Validation of Automatic Passenger Counting: Introducing the t-Test-Induced Equivalence Test

Michael Siebert

Interautomation Deutschland GmbH, Berlin, Germany

David Ellenberger

Interautomation Deutschland GmbH, Berlin, Germany
Department of Medical Statistics, University Medical Center Göttingen, Germany

Abstract

Automatic passenger counting (APC) in public transport has been introduced in the 1970s and has been rapidly emerging in recent years. Still, real-world applications continue to face events that are difficult to classify. The induced imprecision needs to be handled as statistical noise and thus methods have been defined to ensure that measurement errors do not exceed certain bounds. Various recommendations for such an APC validation have been made to establish criteria that limit the bias and the variability of the measurement errors. In those works, the misinterpretation of non-significance in statistical hypothesis tests for the detection of differences (e.g. Student’s t-test) proves to be prevalent, although existing methods which were developed under the term equivalence testing in biostatistics (i.e. bioequivalence trials, Schuirmann, 1987) would be appropriate instead. This heavily affects the calibration and validation process of APC systems and has been the reason for unexpected results when the sample sizes were not suitably chosen: Large sample sizes were assumed to improve the assessment of systematic measurement errors of the devices from a user’s perspective as well as from a manufacturers perspective, but the regular t-test fails to achieve that. We introduce a variant of the t-test, the revised t-test, which addresses both type I and type II errors appropriately and allows a comprehensible transition from the long-established t-test in a widely used industrial recommendation. This test is appealing, but still it is susceptible to numerical instability. Finally, we analytically reformulate it as a numerically stable equivalence test, which is thus easier to use. Our results therefore allow to induce an equivalence test from a t-test and increase the comparability of both tests, especially for decision makers.

Keywords: automatic passenger counting, APC validation, APC accuracy, revenue sharing, equivalence testing, post-hoc power adaptions

---

1michael.siebert@interautomation.de
2david.ellenberger@med.uni-goettingen.de
Highlights

- Review of the development in automatic passenger counting (APC) and characterization of the state-of-the-art APC validation and accuracy.

- Illustration of common pitfalls in hypothesis testing, that especially affect the APC approval process and were not handled so far.

- Introduction of statistical methods to guarantee that the type I and the type II errors are controlled as well as an appropriate sample size calculation.

- Creation of a new comprehensible analytical transition from the t-test to equivalence testing. Decision makers can now precisely compare both methods with the option to replace the t-test, even in high impact applications.

- Successful and fundamental modification of a widely used industrial recommendation that affects revenue sharing in public transportation in the billions.

1. Introduction

Assessment of passenger counts is of paramount importance for public transport agencies in order to plan, manage and evaluate their transit service. Application covers many topics, for example short- and long-term forecasting, optimizing passenger behaviour and daily operations, or sharing of revenue among operators. Issues of passenger demand have a long-lasting history (see e.g. Kraft and Wohl, 1968). In recent years, modelling of passenger counts has emerged rapidly due to the availability of large-scale automatic data collection. Data on (automatic) passenger counts has direct impact on both the revenue generated by ticket sales as well as state subsidies of public transport companies within one unified ticketing system. To illustrate, in Berlin, buses and underground are operated by the BVG and the S-Bahn Berlin, a complementary rapid transit system, by the Deutsche Bahn AG. Most tickets, e.g. a day ticket or monthly pass are issued by the Transport and Tariff Association of Berlin and Brandenburg (VBB) and allow passengers to use both services among around 40 others. Excluding subsidies, the revenue from the ticket sales alone has been around 1.4 billion EUR in 2016. Revenue magnitudes within the billions are common in public transport (Armstrong and Meissner, 2010).

APC systems have evolved considerably within the past 40 years. Passenger flow data can be acquired with high accuracy outperforming manual ride checkers (Hodges, 1985; Hwang et al., 2006). Devices that operate on 3D image streams are the industrial state-of-the-art technology. Latest generation devices offer an accuracy of around 99% (iris, 2018b; Hella Aglaia, 2018) and technical progress is ongoing. As all measurement devices, APC systems are susceptible to error. For the comparison of counting precision between different sensors objective statistical criteria are therefore required. These criteria are not only needed for comparisons between APC systems but also decision making processes rely on high accuracy APC data (Furth et al., 2005). Upraising usage of APC systems led to the formulation of some precision criteria to ensure validity and reliability. The term APC validation for this type of quality control was

---

3see e.g. [http://de.wikipedia.org/w/index.php?title=Verkehrsverbund%20Berlin-Brandenburg&oldid=175231663;](http://de.wikipedia.org/w/index.php?title=Verkehrsverbund%20Berlin-Brandenburg&oldid=175231663;) accessed 22-March-2018
used by Strathman (1989) and some wider usage of validation concepts awoke since the early 2000s (see e.g. Kimpel et al., 2003; Strathman et al., 2005; Boyle, 2008; Chu, 2010; Köhler et al., 2015).

For real-world APC validations the most relevant criterion is to ensure unbiasedness, i.e. the need to rule out that the APC system makes a relevant a systematic error. Especially regarding revenue sharing this is crucial since small errors – of whatever origin – can already have a large impact: to illustrate, companies with a shared ticketing system like the above mentioned BVG or the S-Bahn Berlin GmbH have revenues, consisting of ticket sales and subsidies, of roughly a billion EUR each\(^4\). If one of these companies somehow systematically counted 1\% too few and the other one 1\% too many passengers and passenger counts accounted to only 10\%\(^5\) of the shared revenue, yet two million EUR would be distributed inappropriately – every year, for these two companies operating in the Berlin area alone. In Germany, it is prevalent that the tickets are sold by the transport association and revenue as well as subsidies are distributed among the transport companies (Beck, 2011), which accounted for 12.8 billion EUR in 2017 (VDV, 2018). Furthermore, such a revenue sharing scheme itself is currently associated with high costs, which are roughly 1 million EUR for the VBB in 2014 (Baum and Gaebler, 2015).

One industrial recommendation regarding APC systems is central to tendering procedures in Germany and also advertised by manufacturers worldwide (Hella Aglaia, 2018; iris, 2018a): the VDV 457 (Köhler et al., 2015), which regularly and in an unmodified form becomes part of transportation contracts, sometimes even in the latest, yet unreleased version that has “this is a pre-release” watermarked on every page. Due to the huge impact of the document, all change requests to the VDV 457 must be approved by a committee within the Association of German Transport Companies (Verband Deutscher Verkehrsunternehmen, VDV).

Results presented in this manuscript are given as follows: In Section 2 we summarize and discuss the development in (automatic) passenger counting. The complete statistical model of APC measurements we introduce in Section 3. In Section 4 we define and examine the revised t-test, which is an attempt to modify resp. extend the t-test to account for the type II error accordingly. There were two reasons for this approach: Firstly, the admission process based on the t-test was already established in the VDV 457, so we wanted to change as little as possible to make the impact of the changes foreseeable and manageable by decision makers. Secondly, being unopinionated was more likely to succeed than simply insisting to use a statistic test because it was popular in other fields like biostatistics. Subsequently, we illustrate that this newly introduced test generally suffers from numerical instability making the approach unsuitable for wide practical use. In Section 5 we introduce the equivalence test and in Section 6 we normalize the test criteria of both the revised t-test and the equivalence test to analytically see, that, after a transposition of parameters, the tests are identical. This so-obtained t-test-induced equivalence test however is, due to only elementary calculations being made, generally not susceptible to numerical instability. We close with some concluding remarks and future prospects in Section 7.

\(^4\)see e.g. http://de.wikipedia.org/w/index.php?title=S-Bahn%20Berlin%20GmbH&oldid=174014190, http://de.wikipedia.org/w/index.php?title=Berliner%20Verkehrsbetriebe&oldid=175043479 ; accessed 19-March-2018

\(^5\)Based on (Beck, 2011): 10\% of the contracts were net-cost contracts. The actual shares of the individual transport companies in the VBB are considered trade secrets of the individual transport companies and are thus not disclosed, not even to a request of the House of Representatives of Berlin (Baum and Gaebler, 2015).
2. APC development and current practice

Traditionally, but also nowadays, passenger counts are collected manually via passenger surveys or human ride checkers, which are both expensive and produce only small samples. Former, the passenger surveys, are possibly biased and unreliable (Attanucci et al., 1981). For latter, the manual counts by ride checkers, the accuracy is doubtable, since already the first-generation automatic counting systems have been regarded to be more accurate (Hodges, 1985). Ride checking is often done by less qualified personnel with high turn-over rates and Furth et al. (2005) instead suggest the use of video cameras to increase accuracy and reliability for any evaluations. Today, automatic data collection (ADC) systems in public transport are classified into automatic vehicle location (AVL), automatic passenger counting (APC), and automatic fare collection (AFC) systems (Zhao et al., 2007). The AVL system provides data on the position and timetable adherence of the bus, metro, or train which needs to be merged with APC data (Furth et al., 2004; Strathman et al., 2005; Saavedra et al., 2011). AFC data is based on ticket sales, magnetic strip cards, or smart cards and has become popular since it is often easily available (Zhao et al., 2007; Lee and Hickman, 2014). Still, it often only provides information on boarding but not alighting, generally underestimates actual passenger counts and may therefore be less accurate than APC data (see e.g. Wilson and Nuzzolo, 2008; Chu, 2010; Xue et al., 2015).

First generation of APC systems were deployed in the 1970s (Attanucci and Vozzolo, 1983) and usage increased in the following decades. Casey et al. (1998) reported that plenty local metropolitan transit agencies use APC systems and Strathman et al. (2005) reported increase rates in APC usage of over 445% within 7 years. Today APC systems are used worldwide and have found their way into official documents, as the above mentioned tendering procedures in Germany. In the United States, transit agencies using APC data have to submit a benchmarking and a maintenance plan for reporting to the FTA’s National Transit Database (NTD) to be eligible for related grant programs (see e.g. Chu, 2010).

A wide range of competing APC technologies has been developed. Detection methods include infrared light beam cells, passive infrared detectors, infrared cameras, stereoscopic video cameras, laser scanners, ultrasonic detectors, microwave radars, piezoelectric mats, switching mats and also electronic weighing equipment (EWE) (Casey et al., 1998; Kuutti, 2012; Kotz et al., 2015). Transit agencies usually mount one or multiple sensors to collect APC data in each door area of public transport vehicles like busses, trams and trains. The number of boarding and alighting passengers are counted separately by converting 3D video streams (infrared beam break) or light barrier methods, which are the most commonly used technologies (Kotz et al., 2015). In recent years also weight based EWE approach utilizing pressure measurements in the vehicle braking / air bag suspension system has emerged to estimate passenger numbers (Nielsen et al., 2014; Kotz et al., 2015). The relatively new approach has proved to provide easy-to-acquire additional information, since modern buses and powertrains are equipped with (intelligent) pressure sensors by default.

First assessment as APC validity, i.e. accuracy, date back to the 1980s when large-scale usage started in the United States and Canada (Hodges, 1985). To assess APC systems several researchers used confidence intervals and tests for paired data to investigate whether any found bias is statistically significant. Tests mostly used is the t-test (Strathman, 1989; Kimpel et al., 2003; Köhler et al., 2015), but also non-parametric Wilcoxon test for paired data has been used in automated data (Kuutti, 2012). Handbooks for reporting to FTA’s National Transit Database have adapted the t-test as well as the industrial recommendations for APC-buying transit agencies like (Köhler et al., 2015). To our knowledge no APC t-test related criterion formulated so far controls the type II error of the statistical test. Some authors report
concepts that resemble equivalence testing. Furth et al. (2006) states “A less stringent test would allow a small degree of bias, say, 2% (partly in recognition that the ‘true’ count may itself contain errors); [...]” which acknowledges the fact that almost no measurement in the real world will have an expected value of exactly zero. In a survey among transit agencies by Boyle (2008) on how they ensure that APC systems meet a specified level of accuracy it is reported “Some [agencies] were more specific, for example, with a confidence level of 90% that the observations were within 10% of actual boardings and alightings.”, which is an early occurrence of an equivalence test concept. Conversely, Chu (2010) introduced an “equivalency test” for APC benchmarking, which however is not to be confused with the equivalence test but rather is the application of the objected t-test to average passenger trip length. Additional adjustment factors on the raw APC counts are given without defining any equivalency criteria, an issue this paper shall address properly.

Various alternative criteria exist also to the t-test to assess accuracy resp. unbiasedness on the one hand and precision resp. reliability on the other hand. Nielsen et al. (2014) also investigate absolute differences in addition to evaluate the bias when analysing a weight-based APC approach. Restrictions on the absolute deviation from zero also limit the variability of the APC system. Criteria specifically on the variance of the APC have been made indirectly through the error rate or more specifically through specifying the allowed distribution of errors (see e.g. criteria b and c in Köhler et al., 2015), (Appendix E in Furth et al., 2003; Boyle, 2008). To the best of our knowledge, the most comprehensive and maintained industrial recommendation on APC validation and usage is the above mentioned VDV Schrift 457 (Köhler et al., 2015). The document gives guidance on most relevant APC topics, including sampling and standardization of APC validation. One major aspect of APC validation is the demonstration of adequate APC accuracy regarding which Köhler et al. (2015) state for the approval process: for an APC system to pass the admission process, its systematic error has to be at most 1%, which is verified by (a variant of) the t-test. Worldwide, there are similar formulations for the validation and thus admission of APC systems, like Furth et al. (2006); Boyle (2008); Chu (2010).

However, scepticism within the industry arose seemingly good performing APC systems started to fail the test. In February 2015, with the help of a brute force algorithm, we constructed a proof of concept for a failed (APC) t-test: the error is almost zero, i.e. with 1000 (or arbitrary many more) boarding passengers the sample has three measurements, one with an error of one, the other two with an error of two passengers. In that case, the APC system fails the t-test. Such a proof of concept led the count precision work group (Arbeitsgemeinschaft Zählgenauigkeit) of the VDV to add the equivalence test with additional restrictions, as an exceptional alternative test alongside the t-test in the VDV 457 v2.0 release in June 2015 (Köhler et al., 2015) to account for APC systems with a low error standard deviation. Indeed, the above mentioned proof of concept would now be accepted by the new, hybrid test, but as it turned out later, current or near future APC systems would not profit, since the parameter choice was too hard to pass. Further, a remark was added to the VDV 457 that “in the advent of technological advance and increased counting precision, the admission process is still subject to change”: at that time, there was still little insight into why a seemingly suitable and popular statistical test exposed such a seemingly arbitrary behaviour and it was not entirely clear how the equivalence test compared to the long-established t-test.

Detailed investigations showed that the VDV 457 t-test variant only accounted for the type I error, defined to be 5% to 10%, which is the risk for an APC systems manufacturer to fail the test with a system with having a systematic error of zero. In the t-test terminology, this parameter is known as statistical significance $\alpha$. Conversely, the type II error $\beta$ is the risk of an APC system with a systematic error greater
than 1% to obtain admission, which is the complement to the statistical power $1 - \beta$. The type II error and thus the statistical power was neither accounted for in the sample size planning nor in the testing procedure. Through the sample size formula it was implicitly 50%, assuming the a priori estimated standard deviation was correct. Otherwise, the higher the empirical standard deviation, the greater the type II error and vice versa. The statistical framework for APC validation and methods to address the current shortcomings are given in the following Sections.

3. Statistical model

Let $\Omega_0 = \{\omega_i\}, i = 1, \ldots, N$ be the statistical population of door stop events (DSE)\(^6\), which are used to summarize all boarding and alighting passengers at a single door during a vehicle (bus, tram, train) stop. Further, let $\Omega = \{\omega_{ij}\}, i \in \{1, \ldots, N\}, j \in \{1, \ldots, n\}$ be a sample, which consists of either randomly or structurally selected DSE (e.g. by a given sampling plan). The size of the statistical population $N$ may be considered as the number all DSE over the relevant time period, which is typically one or more years, so $N$ can be assumed to be unbounded and thus $N = \infty$. Let $n$ be the sample size and $M_i$, $i \in \{i, \ldots, n\}$ be the manual count and $K_i$, $i \in \{i, \ldots, n\}$ be the automatic count of boarding passengers made by the APC system. The manual count obtained by multiple ride checkers or favourably video camera information (Kimpel et al., 2003) is assumed to be a ground truth to compare against. Alighting passengers are counted as well and results apply analogously, but w.l.o.g. we only consider the boarding passengers. Let $\bar{M} = \frac{1}{n} \sum_{i=1}^{n} M_i$ be the average manual boarding passenger count. Similar to other authors (see e.g. Furth et al. (2003), appendix E; Furth et al. (2006); Nielsen et al. (2014); Köhler et al. (2015)) we consider the random variables

$$D_i := \frac{K_i - M_i}{\bar{M}},$$

which we call relative differences being the difference of the automatically and manually counted boarding passengers relative to the average of the manually counted boarding passengers. The average $\bar{D} := \frac{1}{n} \sum_{i=1}^{n} D_i$ is the statistic of interest that is used in both the t-tests as well as the equivalence test. The expected value $\mu := E(\bar{D})$ is the actual systematic error\(^7\) of an APC system (Furth et al., 2005), since it can systematically discriminate participants of the revenue sharing system or could also be referred to as bias of the measurement device, a term frequently used in APC accuracy evaluations (Strathman, 1989; Kimpel et al., 2003; Furth et al., 2005; Chu, 2010; Nielsen et al., 2014).

Criteria in any APC approval process are often checked by specially trusted authorities authorized to grant admission. They perform their own manual ride check, evaluate the criteria i.e. the statistical test, and either approve or reject the APC system. There are two conflicting interests that need to be dealt with: acquiring maximally accurate and reliable data on the one hand and approve a high number of APC in a fast and cost-efficient process on the other hand. Shortcoming of the first we will call calibration resp. user risk and shortcomings of the latter manufacturer risk. We attribute the user risk to public transportation companies and network authorities, who rely on accurate data, despite that the motivations for APC data collection might be more complex in the real world. The manufacturer risk relates directly to possible recourse claims and negative market reputation if the APC systems fails the

---

\(^6\)Referred to as Haltestellentüreignis (HTE) in VDV 457

\(^7\)Referred to as Verzerrtheit in VDV 457, which means distortedness (as e.g. in market distortion)
admission. The two risks will distinct type I error and type II error of statistical tests. For the t-test, the hypotheses are

\[ H_0 : \text{There is no systematic APC measurement error} \ (\mu = 0) \]
\[ H_1 : \text{There is a systematic APC measurement error} \ (\mu \neq 0) . \] (2) (3)

Let \( \nu \) be the a priori estimated standard deviance, \( \hat{\nu} \) the empirical standard deviance of the sample, \( d_r \) the maximal allowed error (e.g. 1%), \( \alpha_t \) the risk of falsely rejecting the null hypothesis \( H_0 \) (type I error, i.e. rejecting an APC system with an actual systematic error of zero) and \( \beta_t \) the risk of falsely accepting the null hypothesis \( H_0 \) when a particular value of the alternative hypothesis \( H_1 \) is true (type II error, e.g. accepting an APC system with a systematic error of 1%) (Guthrie, 2010).

The sample size estimation for the t-test is given by

\[ n_t = \left( z_{1-\beta_t} + z_{1-\alpha_t/2} \right)^2 \frac{\nu^2}{d_r^2} , \] (4)

with \( z(\cdot) \) being the quantile function of the normal distribution and the test criterion as

\[ |D| \leq z_{1-\alpha_t/2} \frac{\hat{\nu}}{\sqrt{n_t}} . \] (5)

4. Revised t-test

Several discussions about post-hoc power adaptations for the t-test exist. A thorough discussion about those can be found in Hoenig and Heisey (2001). They argue that approaches referred to as Observed Power, Detectable Effect Size, or Biologically Significant Effect Size are “flawed”. For the latter approach, which is described in Cohen (1988), Hoenig and Heisey criticize the assumption that actual power is equal to the intended power and not updated according to experimental results (e.g. sampling variability). Addressing this, we investigate on procedures to control the (actual) type II error to assess non-presence of a crucial difference. Schuirmann (1987) initially referred to approaches of using a negative hypothesis test to make inference that no inequivalence was present as the Power Approach. Analogously to these thoughts, we will consider variations of the type I error \( \alpha \) to make adaptations to the testing procedure and call this approach post-hoc power calculation. This was explicitly mentioned by Schuirmann but was not derived further by stating a lack of practical relevance: “In the case of the power approach, it is of course possible to carry out the test of the hypothesis of no difference at a level other than 0.05 and/or to require an estimated power other than 0.80, but this is virtually never done.” While this approach may not have been used in the world of pharmaceutics, it is of relevance for the validation of devices for automatic passenger count and likely other applications in industrial statistics. In general, as well in practice, after the data collection, the a priori estimated standard deviation and the empirical standard deviation differ to some extent and we strongly believe that it cannot be relied on that difference being negligible. Therefore, we want to ensure that the risk of the user (the type II error) does not exceed a prespecified level, which is usually 5%. So the only parameter to be adapted is the type I error \( \alpha \), which is the risk of the device manufacturer. The appropriate \( \hat{\alpha}_t \), the revised significance, can
thus be determined by solving the equation

\[ n_t = \left( z_{1-\beta_t} + z_{1-\alpha_t/2} \right)^2 \frac{\nu^2}{d^2} , \]  

(6)

and thus

\[ \alpha_t = 2 \left[ 1 - z^{-1} \left( \left( z_{1-\beta_t} + z_{1-\alpha_t/2} \right) \frac{\nu}{\nu} - z_{1-\beta_t} \right) \right] . \]  

(7)

Note that \( n_t \), by choice of the initial \( \alpha_t \) and \( \nu \), is fixed. If the actual sample size does not match the planned sample size, \( n_t \) resp. \( \nu \) need to be adapted. Analogously to equation (5), we define the test criterion for the revised t-test:

\[ |D| \leq z_{1-\alpha_t/2} \frac{\nu}{\sqrt{n_t}} , \]  

(8)

which however, can yield a problematic behaviour in practice: First, the term \( z_{1-\alpha_t/2} \) is undefined for \( \alpha_t > 2 \). Combined with equation (6), this induces a lower bound on \( n_t \):

\[ n_t \geq z_{1-\beta_t}^2 \frac{\nu^2}{d^2} . \]  

(9)

Second, \( z(1) = \infty \) is the source of a different problem: if \( \nu/\nu \) reaches a certain critical value, which is 2.76 for \( \alpha_t = \beta_t = 5\% \). Therefore, this could be relevant in practice, since it yields \( z_{1-\alpha_t/2} = \infty \) due to rounding errors. In that case, the test criterion from equation (8) is always true and the test is thus always passed as illustrated in Fig. 1.

5. Equivalence testing

Equivalence testing has originated in the field of biostatistics. Often the term bioequivalence testing (e.g. Schuirmann, 1987; Berger and Hsu, 1996; Wellek, 2010) is used. Bioequivalence tests are statistical tools that are commonly used to compare the performance of generic drugs with established drugs using several commonly accepted metrics of drug efficacy. The term equivalence between groups means that differences are within certain bounds, as opposed to complete equality. These bounds are application-specific and are usually to be chosen such that they are below any potential (clinically) relevant effect (Ennis and Ennis, 2010). In many publications the problem was referred to as the two one-sided tests (TOSTs) procedure (Schuirmann, 1987; Westlake, 1981). TOST procedures were developed under various parametric assumptions and additionally distribution-free approaches exist (Wellek and Hampel, 1999; Zhou et al., 2004).

Equivalence tests have begun making their way into psychological research (see e.g. Rogers et al., 1993) and natural sciences: Hatch (1996) applied it for testing in clinical biofeedback research. Parkhurst (2001) discussed the lack of usage of equivalence testing in biology studies and stated that equivalence tests improve the logic of significance testing when demonstrating similarity is important. Richter and
Figure 1: Chances (of an APC system) to pass dependent on the method to determine the revised significance $\hat{\alpha}$, which we have obtained from equation (7), denoted by dashed lines, and by using the function `power.t.test` from R Core Team (2017) denoted by the solid lines. Latter was our initial approach, which we illustrate here for the purpose of completeness. We notice that for far too low sample sizes (red and dark red lines) the former method is stricter. For the dark red dashed line, $n$ is below the lower bound from equation (9) and we thus assume the test to always fail, compare Section 6. For (too) large sample sizes numerical instabilities, which cannot be detected by the user through e.g. error messages, lead to sudden gaps in the function values (green lines for $\nu = 0.322$ and $\nu = 0.323$) for the `power.t.test`-variant, which relies on fixed-point iteration and has a history of unexpected convergence behaviour. The blue lines denote a practically reachable numeric limit when using equation (7): starting with $\nu \geq 0.414$ resp. $\nu / \hat{\nu} \geq 2.76$ all systems are always accepted. For the `power.t.test` approach, the light blue line is slightly above the dark blue one, so it has numerical problems for these values, too. Generally, all values of $\nu$ have to be seen w.r.t. the ratio $\nu / \hat{\nu}$, since these numerical effects can already occur with much smaller sample sizes.

Richter (2002) used equivalence testing in industrial applications and gave instructions how to easily calculate it with basic spreadsheet computer programs. Applications also involved risk assessment (Newman and Strojan, 1998), plant pathology (Garrett, 1997; Robinson et al., 2005) ecological modelling (Robinson and Froese, 2004) analytical chemistry (Limentani et al., 2005), pharmaceutical engineering (Schepers and Wätzig, 2005), sensory and consumer research (Bi, 2005), assessment of measurement agreement (Yi et al., 2008), sports sciences (Vanhelst et al., 2009), applications to microarray data (Qiu and Cui, 2010), genetic epidemiology in the assessment of allelic associations (Gourraud, 2011), and geography (Waldhoer and Heinzl, 2011).

For the equivalence test, the hypotheses are (Schuirmann, 1987; Julious, 2004)

\begin{align*}
H_0 : \text{There is a (relevant) systematic APC measurement error } (|\mu| \geq \Delta) \quad & \quad (10) \\
H_1 : \text{There is no (relevant) systematic APC measurement error } (|\mu| < \Delta) . \quad & \quad (11)
\end{align*}
We define $\Delta$ to be the equivalence margin and the relevant errors for the equivalence test with $\alpha_e$ referring to (half) the risk of the user and $\beta_e$ to the risk of the device manufacturer. We will consider two-sided $1 - 2\alpha_e$ confidence intervals (symmetric around the mean) where $\alpha_e$ is often to be chosen 2.5%. The usage of $1 - \alpha_e$ confidence intervals is also possible (see e.g. Westlake, 1981) but is used less frequently in the recent literature in this topic. Note that, by definition, the meanings of the $\alpha$ and $\beta$ are interchanged between the t-test and the equivalence criterion in referring to the risk of the user and to the risk of the manufacturer.

For the equivalence test sample size estimation exists (see e.g. Liu and Chow, 1992; Julious, 2004) similar to equation (4) of the t-test:

$$n_e = \left( \frac{z_{1-\beta_e/2} + z_{1-\alpha_e}}{\nu} \right)^2 \frac{\nu^2}{\Delta^2} .$$  \hspace{1cm} (12)

We define the test criterion for the equivalence test

$$|D| \leq \Delta - z_{1-\alpha_e} \frac{\hat{\nu}}{\sqrt{n_e}} ,$$  \hspace{1cm} (13)

which is formulation of the Two One-Sided Test Procedure for the crossover design in the case of limits that are symmetrical around zero (see Schuirmann, 1987).

6. t-test-induced equivalence test

An approach to compare the revised t-test from Section 4 to the equivalence test from Section 5 is to normalize and compare their test criteria from equations (8) resp. (13) as well as their sample sizes formulas (4) resp. (12). By combining equations (4), (6) and (8), we obtain a normalized test condition for the revised t-test:

$$|D| \leq z_{1-\alpha_t/2} \frac{\hat{\nu}}{\sqrt{n_t}} = \left( \frac{\nu}{\nu} (z_{1-\beta_t} + z_{1-\alpha_t/2}) - z_{1-\beta_t} \right) \frac{\hat{\nu}}{\sqrt{n_t}}$$

$$= \left( \frac{\nu}{\nu} (z_{1-\beta_t} + z_{1-\alpha_t/2}) - z_{1-\beta_t} \right) \frac{1}{\nu} \frac{\hat{\nu}}{z_{1-\beta_t} + z_{1-\alpha_t/2}} d_r$$

$$= \left( 1 - \frac{\nu}{\nu} \frac{1}{1 + \frac{z_{1-\alpha_t/2}}{z_{1-\beta_t}}} \right) d_r .$$  \hspace{1cm} (14)

Using equations (12) and (13) we obtain

$$|D| \leq \Delta - z_{1-\alpha_e} \frac{\hat{\nu}}{\sqrt{n_e}} = \Delta - z_{1-\alpha_e} \frac{\hat{\nu}}{\nu} \left( \frac{1}{z_{1-\beta_e/2} + z_{1-\alpha_e}} \right)^2 \Delta$$

$$= \left( 1 - \frac{\nu}{\nu} \frac{1}{1 + \frac{z_{1-\beta_e/2}}{z_{1-\alpha_e}}} \right) \Delta ,$$  \hspace{1cm} (15)

which resembles equation (14). If we now chose

$$\beta_e := \alpha_t, \quad \alpha_e := \beta_t \quad \text{and} \quad \Delta := d_r$$  \hspace{1cm} (16)
for equations (12) and (15), they are identical to equations (4) and (14). Therefore, the revised t-test analytically is an equivalence test, with error types swapped and an extended domain: Since only elementary calculations are made and there is no need to handle a varying quantile function \( z_{\nu,\alpha} \), there is neither a lower bound as in equation (9), nor an upper bound due to numeric instability as illustrated in Fig. 1. We call an equivalence test with parameters chosen as in equation (16) a t-test induced equivalence test. For a visual comparison of the t-test and the equivalence test, see Fig. 2.

(a) t-test: \( \alpha_t = 5\% \)

(b) Revised t-test: \( \beta_t = 5\% , d_r = 1\% \) resp.
equivalence test \( \alpha_e = 5\% , \Delta = 1\% \)

(c) type I and type II errors

Figure 2: Chances of an APC system with a both sample and actual error standard deviation of \( \hat{\nu} = 0.15 \) to pass the t-test (a) or the revised t-test resp. equivalence test (b) over the actual systematic error. Different lines denote different sample sizes obtained from different a priori estimated error standard deviations \( \nu \). The golden, solid curve always represents a correctly estimated sample size, the green curve a sample which is too large and the reddish curves samples which are too small. The dashes in the dark red line in (b) denote the consequences of equation (9): only for the equivalence test, the outcome is defined for equation (13) and the test may be considered as always failed if \( n = 320 < z_{\nu,\alpha}^2 \). \( \hat{\nu}^2 / d_r^2 \) = 440.99. Using the original VDV 457 with an (implicit) power of 50\% yields the dashed golden curve in (a). The vertically striped areas are additionally correctly accepted, the horizontal striped areas are additionally incorrectly accepted. The thick blue lines denote the relative error of \( d_e = 1\% \). For comparison: in (c) the incorrect decisions of a reference test are red, the correct decisions are coloured cyan. The reason for red areas to exist are economic considerations to limit the test costs: further increasing the sample size towards infinity would make the red areas disappear, at least for the revised t-test resp. the equivalence test. For the t-test, the areas with systematic error \( \mu > 1\% \) and \( \mu < -1\% \) remain blue, but the inner turns red. This behaviour is counterintuitive to the idea that the error of a statistic test goes to zero as the sample size goes to infinity.
7. Conclusion

We illustrated that the t-test as a criterion for APC approval may exhibit undesirable properties, even as the sample size grows beyond a certain level. Further, we have shown that when trying to compensate this behaviour by using post-hoc power calculations with a revised t-test, issues of numeric stability and domain limitations arise. Finally, we have proven analytically that the t-test-induced equivalence test, being numerically stable with a practically unlimited domain, can supersede the revised t-test.

The equivalence test is popular in various fields and, from a user’s perspective, easier to apply than post-hoc power calculations. Our results thus not only apply to APC systems: every use of the t-test can now comfortably be reconsidered and – on demand – be replaced by the t-test-induced equivalence test.

Our work simplifies the decision making process considerably, especially when it affects the worldwide revenue sharing in public transport, where there have been made 243 billion public transport journeys in the year of 2015 alone (UITP, 2017). For this reason, a large German public transportation company, which was significantly involved in the creation of the original, t-test based recommendation, commissioned an additional, complementary experts report, which eventually confirmed our findings. A change proposal for the VDV 457 was made and its acceptance is soon to come with the t-test being replaced as APC validation criteria in 2019. The (t-test-induced) equivalence test is thus in the course of becoming the new recommendation.

Finally, we hope that long lasting arguments within the industry about seemingly arbitrary admission results now end and also that our work will enable a broader audience to understand and profit from equivalence testing.

Acknowledgements

This research is financially supported by the European Regional Development Fund.

References

Armstrong, A., Meissner, J., 2010. Railway revenue management: overview and models.

Attanucci, J., Burns, I., Wilson, N., 1981. Bus Transit Monitoring Manual. volume 1. Technology Sharing Program, Office of the Secretary of Transportation.

Attanucci, J., Vozzolo, D., 1983. Assessment of operational effectiveness, accuracy, and costs of automatic passenger counters. HS-037 821.

Baum, A., Gaebler, C., 2015. Ausgleichszahlungen im Rahmen der VBB-Einnahmeaufteilung. URL: http://www.stiftung-naturschutz.de/fileadmin/img/pdf/Kleine_Anfragen/S17-16190.pdf.

Beck, A., 2011. Barriers to entry in rail passenger services: empirical evidence for tendering procedures in Germany. European Journal of Transport and Infrastructure Research 11.

Berger, R.L., Hsu, J.C., 1996. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science 11, 283–319.
Bi, J., 2005. Similarity testing in sensory and consumer research. Food Quality and Preference 16, 139–149.

Boyle, D.K., 2008. Passenger counting systems. 77, Transportation Research Board.

Casey, R.F., Labell, L.N., Carpenter, E.J., LoVecchio, J.A., Moniz, L., Ow, R.S., Royal, J.W., Schwenk, J.C., Schweiger, C.L., Marks, B., 1998. Advanced public transportation systems: the state of the art. update ’98. Technical Report FTA-MA-26-7007-98-1. United States. Federal Transit Administration.

Chu, X., 2010. A guidebook for using automatic passenger counter data for National Transit Database (NTD) reporting. Technical Report. National Transit Resource Center.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Hillsdale. NJ: Lawrence Earlbaum Associates 2.

Ennis, D.M., Ennis, J.M., 2010. Equivalence hypothesis testing. Food Quality and Preference 21, 253–256.

Furth, P., Muller, T., Strathman, J., Hemily, B., 2004. Designing automated vehicle location systems for archived data analysis. Transportation Research Record: Journal of the Transportation Research Board, 62–70.

Furth, P., Strathman, J., Hemily, B., 2005. Part 4: Marketing and fare policy: Making automatic passenger counts mainstream: Accuracy, balancing algorithms, and data structures. Transportation Research Record: Journal of the Transportation Research Board, 205–216.

Furth, P.G., Hemily, B., Muller, T., Strathman, J.G., 2003. Uses of archived avl-apc data to improve transit performance and management: Review and potential. TCRP Web Document 23.

Furth, P.G., Hemily, B., Muller, T.H., Strathman, J.G., 2006. Using archived avl-apc data to improve transit performance and management. TCRP Report.

Garrett, K., 1997. Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. Phytopathology 87, 372–374.

Gourraud, P.A., 2011. When is the absence of evidence, evidence of absence? use of equivalence-based analyses in genetic epidemiology and a conclusion for the kif1b rs10492972* c allelic association in multiple sclerosis. Genetic epidemiology 35, 568–571.

Guthrie, W.F., 2010. 7.2.2.2. Sample sizes required. NIST/SEMATECH Engineering Statistics Handbook, NIST/SEMATECH. URL: http://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm.

Hatch, J.P., 1996. Using statistical equivalence testing in clinical biofeedback research. Applied Psychophysiology and Biofeedback 21, 105–119.

Hella Aglaia, 2018. Public Transport: HELLA Aglaia People Sensing. URL: http://people-sensing.com/public-transport.

Hodges, C.C., 1985. Automatic Passenger Counter Systems: The State of the Practice. Technical Report DOT-I-87-36 (398). URL: https://rosap.ntl.bts.gov/view/dot/398.
Hoenig, J.M., Heisey, D.M., 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician 55, 19–24.

Hwang, M., Kemp, J., Lerner-Lam, E., Neuerburg, N., Okunieff, P.E., 2006. Advanced public transportation systems: the state of the art update 2006. Technical Report. United States. Federal Transit Administration.

iris, 2018a. Iris in Mannheim. URL: https://www.irisgmbh.de/en/referenzen/rnv-rhein-neckar-verkehr-gmbh.

iris, 2018b. Mobile passenger counting with IRMA. URL: https://www.irisgmbh.de/wp-content/uploads/16-Folder-Evolution-EN-1.pdf.

Julious, S.A., 2004. Sample sizes for clinical trials with normal data. Statistics in Medicine 23, 1921–1986. URL: http://dx.doi.org/10.1002/sim.1783, doi:10.1002/sim.1783.

Kimpel, T., Strathman, J., Griffin, D., Callas, S., Gerhart, R., 2003. Automatic passenger counter evaluation: Implications for national transit database reporting. Transportation Research Record: Journal of the Transportation Research Board, 93–100.

Köhler, S., Bobinger, S., Branick, R., Cerfontaine, B., Krogull, S., Luther, A., Ritschel, M., Schulze, M., Starck, M., Bruns, W., 2015. Automatische Fahrgastzählsysteme: Handlungsempfehlungen zur Anwendung von AFZS im öffentlichen Personenverkehr. VDV-Schriften, Verband Deutscher Verkehrsunternehmen (VDV), Köln. URL: https://gso.gbv.de/DB=2.1/PPNSET?PPN=832975419.

Kotz, A.J., Kittelson, D.B., Northrop, W.F., 2015. Novel vehicle mass-based automated passenger counter for transit applications. Transportation Research Record: Journal of the Transportation Research Board, 37–43.

Kraft, G., Wohl, M., 1968. New directions for passenger demand analysis and forecasting.

Kuutti, J., 2012. Testijärjestely ihmisvirtasensorien vertailua varten; a test setup for comparison of people flow sensors. URL: http://urn.fi/URN:NBN:fi:aalto-201209203108.

Lee, S.G., Hickman, M., 2014. Trip purpose inference using automated fare collection data. Public Transport 6, 1–20. URL: https://doi.org/10.1007/s12469-013-0077-5, doi:10.1007/s12469-013-0077-5.

Limentani, G.B., Ringo, M.C., Ye, F., Bergquist, M.L., MCSorley, E.O., 2005. Beyond the t-test: statistical equivalence testing.

Liu, J.p., Chow, S.C., 1992. Sample size determination for the two one-sided tests procedure in bioequivalence. Journal of Pharmacokinetics and Pharmacodynamics 20, 101–104.

Newman, M.C., Strojan, C., 1998. Risk assessment: Logic and measurement. CRC Press.

Nielsen, B.F., Frølich, L., Nielsen, O.A., Filges, D., 2014. Estimating passenger numbers in trains using existing weighing capabilities. Transportmetrica A: Transport Science 10, 502–517.
Parkhurst, D.F., 2001. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation: Equivalence tests improve the logic of significance testing when demonstrating similarity is important, and reverse tests can help show that failure to reject a null hypothesis does not support that hypothesis. AIBS Bulletin 51, 1051–1057.

Qiu, J., Cui, X., 2010. Evaluation of a statistical equivalence test applied to microarray data. Journal of biopharmaceutical statistics 20, 240–266.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Richter, S.J., Richter, C., 2002. A method for determining equivalence in industrial applications. Quality Engineering 14, 375–380.

Robinson, A.P., Duursma, R.A., Marshall, J.D., 2005. A regression-based equivalence test for model validation: shifting the burden of proof. Tree physiology 25, 903–913.

Robinson, A.P., Froese, R.E., 2004. Model validation using equivalence tests. Ecological Modelling 176, 349–358.

Rogers, J.L., Howard, K.I., Vessey, J.T., 1993. Using significance tests to evaluate equivalence between two experimental groups. Psychological bulletin 113, 553.

Saavedra, M., Hellinga, B., Casello, J., 2011. Automated quality assurance methodology for archived transit data from automatic vehicle location and passenger counting systems. Transportation Research Record: Journal of the Transportation Research Board, 130–141.

Schepers, U., Wätzig, H., 2005. Application of the equivalence test according to a concept for analytical method transfers from the international society for pharmaceutical engineering (ispe). Journal of pharmaceutical and biomedical analysis 39, 310–314.

Schuirmann, D.J., 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Pharmacodynamics 15, 657–680.

Strathman, J., Kimpel, T., Callas, S., 2005. Validation and sampling of automatic rail passenger counters for national transit database and internal reporting at trimet. Transportation Research Record: Journal of the Transportation Research Board, 217–222.

Strathman, J.G., 1989. An Evaluation of Automatic Passenger Counters: Validation, Sampling, and Statistical Inference. Technical Report.

UITP, 2017. Statistics brief, urban public transport in the 21st century. URL: http://www.uitp.org/sites/default/files/cck-focus-papers-files/UITP_Statistic%20Brief_national%20PT%20stats.pdf.

Vanhelst, J., Zunquin, G., Theunynck, D., Mikulovic, J., Bui-Xuan, G., Béghin, L., 2009. Equivalence of accelerometer data for walking and running: treadmill versus on land. Journal of sports sciences 27, 669–675.
Waldhoer, T., Heinzl, H., 2011. Combining difference and equivalence test results in spatial maps. International journal of health geographics 10, 3.

Wellek, S., 2010. Testing statistical hypotheses of equivalence and noninferiority. CRC Press.

Wellek, S., Hampel, B., 1999. A distribution-free two-sample equivalence test allowing for tied observations. Biometrical journal 41, 171–186.

Westlake, W., 1981. Bioequivalence testing—a need to rethink. Biometrics 37, 589–594.

Wilson, N.H., Nuzzolo, A., 2008. Schedule-based modeling of transportation networks: theory and applications. volume 46. Springer Science & Business Media.

Xue, R., Sun, D.J., Chen, S., 2015. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. Discrete Dynamics in Nature and Society 2015.

Yi, Q., Wang, P.P., He, Y., 2008. Reliability analysis for continuous measurements: equivalence test for agreement. Statistics in Medicine 27, 2816–2825.

Zhao, J., Rahbee, A., Wilson, N.H., 2007. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Computer-Aided Civil and Infrastructure Engineering 22, 376–387.

Zhou, J., He, Y., Yuan, Y., 2004. Online journal of pharmacokinetics©. Online Journal of Pharmacokinetics 3, 1–12.