a users viewpoint it is more important to not underestimate the violator proportion, than to prevent overestimation. Therefore we always assume an SLA violation if the estimate is larger than P_{thres}-ε, regardless of the fact that the true value could be below P_{thres} (see Figure 1). The probability that the real violator proportion is lower than the estimate plus an estimation error can be expressed in a single-sided confidence interval (see III.F).

We call this approach the "user friendly" approach because the user can be sure that all SLA violations are detected with the given confidence level. However some SLA violations detected will not be real SLA violations (false positives). Contrary the "provider friendly" approach would assume an SLA violation occurred only if the estimated proportion is greater than P_{thres}+ε. This eliminates false positives but would leave some SLA violations undetected. A neutral approach would be to assume an SLA violation if the estimated proportion is larger than P_{thres}. In the rest of the paper we follow the "user friendly" approach.

The process of identifying SLA violations is equal to a statistical hypothesis test whether the estimated proportion is significantly different from P_{thres}. For the user friendly approach the question is: is the estimated proportion significantly below the threshold? The corresponding null and alternative hypotheses are:

$$H_0 : \hat{P} - P = 0, H_A : \hat{P} - P < 0 \tag{3}$$

For the "provider friendly" approach the question is: is the estimated proportion significantly above the threshold? The corresponding null and alternative hypotheses are:

$$H_0 : \hat{P} - P = 0, H_A : \hat{P} - P > 0 \tag{4}$$

C.*Time-based vs. Count-based Intervals*

For SLA validation only the traffic for which a guarantee has been given is of interest and should form the basis for the measurement. If we define time-based measurement intervals, we do not know how much traffic is observed in the time intervals. The number of packets from the customer can vary from zero to the maximum amount for the given link rate. Therefore we here use count-based measurement intervals and define the measurement interval as number of packets *N*.

D.*SLA Specification*

Here we describe the SLA parameters (similar specifications can be done for other performance metrics like loss or jitter):

- Delay Threshold (d_{max}): the maximum delay value which still leads to a good user experience
- Proportion Threshold (P_{thres}): the maximum acceptable percentage of packets violating the delay threshold (proportion of the overall traffic)
- Error (ε): the (absolute or relative) estimation error
- Confidence Level: the probability that the real proportion is within the confidence limits given by the estimate and error bounds

III. SAMPLING SCHEMES

In this section we first introduce and compare different sampling techniques with respect to the sampling error. Then we develop the mathematical model for statistical SLAs. Finally we derive equations for determining the minimum sample rate necessary to achieve the desired accuracy.

A.*Initial Assumptions and Notation*

Assumption 1: The sample size *n* is large enough to assume the estimate is asymptotically normal distributed.

$$n \geq \frac{9}{P \cdot (1-P)} \tag{5}$$

Assumption 2: The population size N is large.

$$N - 1 \approx N \tag{6}$$

Throughout this document we use the terms and notation introduced in Table 1.

Table 1: Notation

	Population	Sample	Estimate
Number of packets	N	n	$\hat{N}$
Number of violators	M	m	$\hat{M}$
Proportion of violators	$P = \frac{M}{N}$	$p = \frac{m}{n}$	$\hat{P}$

B.*n-out-of-N Sampling*

With n-out-of-N sampling we select exactly *n* packets out of the population of *N* packets observed in the measurement interval. We can estimate the number of violators in the measurement interval from the number of violators in the sample as follows:

$$\hat{M} = \frac{N}{n} \cdot m = \frac{N}{n} \cdot \sum_{i=0}^{n} x_i \qquad \text{with} \quad x_i \sim Be(P) \quad (7)$$

The random variable x_i denotes the conformance of the sampled packets to the SLA ($x_i = 0$ if packet delay $d \le d_{max}$ and $x_i = 1$ if $d > d_{max}$). x_i can be modelled as Bernoulli distributed random variable (RV) with probability of success *prob=M/N*.

The number of violators *m* in the sample can be modelled as number of hits in an experiment with *n* trials. Since we cannot select a packet again, we have to consider a selection without replacement, i.e. *m* can be considered as RV with a hyper geometric distribution. The proportion *P* of violators in the measurement interval is estimated by the proportion *p* of violators in the sample. Since *n* is constant for n-out-of-N sampling we get the following standard error (see appendix IV.B):

$$\sigma_{\hat{P}} = \sqrt{\frac{P \cdot (1-P)}{n}} \cdot \sqrt{1 - \frac{n}{N}} \quad (8)$$

The estimation accuracy depends on the sample fraction and on the real violator proportion. If the sample fraction is small (<5%), we can neglect the finite population correction (last factor). The real violator proportion is unknown and has to be approximated from the sample or replaced by worst-case parameters in order to make an accuracy prediction in advance. We compare different methods to estimate the standard error in section III.F and show that for our user-friendly scenario using the proportion threshold P_{thres} for the standard error estimation provides an elegant alternative.

C.*Probabilistic Sampling*

With probabilistic sampling each packet is selected with a given probability regardless of the fact how many packets have been already selected before. Therefore the real sample size varies for each run, and in most cases will not be equal to the target sample size *n*. Because the real sample size approaches *n* for large parent populations it is expected that this effect gets smaller the longer the measurement interval is.

An approach for modeling probabilistic packet sampling for volume estimation is shown in [6]. In this approach the variability of the sample size is neglected. We later check with empirical tests whether this can be justified in practice. In [6] the number of packets that belong to a specific flow is estimated by modelling the selection process with a Bernoulli distributed random variable ω_i with success probability $f=n/N$. ω_i becomes 1 if the packet is selected and 0 if the packet is not selected. If we consider the packet property "violate SLA" instead of "belong to flow f", we can apply the same model and get the following estimate for the number of violators[1].

$$\hat{M} = \frac{N}{n} \cdot m = \frac{N}{n} \cdot \sum_{i=0}^{M} \omega_i \qquad \text{with} \quad \omega_i \sim Be\left(\frac{n}{N}\right) \quad (9)$$

By neglecting the variability of *n*, we get an unbiased estimate and the standard error is (see appendix IV.C):

$$\sigma_{\hat{P}} = \sqrt{\frac{P}{n}} \cdot \sqrt{1 - \frac{n}{N}} \quad (10)$$

We observe that the standard error for n-out-of-N sampling is equal to the standard error for probabilistic sampling multiplied by a factor $\sqrt{(1-P)}$.

Since $0 \le \sqrt{(1-P)} \le 1$, we can deduce that n-out-of-N sampling provides a smaller standard error and therefore a better accuracy than probabilistic sampling. Nevertheless, the difference depends on the violator proportions in the measurement interval and can become very small in case the violator proportion is low.

D.*Systematic Sampling*

If all packet delays in the measurement interval were independent, we could apply the same mathematical model as for n-out-of-N sampling. But if correlations occur, the systematic selection process can interfere with periodicities in the packet sequence. In such cases we may get a non-representative accumulation of packets with specific properties (e.g. packets with high delays) in the sample and therefore a biased estimation. The nature of this bias heavily depends on the specific traffic mix. Therefore we cannot derive a generic model for the accuracy as we did for the random selection methods.

E.*Hash-based Sampling*

Hash-based packet selection is proposed in [5]. A deterministic function on the packet content is used to calculate a hash value. If the hash value falls in a specific range the packet is sampled. With this a random probabilistic sampling is emulated. In [5] it is shown that sufficient randomness (e.g. regarding source and destination) can be achieved if parts of the packet content are included in the hash calculation. Further investigations are needed to proof this also for other attributes, but if one can achieve sufficient randomness to apply a statistical model, the model for probabilistic sampling can be applied.

F.*Statistical SLAs*

Confidence Limits and Standard Error

In accordance to the "user-friendly" approach we

[1] Please note that we use a different notation than introduced in [6] to be consistent with the notation throughout this document.

define a single sided confidence interval (CI). If n is large enough to fulfill condition (5), the single-sided CI can be formulated as follows[2]:

$$\Pr\left(P \le \hat{P} + z_{1-\alpha}\cdot\sigma_{\hat{P}} + \frac{0.5}{n}\right) = 1-\alpha \qquad (11)$$

The single sided CI states that with a probability of $1-\alpha$ the real proportion is below the upper confidence limit. Based on (11) we can define the absolute error ε:

$$\varepsilon = \hat{P} - P \ge -z_{1-\alpha}\cdot\sigma_{\hat{P}} - \frac{0.5}{n} \qquad (12)$$

With the user-friendly approach a violation is assumed, if the estimated violator proportion plus the potential estimation error is above the desired threshold.

$$\hat{P} + \varepsilon \ge P_{thres} \qquad (13)$$

Required Sample Size

From (8) and (12) we can derive the minimum needed sample size to achieve the accuracy level given by ε and the confidence level $1-\alpha$. Because the solution is complex (see appendix IV.D) we do an approximation using assumption (6) and assuming that ε is small ($\le$0.01):

$$n \ge \frac{\varepsilon + z_{1-\alpha}^2\cdot P\cdot(1-P)}{\varepsilon^2 + \frac{z_{1-\alpha}^2\cdot P\cdot(1-P)}{N}} \qquad (14)$$

The approximation slightly overestimates n (depending on ε and N).

Approximation of Standard Error

Equations (8) and (10) show that the estimation accuracy depends on the sampling rate n and on the real proportion P. P usually varies over different intervals. In order to maintain a given accuracy level it is necessary to adapt n based on P. But the real proportion P in the parent population is never known. Nevertheless, there are different ways to approximate the standard error.

1. Maximum Value

We can approximate the variance by using 0.25, the maximum value for $P(1-P)$:

$$\sigma_{\hat{P}} \approx \sqrt{\frac{0.25}{n}}\cdot\sqrt{1-\frac{n}{N}} = \frac{1}{2\cdot\sqrt{n}}\cdot\sqrt{1-\frac{n}{N}} \qquad (15)$$

This approximation is independent of any real value and can be used before the sampling process to estimate the required sample size. Nevertheless, this maximum value may be much higher than the real standard error, leading to an unnecessary high sample size.

2. Estimation from actual sample

Estimating the violator proportion P by the proportion p in the sample provides a more accurate approximation:

$$\sigma_{\hat{P}} \approx \sqrt{\frac{p\cdot(1-p)}{n}}\cdot\sqrt{1-\frac{n}{N}} \qquad (16)$$

With this, the achieved accuracy can only be computed after the sampling process, because only then the delay values of the sampled packets are known, which are required to calculate p.

3. Estimation from previous sample

P can also be estimated by using sample values from previous measurement intervals and a prediction function.

$$\sigma_{\hat{P}} \approx \sqrt{\frac{p'_t\cdot(1-p'_t)}{n}}\cdot\sqrt{1-\frac{n}{N}} \text{ with } p'_t = f(p_{t-1}) \qquad (17)$$

With this there are two errors, an estimation error, when estimating P_{t-1} by p_{t-1} from the sample and a prediction error when p_t is predicted from p_{t-1}.

A similar problem is addressed in [3] for measuring traffic load by using an autoregressive (AR) model. Although the approach is quite valuable, one has to consider the prediction and the estimation error and the adaptation of the sample size requires a dynamic control of the measurement configuration.

4. Approximation with P_{thres}

The maximum value approach clearly leads to a much higher sampling rate than needed while for the estimation approach the error is not known anymore. Figure 2 illustrates the later for both absolute and relative errors (the relative error is the absolute error divided by the real proportion P). The figure shows the minimum required sampling size to maintain an absolute error ε=1% and a relative error of 33% over the real proportion P (N=50,000). The numbers have been chosen such that $P_{thres}\cdot\varepsilon_{rel}=\varepsilon_{abs}$ for P_{thres}=3%.

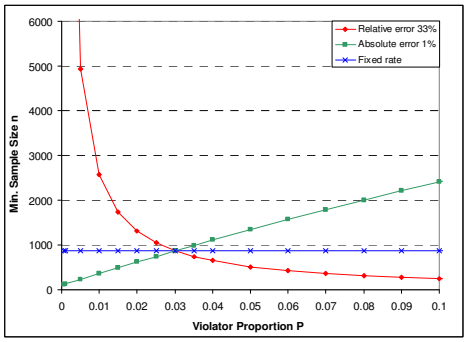

Figure 2: Approximation with P_{thres}

When P is estimated by a previous measurement the following problem occurs. If the relative error is used and the estimated P was large but P is small in the current interval the sampled rate will be too low. If the absolute error is used, the estimated P was small but the current P is large the sample rate will be too low. In both cases the resulting error is not known anymore.

Instead of using an adaptive sampling rate based on the estimated P we propose to choose the fixed sampling rate that is only based on P_{thres}. The intuitive reason is that at the point P_{thres} we must know the exact error to detect possible SLA validations. For higher proportions we do not care that the absolute error is larger and for lower proportions we do not care that the relative error is larger. With respect to Figure 2 this means it is sure that

[2] The upper level of the confidence interval is increased by $.5/n$ for continuity correction.

for estimated proportions below or equal 2% the SLA has not been violated while estimated proportions above 3% indicate SLA violations.

We show that our approach is correct using proof by contradiction. First we assume that our approach may fail to detect a violation if

$$\hat{P} + \varepsilon_T < \hat{P} + \varepsilon_R \quad (18)$$

where ε_T is calculated using P_{thres} and ε_R is calculated using the real P. A violation occurs if $P > P_{thres}$ and therefore:

$$P = a + P_{thres} \quad \text{where} \quad 0 \le a \le 1 \quad (19)$$

Under assumption (19) equation (18) is always false meaning that our approach detects all violations that would be detected with the real standard error. The detailed proof is given in appendix IV.E. The approximation with P_{thres} leads to higher sampling rates but assures that the statistical guarantees are always met. Furthermore it has the advantage that no further processing is required during runtime.

IV. Conclusion

In this paper we describe a novel approach for SLA validation based on direct sampling of the customer traffic. We model the validation problem as proportion estimation of non-conformant traffic. The passive approach ensures that a statement about the real customer traffic is made, which provides a significant advantage over active measurements that can only provide general and possibly biased quality statements.

We compared different sampling schemes according to their sampling errors and found that n-out-of-N has the smallest error but for small proportions probabilistic sampling is close. We argue that the effectiveness of systematic and content-based sampling depends on the traffic under investigation i.e. the results may be biased depending on the traffic. Assuming there is no bias we show that systematic sampling can be compared with n-out-of-N sampling while content-based sampling can be compared to probabilistic sampling. We also present an elegant solution for estimating the sampling error prior to the sampling and for computing the minimum sample rate required depending on SLA parameters.

We plan to empirically compare the different sample schemes and to evaluate the approach using real gaming traffic. We also plan to extend our approach to estimate multiple thresholds at the same time.

Acknowledgments

Thanks to Grenville Armitage for helping us improve this paper.

References

[1] G. Armitage. An Experimental Estimation of Latency Sensitivity In Multiplayer Quake 3. ICON 2003, Sydney, Australia, September 2003

[2] N. Brownlee. Traffic Flow Measurement: Experiences with NeTraMet. RFC2123, March 1997

[3] B.-Y. Choi, J. Park, Z.-L. Zhang. Adaptive Random Sampling for Load Change Detection. Poster at ACM SIGMETRICS 2002, Marina Del Rey, California, USA, June 15-19, 2002

[4] B.-Y. Choi, S. Moon, R. Cruz, Z.-L. Zhang, C.Diot. Practical Delay Measurements for ISPs. SPRINT ATL research report RR03-ATL-051910, May 2003

[5] N. Duffield, M. Grossglauser. Trajectory Sampling for Direct Traffic Observation. Proceedings of ACM SIGCOMM 2000, Stockholm, Sweden, August 28 - September 1, 2000.

[6] N. Duffield, C. Lund, and M. Thorup: Properties and Prediction of Flow Statistics from Sampled Packet Streams, ACM SIGCOMM Internet Measurement Workshop 2002, Marseille, France, November 6-8, 2002

[7] http://www.isp-planet.com/news/2002/gamez_021202_p2.html. December 2002

[8] T. Zseby, S. Zander, G. Carle. Evaluation of Building Blocks for Passive One-way-delay Measurements. Passive and Active Measurement Workshop (PAM 2001), Amsterdam, Netherlands, April 23-24, 2001

APPENDIX

B. *Derivation of Expectation and Variance for n-out-of-N Sampling*

Expectation and variance of hyper geometric random variable m:

$$E[m] = n \cdot P \qquad V[m] = \frac{N-n}{N-1} \cdot n \cdot P \cdot (1-P)$$

Expectation and variance of estimate $\hat{P}$:

$$E\left[\hat{P}\right] = E\left[\frac{m}{n}\right] = \frac{1}{n} \cdot E[m] = \frac{1}{n} \cdot n \cdot P = P \rightarrow \text{unbiased}$$

$$V\left[\hat{P}\right] = V\left[\frac{m}{n}\right] = \frac{1}{n^2} \cdot V[m] = \frac{P \cdot (1-P)}{n} \cdot \frac{N-n}{N-1}$$

With $\frac{N-n}{N-1} \approx \frac{N-n}{N} = 1 - \frac{n}{N}$ the standard error can be derived as follows:

$$\sigma_{\hat{P}} = \sqrt{\frac{P \cdot (1-P)}{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

C. *Derivation of Expectation and Variance for Probabilistic Sampling*

Estimate for the number of violators:

$$\hat{M} = \frac{N}{n} \cdot m = \frac{N}{n} \cdot \sum_{i=0}^{M} \omega_i$$

Expectation and variance calculated in accordance to [6] (neglecting the variability of n):

$$E\left[\hat{M}\right] = E\left[\frac{N}{n} \cdot m\right] = \frac{N}{n} \cdot \sum_{i=1}^{M} E[\omega_i] = \frac{N}{n} \cdot M \cdot E[\omega_i] = \frac{N}{n} \cdot M \cdot \frac{n}{N} = M$$

$\rightarrow$ unbiased

$$V\left[\hat{M}\right] = V\left[\frac{N}{n} \cdot m\right] = \frac{N^2}{n^2} \cdot \sum_{i=1}^{M} V[\omega_i] = \frac{N^2}{n^2} \cdot M \cdot V[\omega_i] =$$

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

Derivation of expectation and variance for estimate $\hat{P}$:

$$E\left[\hat{P}\right] = E\left[\frac{\hat{M}}{N}\right] = \frac{1}{N} \cdot E\left[\hat{M}\right] = \frac{M}{N} = P$$

$$V\left[\hat{P}\right] = V\left[\frac{\hat{M}}{N}\right] = \frac{1}{N^2}\cdot V\left[\hat{M}\right] = \frac{M}{N^2}\cdot\left(\frac{N}{n}-1\right) = \frac{P}{N}\cdot\left(\frac{N}{n}-1\right) =$$
$$P\cdot\left(\frac{1}{n}-\frac{1}{N}\right) = P\cdot\left(\frac{N-n}{n\cdot N}\right) = \frac{P}{n}\cdot\left(1-\frac{n}{N}\right)$$

Derivation of standard error:

$$\sigma_{\hat{P}} = \sqrt{\frac{P}{n}}\cdot\sqrt{1-\frac{n}{N}}$$

D.*Derivation of the Minimum Sampling Rate*

From (12) and assumption 2 we can derive the minimum needed sample size to achieve the accuracy level given by the maximum error ε and the confidence level $1\text{-}\alpha$:

$$\left(\varepsilon^2 + \frac{z_{1-\alpha}{}^2 P(1-P)}{N}\right)\cdot n^2 - \left(\varepsilon + z_{1-\alpha}{}^2 P(1-P)\right)\cdot n + 0.25 \le 0$$

$$\Rightarrow \quad n \ge \frac{\varepsilon + z_{1-\alpha}^2\cdot P\cdot(1-P) + \sqrt{\left(\varepsilon + z_{1-\alpha}^2\cdot P\cdot(1-P)\right)^2 - \left(\varepsilon^2 + \frac{z_{1-\alpha}^2\cdot P\cdot(1-P)}{N}\right)}}{2\cdot\left(\varepsilon^2 + \frac{z_{1-\alpha}^2\cdot P\cdot(1-P)}{N}\right)}$$

E.*Approximation of Standard Error with P_{thres}*

Estimation Error calculated with real proportion:

$$\varepsilon_R = z_{1-\alpha}\cdot\sqrt{\frac{P\cdot(1-P)}{n}}$$

Estimation Error calculated with P_{thres}:

$$\varepsilon_T = z_{1-\alpha}\cdot\sqrt{\frac{P_{thres}\cdot(1-P_{thres})}{n}}$$

If we know the real error, a violation occurs if:

$\hat{P} + \varepsilon_R \ge P_{thresh}$ (user-friendly approach)

If we estimate the error with P_T a violation occurs if:

$\hat{P} + \varepsilon_T \ge P_{thresh}$

We use a proof by contradiction to show that this approach is able to detect all violations. If $\hat{P} + \varepsilon_T < \hat{P} + \varepsilon_R$ it could happen that we do not detect a violation that would have been detected using the real error. But a violation occurs only if $P>P_{thres}$:

$\hat{P} + \varepsilon_T < \hat{P} + \varepsilon_R$ if $P>P_{thres}$?

$\Rightarrow \quad \varepsilon_T < \varepsilon_R$

$$\Rightarrow \quad z_{1-\alpha}\cdot\sqrt{\frac{P_{thres}\cdot(1-P_{thres})}{n}} < z_{1-\alpha}\cdot\sqrt{\frac{P\cdot(1-P)}{n}}$$

with $P=a+P_{thres}$ and $0 \le a \le 1$ (because $P>P_{thres}$) we get:

$$P_{thres} - P_{thres}^2 < a + P_{thres} - (a + P_{thres})^2$$

$$\Rightarrow \quad P_{thres} < -\frac{1}{2}(1-a)$$

Since P_{thres} is positive and $0 \le a \le 1$ the equation is always false and we can conclude that all real violations, that are detected knowing the real error, are also detected with the approximated error.