NIST Technical Note 2106

Comparing Instruments

James Yen
Dennis Leber
Leticia Pibida

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2106

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Technical Note 2106

Comparing Instruments

James Yen
Dennis Leber
Statistical Engineering Division
Information Technology Laboratory

Leticia Pibida
Radiation Physics Division
Physical Measurement Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2106

September 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology Technical Note 2106
Natl. Inst. Stand. Technol. Tech. Note 2106, 57 pages (September 2020)
CODEN: NTNOEF

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2106

Abstract

This document details methods to compare instrument performance. Comparison methods for instruments outputting binary (0-1) responses as well as for instruments outputting continuous numeric responses are shown, first for two instruments and then for multiple instruments. Hypothesis tests and confidence intervals for instrument differences are demonstrated. Several nonparametric procedures are shown along with reasons why they might be needed. Finally, equivalence testing is demonstrated for those cases where testing for equivalence rather than for differences is indicated.

Keywords

Analysis of variance (ANOVA); binomial test response; confidence interval; continuous test response; equivalence testing; hypothesis test; nonparametric tests; normal approximation; Student's *t* test.

Table of Contents

List of Tables

List of Figures

Comparing Instruments

Often, experimenters wish to compare the performances of two or more instruments. For example: what is the difference between how well two detectors perform, or is there any significant differences in the instrument response between a group of several instruments?

1 Statistical Preliminaries

This document will begin with brief primers about hypothesis tests and confidence intervals, as those tools will be used repeatedly in this document. There will be an additional primer about using the normal approximation to create a *z* test, which is a type of hypothesis test used repeatedly in this document.

1.1 Hypothesis tests

In this document we describe several hypothesis tests for comparing instrument performance. NIST Technical Note 2045 [1], the NIST/SEMATECH e-Handbook *[2],* and Mendenhall and Sincich [3] all give an introduction to hypothesis testing. However, since the focus of this document is the comparison of instrument performance rather than, say, confirming performance thresholds, the terminology of the hypothesis tests will be geared toward that application. In this document, the null hypothesis will usually have the form:

H_0: All instruments perform equally well.

The alternative hypothesis will have the form:

H_A: Not all instruments perform equally well.

What is called in the statistical literature a *type I error* occurs when the hypothesis test result leads us to conclude that instruments perform differently when they are not actually different. Hypothesis tests can be specified so that the probability of a *type I error* is bounded above by a number α. Such a hypothesis is said to have *significance level* equal to α. The most commonly used value of α is 0.05, although 0.0, 0.10, and 0.20 values are also used depending on the type of instrument that is being tested and its use. These significance levels are used to weigh the risk associated with a *type I error*.

For our hypothesis tests, what we call a *type II error* occurs when there really are differences between the instruments, but the hypothesis test fails to reject the null hypothesis. The probability of such an error is usually denoted by β. The quantity $1-\beta$ is called the *power* of a test, which is the probability of rejecting the null hypothesis when there is an actual difference. The *power* of a hypothesis test is not a single number but depends on the magnitude of such an actual difference; a large difference between instrument will be much easier to detect than a small difference.

The power function of a hypothesis test is the power of the test as a function of the underlying difference.

There are trade-offs between type I and type II errors. To carry it to the most extreme, a hypothesis test that never rejects the null hypothesis will have zero type I errors, but have no power to detect any significant differences, thus leading to unacceptably high probability of type II errors.

1.1.1 *P-values*

When one does a hypothesis test using statistical or computational software, the output usually contains a *p-value*. (Except in certain unusual cases, a *p-value* will not be available without the use of statistical software or statistical functions in computational software.) The *p-value* is used to indicate a probability that is calculated after the collected data are analysed. The *p-value* is the probability of obtaining a statistic as extreme or more extreme than what actually occurred, given that the null hypothesis is actually true. In this document on instrument comparison, the *p-value* can usually be interpreted as the probability of measuring a disparity as large or larger than that seen, under the assumption that the instruments are actually equivalent. ***If the p-value is smaller than α, ($p<\alpha$), then the null hypothesis is rejected by the test with significance level α.*** For instance, if the *p-value* = 0.04, then the null hypothesis would be rejected at a significance level $\alpha = 0.05$, but not at a significance level $\alpha = 0.01$. The *p-value* indicates the smallest significance level for which the null hypothesis would be rejected; thus, the smallness of a *p-value* can be seen as an indicator of the weight of evidence against the null hypothesis.

There is a growing attitude in the scientific and statistical communities that one should not rely solely on hypothesis tests and associated *p-values* to draw conclusions. One should also examine other procedures, especially statistical intervals, which are described in the next section, as well as always doing an exploratory data analysis that includes looking at plots of the data.

1.2 Confidence intervals

A hypothesis test gives no information other than whether the null hypothesis of "no difference between the instruments" should be rejected. It does not specify which instrument performed better, the magnitude of the difference, and how practically significant the difference is. However, a statistical interval, often in the form of a confidence interval, can address those very issues, at least in part. See Hahn and Meeker [4] for a fuller discussion of confidence intervals and other intervals. In this document, the intervals will usually be those of performance differences between instruments. In the examples, the performance will be measured as a probability of detection or as a continuous measurement of performance such as the limit of detection of an instrument, or the magnitude of a signal measured by an instrument. The confidence interval for a difference includes values for the difference that are plausible given the data. The length of a confidence interval depends on the specified confidence level. For example, a procedure with a 90 % confidence level should include the correct value 90 % of the time. The higher confidence level that is desired, the larger that confidence interval must be to attain that confidence level. Thus, a 99 % confidence interval is longer than a 95 % confidence interval, which is in turn longer than a 90 % confidence interval.

Often there is a correspondence between confidence intervals and hypothesis testing as follows: ***If a 100×(1-α) % confidence interval for the difference contains zero, then that corresponds to the null hypothesis of "no difference" not being rejected by a corresponding hypothesis test with significance level α. Conversely, if the 100×(1-α) % confidence interval does not contain zero, that corresponds to the null hypothesis being rejected by the corresponding hypothesis test with significance level α.***

1.3 Normal approximation tests (*z* tests)

In various places in this document, we will provide hypothesis tests and procedures that rely on a normal distribution approximation. Due to a theorem from mathematical statistics called the Central Limit Theorem, statistics that are averages of a large number of independent, identically distributed random variables with finite mean and variance approximately follow a Normal distribution, also known as a Gaussian distribution. Suppose that a random variable x follows a Normal distribution with mean μ and variance σ^2 (and standard deviation σ), denoted by $N(\mu, \sigma^2)$. Then, x can be *standardized* by subtracting the mean and dividing by the standard deviation to $z = (x - \mu)/\sigma$, which has a *N(0,1)* distribution, known as a *standard normal* distribution. In practice, μ and σ are not known in advance and in the z statistic formula will be replaced by estimates of those parameters.

A z test involves a statistic z that should approximately follow a standard normal distribution if the Null Hypothesis is true. If $|z|$ is very large, then the Null Hypothesis becomes less plausible. More exactly, the *p-value* for a hypothesis test result is defined as the probability of obtaining a result just as extreme as the observed result if the null hypothesis were true. Since we will be doing comparison tests to see if quantities are the same or different, most of the tests in this document will be two-sided in that given an observed statistic z, the *p-value* will be of the form $P(|Z| \geq |z|)$, where Z is the standard normal variate.

Unfortunately, the cumulative distribution function $P(z<Z)$ of a standard normal variate, usually denoted $\Phi(Z)$, cannot be written in closed form. However, there are tables of the percentile points of the standard normal distribution in most statistical textbooks as well as in many online resources. **In particular, for the two-sided hypothesis tests used repeatedly in this document, it is useful to know the <u>two-sided</u> critical points of the standard normal distribution are 1.28 for significance level $\alpha = 0.20$, 1.645 for significance level $\alpha = 0.10$, 1.96 for $\alpha = 0.05$, and 2.576 for $\alpha = 0.01$.** Most statistical and computational software packages possess a function that outputs $\Phi(Z)$ for any real value of Z (as well as its inverse function $\Phi^{-1}(p)$)).

2 Binary data

2.1 Comparing two instruments with binary response

We will first explore the case of comparing two instruments having binary response. As an example, suppose that there are two detectors that register a "present/not present" response to a radioactive source. For the purposes of this section, suppose that during the measurement, a radioactive source is present so that detecting it would be considered a success, and not detecting would be a failure. Suppose there are n_1 trials for Instrument 1 and n_2 trials for Instrument 2. The number of detections by Instrument 1 would be x_1, with the resulting proportion of success being $p_1 = x_1/n_1$; for Instrument 2, there would be x_2 successes out of $n_2$trials with the resulting proportion $p_2 = x_2/n_2$. **The sample proportions $p_1$and p_2 are our best estimates of the true proportions P_1 and P_2, respectively.**

The number of successes for an instrument is modeled by a Binomial distribution modeling the number of successes in a set of independent Bernoulli trials; refer to NIST Technical Note 2045 [1], Ross [5] or most statistics textbooks for more background and justification for these models.

2.1.1 Testing for differences between proportions

2.1.1.1 Large samples: *z*-test

For relatively large sample sizes, we can create a test statistic that is approximately distributed as a standard normal variate under the null hypothesis. If the null hypothesis H_0 of "no difference between instruments" is true, let $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$ be the proportion of successes in the combined sample. Define

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \qquad (2.1)$$

Under the null hypothesis H_0 of "no difference between instruments", with relatively large samples sizes, *z* is approximately normal with mean 0 and variance 1. We can apply the previous section on normal approximation tests to this *z* statistic; the null hypothesis H_0 of "no difference between instruments" is rejected if the statistic *z* is too large or too small to be consistent with belonging to the standard normal distribution. **At the α significance level, that translates to rejecting the null hypothesis if *z* falls outside the middle *100*(*1-α/2*) percent of the standard normal distribution.** The figure below demonstrates that situation for significance level *α=0.10*; the area beneath the standard normal density curve between the two-sided critical points *-1.645* and *1.645* has area 1- α=0.90. The null hypothesis of no difference is rejected if

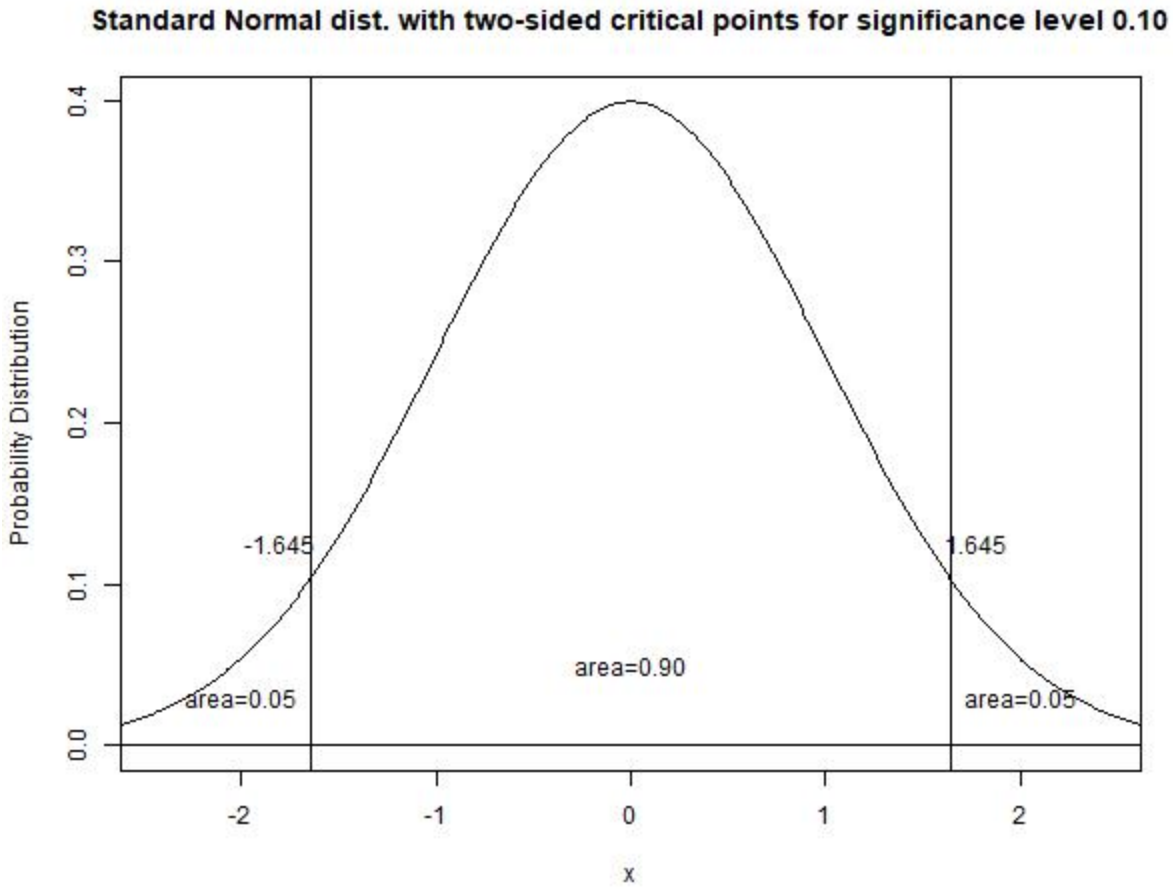

Figure 1: A standard normal distribution, shown with its two-sided critical points (-1.645 and 1.645) for significance level α = 0.10. If the z statistic is greater than 1.645 or smaller than -1.645, then the null hypothesis of no instrument differences is rejected.

the *z* statistic *z<-1.645* or *z> 1.645*.

To tell if the sample sizes are large enough, one can plug in the values of n_1, n_2, p_1, and p_2 into a formula provided by Mendenhall and Sincich [3]. They state that if the sample sizes n_1 and n_2 are large enough so that the intervals

$$p_1 \pm 2\sqrt{\frac{p_1(1-p_1)}{n_1}} \text{ and } p_2 \pm 2\sqrt{\frac{p_2(1-p_2)}{n_2}} \qquad (2.2)$$

do not contain 0 or 1, then the normal approximation is reasonably accurate. Otherwise, if the sample sizes are small enough that the above intervals contain 0 or 1, then the use of a nonparametric test such as Fisher's Exact Test (Section 2.1.1.2) is indicated.

Example: Suppose that we are trying to test the null hypothesis H_0 of "no difference between Instrument 1 and Instrument 2" at a significance level α=0.10. An experiment is performed

where Instrument 1 successfully detects a source 14 times out of 20 trials, while Instrument 2 detects a source 10 times out of 20 trials. Thus we have $p_1 = \frac{14}{20} = 0.7$, and $p_2 = \frac{10}{20} = 0.5$, and $\bar{p} = \frac{14+10}{20+20} = 0.6$. First, checking on the sample size criterion, we find that

$p_1 \pm 2\sqrt{\frac{p_1(1-p_1)}{n_1}} = 0.7 \pm 2\sqrt{\frac{0.7(1-0.7)}{20}} = 0.7 \pm 0.2 = (0.5, 0.9)$ and

$p_2 \pm 2\sqrt{\frac{p_2(1-p_2)}{n_2}} = 0.5 \pm \sqrt{\frac{0.5(1-.5)}{20}} = 0.5 \pm 0.22 = (0.28, 0.72)$.

Neither interval contains 0 or 1, so the normal approximation should be valid.

The *z* statistic is

$z = (0.7 - 05)/\sqrt{0.6(1-0.6)\left(\frac{1}{20} + \frac{1}{20}\right)} = 0.2/\sqrt{0.024} = 1.291$. The magnitude of *z* is smaller than 1.645, which is the two-sided standard normal critical point for significance level α=0.10; therefore, the null hypothesis of no performance difference between instruments is not rejected at this significance level. Using the normal cumulative distribution in statistical software gives the additional information that the *p*-value of *z*=1.291 is 0.19>0.10; since the *p*-value is larger than the significance level α=0.10, the null hypothesis is not rejected.

2.1.1.2 Fisher's Exact Test

When sample sizes are too small for the normal approximation tests to be appropriate, an alternative is a nonparametric test such as Fisher's Exact Test [2]. Suppose that the results of a two-instrument test can be tabulated in a 2×2 contingency table as follows:

	Successes	Failures	Column Totals
Instrument 1	*A*	*B*	*A+B*
Instrument 2	*C*	*D*	*C+D*
Row Totals	*A+C*	*B+D*	*N=A+B+C+D*

Fisher's Test enumerates how compatible the experimental results are with the null hypothesis. Under the null hypothesis of no real performance difference between the instruments, whatever differences did occur are due to chance. If we fix the marginal totals in the table and presume that the real success rate of both instruments is *(A+C)/N*, then the probability of each frequency in the contingency table is given by the hypergeometric distribution [5]:

$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}}$, where $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ is the number of combinations of *k* objects that can be chosen from a collection of *m*. Since we are focused on comparing instruments, the relevant *p*-value is the two-sided *p*-value, being the total probability of possible outcomes of equal or

greater instrument difference given fixed marginal totals. How those outcomes are tabulated is best explained using an example (below).

Example

Suppose that for an experiment testing two instruments, only 5 independent responses are available for each instrument. It is desired to have significance level $\alpha=0.05$. Suppose that Instrument 1 is successful in 4 of 5 trials, while Instrument 2 is successful in only 1 of 5 trials.

	Successes	Failures	Column Totals
Instrument 1	4	1	5
Instrument 2	1	4	5
Row Totals	5	5	10

Suppose that the recorded row and column totals are fixed, so that each instrument has 5 trials, and there is a total of 5 successes and 5 failures among these 10 total trials. A result showing the same performance difference between instruments as the actual experimental result has probability $\frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}} = \frac{25}{252} = 0.099$.

The only cases that are more extreme than the actual results in opposing the null hypothesis, while retaining the same row and column totals, are the cases where one instrument has 5 successes and the other instrument has 0 successes. Each of these cases has probability $\frac{\binom{5}{5}\binom{5}{0}}{\binom{10}{5}} = \frac{1}{252} = 0.004$. Since each case could have either Instrument 1 or Instrument 2 having the more successes, the relevant two-sided *p*-value is $2\times0.099 + 2\times0.004=0.206$, so the result is not significant at the $\alpha=0.05$ level. Note that if there are only 5 observations for each instrument, the only experimental result that would reject the null hypothesis at the significance level $\alpha=0.05$ is the most extreme possible difference result of one instrument being successful in all 5 trials and the other instrument failing in all 5 trials (in which case the resulting *p*-value is $2\times0.004=0.008$). This highlights how having a small sample size limits the extent of statistical conclusions one can make, as well as the limitations of Fisher's Test. The *p*-values for Fisher's Exact Test can be laborious to calculate by hand for larger sample sizes and more moderate probabilities, but then these are the cases that are ideal for the *z*-test. Modern statistical software should make both tests easy to perform.

Regardless of which hypothesis test is utilized, users should perform the confidence interval described in the next section to further understand the nature of any differences.

2.1.2 Confidence intervals of difference between proportions

The hypothesis tests described in the previous sections give no information other than whether the null hypothesis of "no difference between the instruments" should be rejected. However, a

statistical interval can address at least in part which instrument performed better, the magnitude of the difference, and how practically significant the difference is.

Let $z_{1-\frac{\alpha}{2}}$ be the 100(1-α/2) percentile point of the standard normal distribution. Then a widely used 100(1-α) % level confidence interval for $P_1 - P_2$ based on a normal approximation is:

$$(p_1 - p_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}. \qquad (2.3)$$

Unfortunately, the interval shown in Equation (2.3), which is also known as a Wald-type interval, has been shown not to achieve its desired coverage of the true difference in many cases. Agresti and Caffo [6] propose a modification of the Wald interval as follows: Produce a "modified" pseudo-data set by adding one success and one failure to each instrument's results and plug the statistics from the resulting pseudo-data set into Equation (2.3). In more detail, let the "modified" sample proportions be $\tilde{p}_1 = \frac{x_1+1}{n_1+2}$, and $\tilde{p}_2 = \frac{x_2+1}{n_2+2}$. Then a 100(1-α) % level confidence interval for $P_1 - P_2$ with good performance qualities is:

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}. \qquad (2.4)$$

Agresti and Caffo do not claim that this interval satisfies any theoretical optimality criteria; rather, they demonstrate from extensive simulation studies that it attains the proper coverage probabilities for virtually any choice of n_1, n_2, P_1, and P_2 that they examined.

This interval does not have a one-to-one correspondence with the normal approximation-based *z*-test hypothesis testing procedure described in the previous section. **In this case, whether this confidence interval contains or does not contain zero does not correspond exactly to that hypothesis test accepting or rejecting, respectively, the null hypothesis of equal proportions.** Each procedure was chosen to satisfy its particular purpose, although they will agree in the vast majority of cases. Agresti and Caffo [6] do not advocate using the interval (2.4) as an implicit hypothesis test instead of the *z*-test in (2.1) because doing so would be needlessly conservative in cases where the hypothetically common proportion of success $P_1 = P_2$ in the null hypothesis is close to 0 or to 1.

Example.

Let us return to our previous example of the two instruments, where Instrument 1 successfully detects a source 14 times out of 20 trials, while Instrument 2 detects a source 10 times out of 20 trials. Adding pseudo-observations of one success and one failure to each instrument's data produces "modified" sample proportions $\tilde{p}_1 = \frac{15}{22} = 0.68$, and $\tilde{p}_2 = \frac{11}{22} = 0.5$. Using the formula in (2.4), a 90 % confidence interval for the difference $P_1 - P_2$ is

$$(0.68 - 0.5) \pm z_{1-\frac{0.1}{2}} \sqrt{\frac{0.68(1-0.68)}{22} + \frac{0.5(1-0.5)}{22}}$$

$= 0.18 \pm 1.645\ (0.146) = 0.18 \pm 0.24 = (-0.06, 0.42)$.

Note that this confidence interval contains zero at the 90% confidence level. However, we do see that the range of plausible values for the difference $P_1 - P_2$ includes numbers as large as 0.42 on the positive end, but only extends to -0.06 on the negative end. Thus, it is much more plausible for P_1 to be a little larger than P_2, rather than the reverse, as would be expected given p_1 was larger than p_2.

2.1.3 Sample Size Requirements

2.1.3.1 Sample Sizes for testing equality of proportion

In terms of practical significance, it may be desired to specify how many trials are needed to detect a significant difference at a given significance level. For this to make sense, what must be specified first is the minimum difference the test is required to discover, as well as the minimum power and maximum *Type I* error needed for the Hypothesis test. We will follow the treatment of Chapter 3 of Fleiss [7]. In a problem of this type, one instrument has a relatively known success rate of P_1, and a comparison study with a new instrument is worth doing only if it is able to detect a difference between the old instrument and a new instrument with rate P_2.

Suppose we need such a test to have significance level α and power $1-\beta$, for a desired α and β. We will assume equal sample sizes for both instruments.

Let $\bar{P} = (P_1 + P_2)/2$. Let

$$n' = \frac{\left(z_{1-\frac{\alpha}{2}} \sqrt{2\bar{P}(1-\bar{P})} - z_{\beta} \sqrt{P_1(1-P_1) + P_2(1-P_2)}\right)^2}{(P_2 - P_1)^2} \qquad (2.5)$$

Then n' is the same size needed for each sample.

According to Fleiss, there are studies that say the above formula underestimates the needed sample size needed to achieve the desired power. This should be especially relevant when a continuity correction is utilized in the test statistic (it is not in this document). He lists a formula that uses a continuity correction in the formula to adjust the needed sample size, with an accompanying close approximation that we list here:

$$n'' = n' + \frac{2}{|P_2 - P_1|} \qquad (2.6)$$

Example

Refer to the previous example again. Let us presume the sample proportions were the actual population proportions, i.e. $P_1 = 0.7$ and $P_2 = 0.5$. Suppose we desired a hypothesis test with significance level $\alpha = 0.10$ and power $(1-\beta) = 0.75$. According to the formula above, the sample size for each instrument needed is $n' = 64$. If we wanted to even be more conservative, the augmented formula gives a sample size of $n'' = 74$. It is no wonder that our hypothesis test in the first example failed to reject the null hypothesis, as the actual sample of 20 is much smaller.

Note that if we drastically lowered the needed power of the test $(1-\beta)$ to a mere 0.50 and weakened the significance level to $\alpha = 0.20$, then a sample size of $n' = 20$ is needed.

The table below shows the sample sizes needed, showing both n' (Formula (2.5) above) and n'' (from Formula (2.6)), to satisfy several values of α and β, for some combinations of P_1, and P_2. All tables show how the sample size needed depends on the significance level and power required. Comparing the first two sub-tables shows how the required sample sizes can depend on the magnitudes of P_1 and P_2, even though $P_1 - P_2$ is the same for both sub-tables. Finally, the third sub-table shows that the sample sizes needed can be much smaller if one instrument has a very sizeable advantage over the other.

α	1-β	P_1	P_2	n'	n''
0.10	0.60	0.7	0.5	43	53
0.10	0.70	0.7	0.5	56	66
0.10	0.80	0.7	0.5	73	83
0.05	0.60	0.7	0.5	59	69
0.05	0.70	0.7	0.5	73	83
0.05	0.80	0.7	0.5	93	103

α	1-β	P_1	P_2	n'	n''
0.10	0.60	0.9	0.7	29	39
0.10	0.70	0.9	0.7	37	47
0.10	0.80	0.9	0.7	49	59
0.05	0.60	0.9	0.7	39	49
0.05	0.70	0.9	0.7	49	59
0.05	0.80	0.9	0.7	62	72

α	1-β	P_1	P_2	n'	n''
0.10	0.60	0.95	0.6	10	16
0.10	0.70	0.95	0.6	13	19
0.10	0.80	0.95	0.6	17	22
0.05	0.60	0.95	0.6	14	20
0.05	0.70	0.95	0.6	17	23
0.05	0.80	0.95	0.6	21	27

Table 1: Sample sizes needed for each group to satisfy several values of α and β, for some combinations of P_1 and P_2.

2.1.3.2 Sample Sizes for estimating the difference of proportion

Suppose there is a need to estimate the difference between the two instrument means $P_1 - P_2$ to within a margin of *H* with probability *1-α*. A reasonable magnitude for *H* depends on the context and application of the experiment. Suppose there is prior knowledge about the approximate magnitude of P_1 and P_2. Presuming equal sample sizes, the number of trials for each sample required is [3]:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}}{H}\right)^2 [\, P_1(1-P_1) + P_2(1-P_2)]. \qquad (2.7)$$

If there is no prior knowledge of the approximate magnitude of P_1 and P_2, since the sample size formula is maximized by $P_1 = P_2 = 0.5$, then a conservative estimate of the number of trials for each sample required can be obtained plugging in $P_1 = P_2 = 0.5$:

$$n = \frac{1}{2}\left(\frac{z_{1-\frac{\alpha}{2}}}{H}\right)^2 \qquad (2.8)$$

If P_1 and P_2 are much different from 0.5, then this sample size may be much larger than needed.

Example.

Refer again to the example used in the previous sections. Suppose we had prior knowledge that the two instrument means were approximately $P_1 = 0.7$ and $P_2 = 0.5$. Suppose we needed to estimate the difference between the two instrument means $P_1 - P_2$ to within a margin of 0.2 with probability 0.90. Then the number of trials for each sample required is:

$$n = \left(\frac{1.645}{0.2}\right)^2 [0.7(1-0.7) +).5(1-0.5)] = 67.64\ (\,0.46) = 31.1 \approx 31.$$

We round *n*=31.1 to *n*=31 trials needed for each sample.

Suppose that we had no prior knowledge about P_1 and P_2, and used $P_1 = P_2 = 0.5$ as a conservative procedure. Since the estimates we used earlier are relatively close to, or even equal to 0.5, the resulting number *n* = $67.64\ (0.5) = 33.8$, which we round to *n* = 34, is not very different.

2.2 Comparing multiple instruments with binary response

2.2.1 Chi-square test for equality of proportions

Suppose that there are *k* > 2 instruments that when tested yield Detect/No Detect binary responses. Using the same notation as the previous section, suppose that instrument i has x_i

successes in n_i trials, for a sample proportion of $p_i = x_i / n_i$, for *i=1,…,k.* How do we compare these proportions?

The first step in the analysis is to test whether all instruments have the same probability of success. Here, the null hypothesis has the form:

H_0 : All instruments have the same proportion of success.

The alternative hypothesis is:

H_A : Not all the instruments have the same proportion of success.

The test used is an example of a *chi-square* goodness of fit test on a contingency table [2,8]. If all proportions are the same, this common proportion can be estimated by pooling all the trials:

$$\bar{p} = \sum_{i=1}^{k} x_i \; / \sum_{i=1}^{k} n_i \qquad (2.9)$$

Under the null hypothesis, the expected number of successes for the i^{th} instrument is $s_i = n_i p_i$, with corresponding expected number of failures $r_i = n_i(1 - p_i)$. Our test statistic is the sum of the squared deviation of the numbers of successes and failures each from its expected number, and each normalized by the expected number:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - s_i)^2}{s_i} + \sum_{i=1}^{k} \frac{(n_i - x_i - r_i)^2}{r_i}. \qquad (2.10)$$

Another way to write the same formula that may be more familiar to those experienced with contingency table tests is:

$$\chi^2 = \sum_{all\ cells} \frac{(f_o - f_e)^2}{f_e}, \qquad (2.11)$$

Here f_e is the expected frequency, and f_o is the observed frequency of each particular cell out of the *2k* number of cells in a contingency table.

Given a significance level of α, H_0 is rejected if the test statistic χ^2 is larger than the 100(1- α) percentile of the *chi-squared* distribution with *k-1* degrees of freedom. Unlike the hypothesis tests previously discussed in this document, this is a one-sided test in that only the area under the right tail of the reference distribution is the rejection region of the test. The plot below depicts a possible chi-square distribution (with 5 degrees of freedom); the critical point for that distribution (11.07) is that point where 95% of the area under the density curve lies to the left, and 5% of the area to the curve lies to the right. If the χ^2 statistic is greater than 11.07, then the null hypothesis of no instrument differences is rejected. Tables of critical values are available in most statistics textbooks as well as many online resources. Alternatively, one can use a statistical software to compute a *p*-value.

Note: It is well-known that for the case of 2 instruments, this χ^2 test is equivalent to the *z*-test comparing two proportions shown earlier in Section 2.1.1.1.

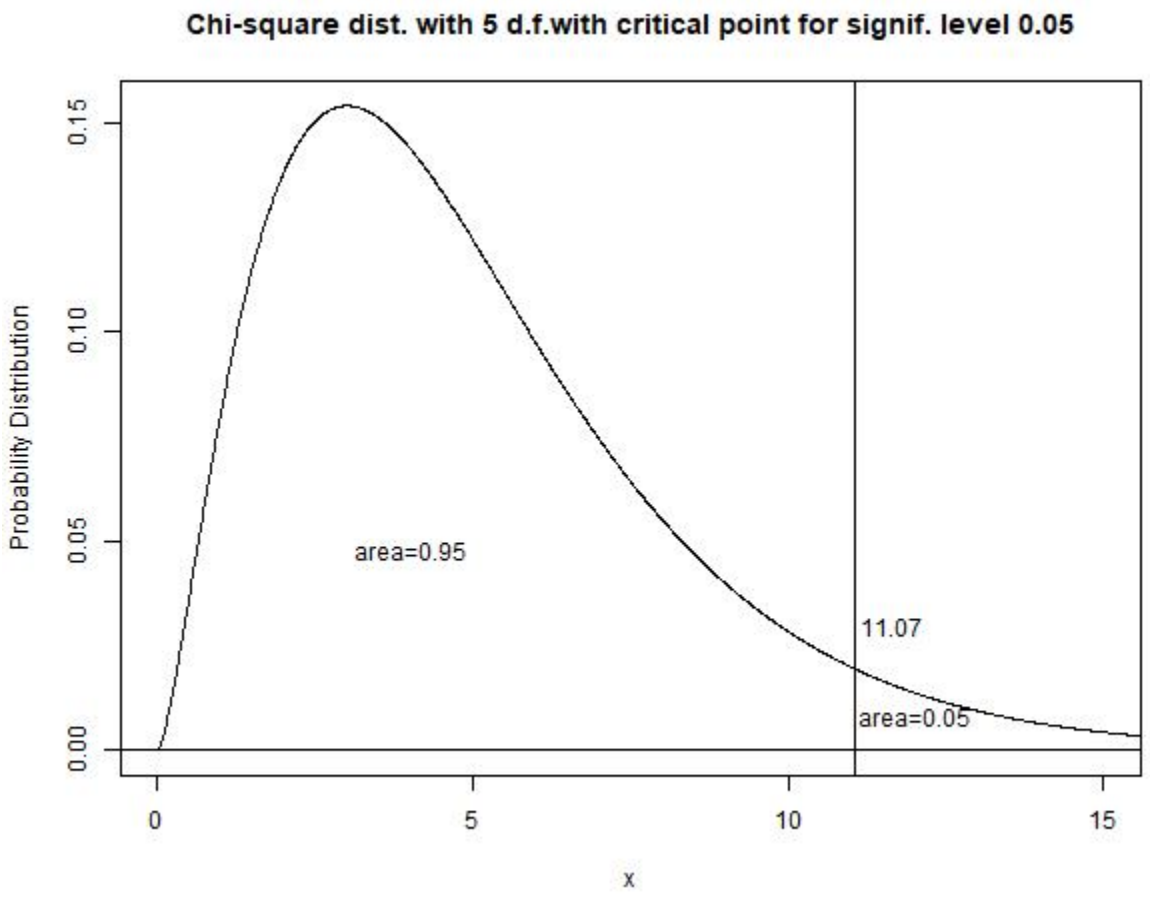

Figure 2: A chi-square distribution with 5 degrees of freedom, shown with its critical point (11.07) for significance level α = 0.05. If the χ^2 statistic is greater than the critical point, then the null hypothesis of no instrument differences is rejected.

Example

Refer to our previous example involving two instruments. Suppose that in addition to the two instruments there was a third instrument that had 6 successes in 20 trials.

Instrument	**1**	**2**	**3**	**Row totals**
Successes	14	10	6	30
Failures	6	10	14	30
Column Totals	20	20	20	60

For this example, $\bar{p} = \frac{14+10+6}{60} = \frac{1}{2}$, and since each instrument had 20 trials, the expected number of successes for each instrument under the null hypothesis is 10. The expected number of failures for each instrument under the null hypothesis is also 10. So, the test statistic for this example is

$$\chi^2 = \frac{[\,(14-10)^2 + (10-10)^2 + (6-10)^2 + (6-10)^2 + (10-10)^2 + (14-10)^2]}{10} = 6.4.$$

The reference distribution is the *chi-square* distribution with *k-1=2* degrees of freedom. The χ^2 statistic of 6.4 is larger than the critical point of 5.99 for significance level α=0.05, so the null hypothesis of "all instruments being the same" is rejected. Statistical software shows that the *p*-value of 6.4 is 0.04.

2.2.2 Comparing Multiple proportions: Marascuilo procedure

If running a multiple proportions contingency table hypothesis test leads you to conclude that not all instruments perform equally, it still does not inform you which instruments are better or worse. The instruments can be ranked by their respective proportions of success, but that does not indicate which proportions are significantly different. The Marascuilo procedure [2] is a way to simultaneously test the differences between all pairs of instruments.

Once again suppose that there are *k* instruments, and that instrument i has x_i successes in n_i trials, for a sample proportion of $p_i = x_i / n_i$, for i=1,…,*k*.

For every pair of sample proportions, compute the absolute value of the difference $|p_i - p_j|$ for every $i \neq j$ in 1, …, *k*. This will be the test statistic for that pair. There will be $k(k-1)/2$ different proportion differences.

Next for every pair of $i \neq j$, for a chosen overall significance level α, the critical value of that test statistic will be

$$r_{ij} = \sqrt{\chi^2_{1-\alpha,k-1}} \sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}} \qquad (2.12)$$

Here $\chi^2_{1-\alpha,k-1}$ is the 100(1-α) percentile point of a *chi*-squared distribution with *k-1* degrees of freedom. As before the numerical value of $\chi^2_{1-\alpha,k-1}$ can be found in the literature, online, or using a statistical software.

For each of the *k*(*k*-1)/2 pairs of proportions, if $|p_i - p_j| > r_{ij}$, then those two proportions are significantly different.

Example

Returning to the same example used in the contingency table test, there are three instruments that have 14, 10, and 6 successes out of 20 trials for each experiment. Thus, $p_1 = 0.7, p_2 = 0.5$, and $p_3 = 0.3$. There are $k(k\text{-}1)/2 = 3$ possible pair differences.

A significance level of α=0.05 will be used. Then $\chi^2_{1-\alpha,k-1} = \chi^2_{0.95,2} = 5.99$.

- For Instruments 1 and 2, $|p_1 - p_2| = 0.2 < r_{12} = \sqrt{5.99}\sqrt{\frac{0.7(1-0.7)}{20} + \frac{0.5(1-0.5)}{20}} = 0.37$, so Instruments 1 and 2 <u>are not</u> significantly different.

- For Instruments 1 and 3, $|p_1 - p_3| = 0.4 > r_{13} = \sqrt{5.99}\sqrt{\frac{0.7(1-0.7)}{20} + \frac{0.3(1-0.3)}{20}} = 0.35$, so Instruments 1 and 3 <u>are</u> significantly different.

- For Instruments 2 and 3, $|p_2 - p_3| = 0.2 < r_{23} = \sqrt{5.99}\sqrt{\frac{0.5(1-0.5)}{20} + \frac{0.3(1-0.3)}{20}} = 0.37$, so Instruments 2 and 3 <u>are not</u> significantly different.

3 Continuous measurements

The binary nature of the data discussed earlier can sometimes make it problematic to find definitive answers with small sample sizes. If there is continuous data available, it can be much more informative than simple counts of (0/1) trials.

As a hypothetical example, suppose that instead of (Detect/No Detect) responses from an instrument from a fixed distance from a source, there are multiple trials of the following experimental procedure: a source and a detector are gradually moved closer together until the instrument detects the source, with the detection distance recorded. Here we presume the instrument will always eventually detect the source, otherwise the result can be denoted as zero or some previously agreed to quantity. In practice, such numerical measurements are never truly continuous due to the limited resolution of measurement, and in this scenario, it would be practical to move the detector and instrument closer in steps rather than as a continuous process.

Another example of continuous measurements would be the time till detection of the source by the detector. (This test would be more likely for a chemical detector or if a radiation detector integrates the signal until there are enough counts above background to be detected.)

The increased complexity of continuous data over binary data brings an accompanying complexity in how instruments can be compared. Figure 3 contains some schematics of how instrument performances can differ. The upper left plot depicts hypothetical data of an experiment where the measurements taken from Instrument 1 and Instrument 2 have the same mean and variability. The upper right plot depicts an experiment where the data from the two instruments have the same variability, but different means. In the lower left plot, the measurements taken on both Instrument 1 and Instrument 2 are centered on the same mean, but those of Instrument 2 are much more variable. The lower right plot shows a situation where the measurements from Instrument 2 are both larger on the average and more variable than those of Instrument 1. This section on comparing instruments that produce continuous measurements will concentrate first on location-based methods for comparing instruments based on analysis of their mean measurements, before moving to methods for comparing instruments by examining the respective variances of the instrument measurements.

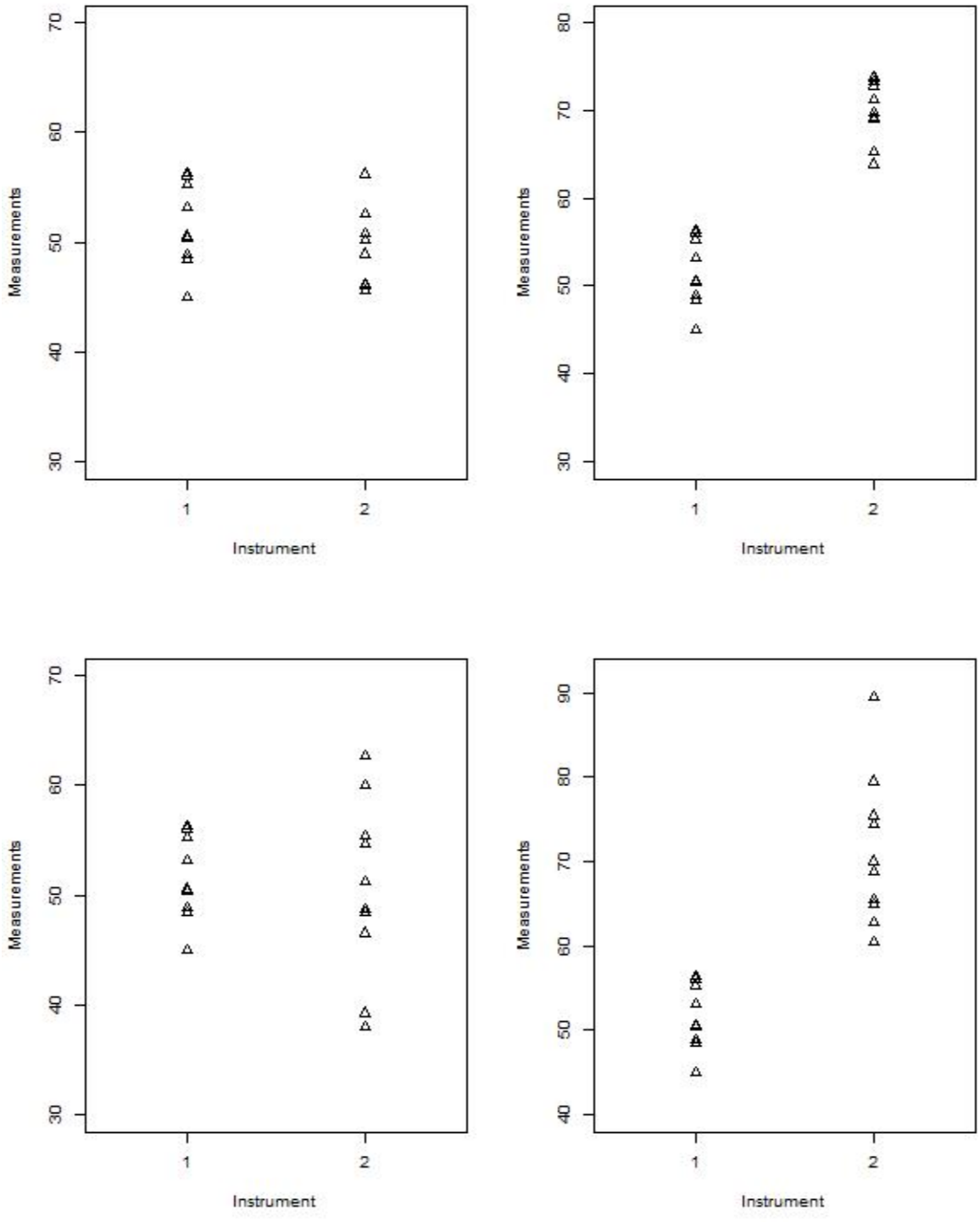

Figure 3: This set of graphs are schematics of some different ways that the measurements of two instruments can compare with each other. In the two graphs on the left, the two instruments have the same measurement means, while in the graphs on the right, the two instruments have different measurement means. In the graphs on the top row, the two instruments have the same measurement variances, while in the graphs on the bottom row the two instruments have different measurement variances.

3.1 Comparing two instruments with continuous measurements

Suppose we want to compare two instruments where the data are in the form of numeric measurements. For instance, they may be from multiple trials of the distance till first detection of the source, as described in the previous section. For the rest of this section we have the following notation for the observations:

Suppose for Instrument 1, there are $m \geq 2$ measurements denoted as: $x_1. \ldots, x_m$.

Suppose for Instrument 2, there are $n \geq 2$ measurements denoted as: $y_1, \ldots, y_n$.

Let $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ be the mean of the measurements for Instrument 1.

Let $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ be the mean of the measurements for Instrument 2.

Let $s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$ be the sample variance of the measurements for Instrument 1.

Let $s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$ be the sample variance of the measurements for Instrument 2.

3.1.1 Hypothesis tests for comparing two instruments

When comparing two instruments where the data are in the form of numeric measurements, the simplest way to compare is to test whether they have the same mean. The classical test is the two-sample Student's *t*-test, of which we describe two variations here.

Suppose that the measurements $x_1. \ldots, x_m$ come from a distribution with mean μ_1, and the measurements $y_1, \ldots, y_n$ come from a distribution with mean μ_2.

For the *t*-test, the null and alternative hypotheses will be:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

3.1.1.1 Student's *t*-test: Equal variance case.

Suppose it is presumed or verified (see Section 3.3 on comparing variances) that the measurements for both instruments have approximately the same variance. Then we can use a *pooled* estimator for the variance that pools the sample variances from both instruments:

$$s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}.$$

Our test statistic is

$$T = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

For a given significance level α, reject the null hypothesis that the two instruments are equal if

$$|T| > t_{1-\alpha/2,\nu}\,,$$

Where $t_{1-\alpha/2,\nu}$ is the 100(1- α) percentile of the Student's *t* distribution with *v* degrees of freedom. For this test, $\nu = m + n - 2$.

Tables of *t*-test critical values are widely available in any statistics textbook, as well as online resources such as the NIST-Sematech handbook [2].

3.1.1.2 Student's *t*-test: Unequal variance case.

If the two variances are presumed or tested not to be equal, or if there is doubt about the equality of the variances, then one should not utilize the pooled variance estimate. In that case the preferred form of a *t*-test statistic is

$$T = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}} \tag{3.1}$$

For a given significance level *α*, reject the null hypothesis that the two instruments are equal if

$$|T| > t_{1-\alpha/2,\nu}\,,$$

Here the degrees of freedom used is given by the Welch-Satterthwaite formula [2]:

$$\nu = \frac{\left(\frac{s_x^2}{m}+\frac{s_y^2}{n}\right)^2}{\left(\frac{s_x^2}{m}\right)^2 \Big/ (m-1) \; + \; \left(\frac{s_y^2}{n}\right)^2 \Big/ (n-1)} \tag{3.2}$$

The calculated estimate of *v* can be rounded to the nearest or to the next lowest integer; alternatively, many statistical software packages have *t*-distribution functions that do not require the degrees of freedom to be integers.

3.1.1.3 Student's *t*-test Example:

Suppose for each of two instruments, a sample of 15 measurements of the distance till detection was measured in cm.

The measurements for Instrument 1 $(x_1, \ldots, x_{15})$:

91 95 107 105 102 85 88 92 101 99 102 85 114 91 95

For Instrument 2, the measurements are $(y_1, \ldots, y_{15})$:

93 99 97 101 70 83 97 100 91 73 90 86 95 70 87

Let's proceed with a Student's *t*-test with significance level α=0.05.

Thus, $m = n = 15$, $\bar{x} = 96.8$, $\bar{y} = 88.8$, $s_x^2 = 71.17$, and $s_y^2 = 112.6$.

Equal Variance test

Let us first presume that the variances of the distribution are equal. Then the pooled estimate of variance is

$$s_p^2 = \frac{(15-1)71.17 + (15-1)112.6}{15+15-2} = 91.885$$

Thus, $s_p = \sqrt{91.885} = 9.586$.

Our test statistic is

$$T = \frac{96.8 - 88.8}{9.586\sqrt{\frac{1}{15}+\frac{1}{15}}} = \frac{8.0}{3.5} = 2.29.$$

The critical point $t_{1-\frac{\alpha}{2},m+n-2} = t_{0.975,28}$ is $2.05 < |2.29|$, so the null hypothesis of equal mean is rejected.

Unequal Variance test

In the unequal variance *t*-test, the denominator of the *T* statistic is

$$\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}} = \sqrt{\frac{71.17}{15}+\frac{112.6}{15}} = 3.5$$

The approximate degrees of freedom is estimated by the Satterthwaite formula. Plugging numbers in the formula gives

$$v = \frac{150.1}{5.633} = 26.6,$$

which we can round down to $v = 26$. Thus,

$$|T| = \frac{8.0}{3.5} = 2.29 > t_{0.975,27} = 2.06,$$

so the null hypothesis is rejected in this case as well.

3.1.1.4 Student's *t*-test: Paired tests

There can be situations where data measurements from different instruments can be paired. For instance, suppose there were a series of different configurations of radioactive sources, and the distance till detection was measured simultaneously for two instruments for each configuration. Suppose the measurements for all instruments vary depending on the configuration, but that the differences between instruments are consistent across configurations and can be logically grouped into a sample. It is then more efficient and powerful to reduce the data to a single sample of paired differences $d_1 = x_1 - y_1, \ldots, d_k = x_k - y_k$.

The paired Student's *t* test essentially takes the differences $d_1, \ldots, d_k$ and performs a one sample *t*-test of whether the differences have mean zero.

Let $\bar{d}$=$\sum_{i=1}^{k} d_i \,/k$ be the mean of the sample of differences, and $s_d^2 = \frac{1}{(k-1)} \sum_{i=1}^{k} \left(d_i - \bar{d}\right)^2$ be the sample variance of the differences. The test statistic is

$$T = \frac{\bar{d}}{s_d/\sqrt{k}}.$$

For a selected significance level α, the null hypothesis is rejected if

$$|T| > t_{1-\frac{\alpha}{2},k-1}\,.$$

Example of paired *t*-test:

Suppose that a performance measure for Instrument 1 and for Instrument 2 is measured simultaneously on $k = 20$ different trials. Presume that it makes sense to look at the pooled set of paired differences. The measurements x_i for Instrument 1, the measurements y_i for Instrument 2, as well as the differences $d_i = x_i - y_i$ are listed in Table 2 below. Calculations show

$\bar{d} = 6.6$, and $s_d^2 = 49$. Thus, $T = \frac{6.6}{7/\sqrt{20}} = 4.2$.

With a significance level of $\alpha = 0.05$,

$t_{1-\frac{\alpha}{2},k-1} = t_{0.975,19} = 2.1$, so the null hypothesis of no difference between the instruments is rejected. In fact, statistical software shows that the associated *p*-value is 0.0005, which should be very strong evidence for a difference in instruments.

Note that if one applied either of the unpaired t-tests described above with this data, the p-value would be around 0.265, so a significant difference between instruments would not be detected. That is because the instrument difference would be hidden amidst the large

differences between measurements for each instrument. This highlights the importance of knowing the background behind one's data and tailoring the analysis accordingly.

x_i	y_i	$d_i = x_i - y_i$
102	89	13
67	60	7
109	101	8
97	96	1
72	73	-1
46	39	7
83	70	13
86	92	-6
58	35	23
65	64	1
73	67	6
92	90	2
105	93	2
96	84	12
92	95	-3
60	53	7
88	74	14
78	75	3
65	53	12
64	63	1

Table 2: List of paired data

3.1.2 Confidence intervals for differences between means

As stated earlier in the section on binary results, readers are urged to proceed beyond hypothesis testing and to examine confidence intervals summarizing the difference between instruments' performance. The same data scenario will be repeated:

Suppose that the measurements $x_1. \ldots, x_m$ for Instrument 1 come from a distribution with mean μ_1, and the measurements $y_1, \ldots, y_n$ come for a distribution with mean μ_2. It is desired to find a confidence interval for the difference $\mu_1 - \mu_2$.

We will describe three different confidence intervals to use depending on the assumptions and data scenario. They will be closely linked to the three different kinds of Student's *t* test described earlier. For the first two intervals, we again use the following quantities:

Let $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ be the mean of the measurements for Instrument 1.

Let $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ be the mean of the measurements for Instrument 2.

Let $s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$ be the sample variance of the measurements for Instrument 1.

Let $s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$ be the sample variance of the measurements for Instrument 2.

3.1.2.1 Confidence intervals for differences between means: Equal variance case

Suppose it is presumed or verified (see Section 3.3 on comparing variances) that the measurements for both instruments have approximately the same variance. Then we can use a pooled estimator for the variance that pools both sample variances:

$$s_p^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}.$$

For a given confidence level α, the two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x} - \bar{y} \pm t_{1-\frac{\alpha}{2},m+n-2}\, s_p \sqrt{\frac{1}{m} + \frac{1}{n}}.$$

3.1.2.2 Confidence intervals for differences between means: Unequal variance case:

For a given confidence level α, the two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\bar{x} - \bar{y} \pm t_{1-\frac{\alpha}{2},v} \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}},$$

where v is the effective degrees of freedom estimated by the Welch-Satterthwaite approximation [2].

Examples for confidence intervals for differences between means

Refer to the data set in the *t*-test example in section 3.2. Recall that $m = n = 15$, $\bar{x} = 96.8$, $\bar{y} = 88.8$, $s_x^2 = 71.17$, $s_y^2 = 112.6$, and $s_p = 9.586$.

Example: Equal variances confidence interval

A 95 % confidence interval for the difference between instrument means, assuming equal variances, is

$$96.8 - 88.8 \pm t_{.975,28}\ 9.586\sqrt{\frac{1}{15}+\frac{1}{15}} = 8.0 \pm 7.17 = (0.83, 15.17).$$

A 95 % confidence interval for the mean difference between instruments, assuming unequal variances, is

$$96.8 - 88.8 \pm t_{.975,26}\sqrt{\frac{71.17}{15}+\frac{112.6}{15}} = 8.0 \pm 7.19 = (0.81, 15.19).$$

The interval is slightly wider than the equal-variance interval. Note that for this example, we rounded down the Satterthwaite approximation for effective degrees of freedom $v = 26.6$ to $v = 26$ (this would be appropriate for those using tables of the Student's t distribution). Many statistical software packages will not round v because they can calculate percentiles of Student's t distribution with non-integer degrees of freedom.

3.1.2.3 Confidence intervals for differences between means: Paired case

Assume that the data from the two observations are paired in a way that it makes sense to look at the sample of paired differences $d_1 = x_1 - y_1, \ldots, d_k = x_k - y_k$.

Let $\bar{d} = \sum_{i=1}^{k} d_i\,/k$, and $s_d^2 = \frac{1}{(k-1)}\sum_{i=1}^{k}\left(d_i - \bar{d}\right)^2$.

For a given confidence level α, the two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\bar{d} \pm t_{1-\frac{\alpha}{2},k-1}\frac{s_d}{\sqrt{k}}.$$

Example.

Refer to the data contained in the table on paired t-tests in Section 1.3.1.3. A 95 % confidence interval for the difference between instrument means is

$$\bar{d} \pm t_{1-\frac{\alpha}{2},k-1}\frac{s_d}{\sqrt{k}} = 6.6 \pm \frac{7}{\sqrt{20}} = (3.32, 9.88).$$

3.1.3 Sample size requirements for estimating mean difference

Suppose it is needed to estimate the difference between the two instrument means $\mu_1 - \mu_2$ to within a margin of H with probability 1-α. Suppose also that the variances $\sigma_1^2$and σ_2^2 are known, or at least can be estimated or approximated. Presuming equal sample sizes, the number of measurements for each sample required is:

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}}{H}\right)^2 (\sigma_1^2 + \sigma_2^2)$$

(Mendenhall and Sincich [3]).

Example.

The variances $\sigma_1^2$and σ_2^2 are often not known and have to be estimated from prior knowledge or approximated. Let us go back to our previous data example with the unpaired *t*-test from section 3.1.2 and presume that we had some prior knowledge that σ_1^2 is approximately 100 and σ_2^2 is approximately 120. If it is needed to estimate the difference between the two instrument means $\mu_1 - \mu_2$ to within 10 cm with probability 0.95, then the number of measurements for each sample required is:

$n = \left(\frac{1.96}{10}\right)^2 (100 + 120) = 8.45$, which we round up to $n = 9$.

Suppose there was a more stringent objective of estimating the difference between the two instrument means $\mu_1 - \mu_2$ to within 5 cm with probability 0.95; then the number of measurements for each sample required is $n = \left(\frac{1.96}{5}\right)^2 (100 + 120) = 33.8$, which we round to $n = 34$.

3.2 Comparing multiple instruments with continuous data

3.2.1 Hypothesis test: ANOVA test

Suppose there are multiple instruments where the performance output is in the form of continuous numerical measurements. There will be k instruments. Instrument i has n_i measurements $y_{i1}, \ldots, y_{in_i}$ that originate from a distribution with mean μ_i. Experiment designers should know that it is advantageous for the sample sizes $n_1, \ldots, n_k$ to be equal.

A one-way Analysis-of-variance (ANOVA) hypothesis test to compare how the instruments perform will compare the means of their measurements.

The model underlying the ANOVA is that the j^{th} observation from the i^{th} instrument can be written as:

$$y_{ij} = \mu + a_i + e_{ij}.$$

This model decomposes each observation into three components: an overall mean μ, an instrument effect a_i (the deviation of the i^{th} Instrument mean from the grand mean), and a residual e_{ij}. The ANOVA model presumes that the residuals are independent, approximately normally distributed with mean 0, and have approximately the same variance for each instrument.

The null hypothesis that all instruments perform the same will be

$$H_0: \mu_1 = \cdots = \mu_k$$

The alternative hypothesis is H_A: Not all the μ_i are the same.

In terms of the ANOVA model, the null hypothesis H_0 states that for each instrument $a_i = 0$.

Denote the i^{th} instrument mean as $\overline{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$.

Let $N = \sum_{i=1}^{k} n_i$ be the total number of observations.

The grand mean is $\overline{y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$.

The sum of squares due to Factor, which in our context is the Instrument is

$$SSF = \sum_{i=1}^{k} n_i \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2$$

The sum of squares due to Error (residual) is

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\cdot})^2 .$$

The mean squares are the sum of squares divided by the associated degrees of freedom:

MSF=SSF/(k-1) and *MSE=SSE/(N-k)*.

The test statistic is *F=MSF/MSE*.

Under the null hypothesis of no instrument effects, *MSF* and *MSE* would be essentially estimating the same quantity of variation and would follow an $F_{k-1,N-k}$ distribution, which is an *F* distribution with *k*-1 and *N*-*k* degrees of freedom. Under the alternative hypothesis, the existence of significant instrument effects should cause *MSF* to be larger than *MSE*. **Thus, the null hypothesis is rejected at significance level α if the test statistic *F* is larger than the 100(1-α) percentile point of the $F_{k-1,N-k}$ distribution.** Critical points of the *F* distribution for selected significance levels and various degrees of freedom are available in many statistics textbooks and online resources. Statistical software will usually provide a *p*-value associated with the result.

Example

Suppose for each of three instruments, a sample of 5 measurements of the distance till detection was measured in cm.

The measurements for Instrument 1 ($y_{1,1}, \ldots, y_{1,5}$):

67 88 72 68 69

The measurements for Instrument 2 ($y_{2,1}, \ldots, y_{2,5}$):

99 83 102 101 76

The measurements for Instrument 3 ($y_{3,1}, \ldots, y_{3,5}$):

108 132 114 108 124

Doing the ANOVA calculations show that for the sums of squares *SSF*= 4954.5 and *SSE*= 1314.4. The relevant degrees of freedom are *k*-1= (3-1), and *N*-*k* = (15-3). Thus, the *F*-statistic is

$$F = \frac{MSF}{MSE} = \left(\frac{SSF}{2}\right)/\left(\frac{SSE}{12}\right) = 22.61.$$

For significance level $\alpha = 0.05$, the relevant critical point of the $F_{2,12}$ distribution is 3.9, which is much smaller than the observed *F*-statistic, so the null hypothesis of no mean instrument differences is rejected.

3.2.2 Multiple Comparisons

3.2.2.1 Tukey Procedure

If the *F*-test rejects the null hypothesis of no instrument difference, it does not show how or which instrument means are different from each other. If one particular pair of instruments are singled out for comparison, then a confidence interval for that difference can be calculated. However, it is not appropriate to compute 100(1-α) % intervals for all or multiple pairwise comparisons between means because the 100(1-α) % confidence level will not hold for the set of comparisons.

One can use the Tukey method to estimate intervals for all pairwise mean differences and an overall confidence level of 100(1-α) % for the set of intervals. Suppose that there are *k* instruments, and all the data set up, notation, and quantities are the same as in the previous section on the ANOVA test. Assume all instruments have the same sample size $n_i = n$, so the total number of observations is $N = kn$.

Let *MSE* be the same mean squares as in the previous section. For a given overall significance level *α*, for every pair of i ≠ j, the set of simultaneous 100(1- α) % confidence intervals for the instrument mean differences includes

$$\overline{y_{i.}} - \overline{y_{j.}} \pm q_{1-\alpha;k,N-k} \sqrt{\frac{MSE}{n}}.$$

Here $q_{1-\alpha;k,N-k}$ is the 100(1- α) percentile point of the *studentized range distribution* with parameters *k* and *N-k*. While not as well-known as the *t* or *F* distributions, tables of the studentized range distribution are available in many statistics textbooks and online resources. Many statistical software packages also contain a function for percentile points of the distribution for given parameters

Example.

Let's return to the same data used in the ANOVA test example in the previous section. For that data set, MSE= 1314.4/12 = 109.53, so $\sqrt{\frac{MSE}{n}} = \sqrt{\frac{109.53}{5}} = 4.68$. The multiplier for a set of 95 % simultaneous confidence intervals for the mean differences is $q_{0.95;3,12} = 3.77$, so

$q_{1-\alpha;k,N-k} \sqrt{\frac{MSE}{n}} = 3.77\ (4.68) = 17.66$. If we denote the mean for Instrument *i* as μ_i,, the set of intervals for the mean instrument differences are:

$$\mu_2 - \mu_1\colon\ 19.4\ \pm 17.66 = (1.74, 37.06)$$

$$\mu_3 - \mu_1\colon\ 44.4 \pm 17.66 = (\ 26.74, 62.06)$$

$$\mu_3 - \mu_2\colon\ 25.0 \pm 17.66 = (7.34, 42.66).$$

All instrument means are significantly different from each other, with μ_3 the largest, μ_1 the smallest, and closest together are μ_1 and μ_2.

3.2.2.1 Tukey-Kramer Procedure for unequal sample sizes

The Tukey procedure in the previous section presumed equal sample sizes for each instrument. A modification called the Tukey-Kramer procedure can handle unequal sample sizes. Each interval width is calculated separately depending on the sample size.

Let *MSE* be the same mean squares as in the previous section. For a given overall significance level α, for every pair of $i \neq j$, the set of simultaneous confidence intervals for the instrument mean differences includes

$$\overline{y_{i.}} - \overline{y_{j.}} \pm q_{1-\alpha;k,N-k} \sqrt{\frac{MSE}{2}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

This procedure is known to be conservative in that its true confidence level is larger than 1-α when sample sizes are unequal, meaning that the intervals are longer than they need to be to achieve a true 100(1-α) % confidence level for the set of simultaneous intervals.

3.3 Comparing instrument variances with continuous data

3.3.1 Comparing variances from two instruments

Recall the data scenario from Section 3.1.1 when comparing two instruments where the data are in the form of numeric measurements. Suppose that the measurements $x_1. \ldots, x_m$ come from a distribution with variance σ_1^2, and the measurements $y_1, \ldots, y_n$ come from a distribution with variance σ_2^2. An *F*-test can be used to test whether the variances are equal.

For this *F*-test, the null and alternative hypotheses will be:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_A: \sigma_1^2 \neq \sigma_2^2$$

Let $s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2$ be the sample variance of the measurements for Instrument 1.

Let $s_y^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2$ be the sample variance of the measurements for Instrument 2.

Then the *F*-statistic takes the form $F = s_x^2 / s_y^2$.

If the significance level is α, then the null hypothesis that the two population variances are equal is rejected if $F < F_{\alpha/2,m-1,n-1}$ or $F > F_{1-\alpha/2,m-1,n-1}$, where $F_{\beta,m-1,n-1}$ is the $100(\beta)$th percentile point of the F distribution with $m-1$ and $n-1$ degrees of freedom. Critical points

of the *F* distribution for selected significance levels and degrees of freedom are available in statistical textbooks and software. Note that this is the same *F* distribution in Section 3.2.1, although this test and its relevant test statistic is different.

Example:

We return to the example from Section 3.1.1.1. The measurements for Instrument 1 are $(x_1, \ldots, x_{15})$:

91 95 107 105 102 85 88 92 101 99 102 85 114 91 95

For Instrument 2, the measurements are $(y_1, \ldots, y_{15})$:

93 99 97 101 70 83 97 100 91 73 90 86 95 70 87

Thus, $m = n = 15$, $s_x^2 = 71.17$, and $s_y^2 = 112.6$. The *F*-statistic $F = 71.17 / 112.6 = 0.632$.

If the significance level is, $\alpha = 0.05$, we find that the *F*-statistic is between the two critical points:

$$F_{\alpha/2,m-1,n-1} = F_{0.05,14,14} = 0.336 < F = 0.632 < F_{1-\alpha/2,m-1,n-1} = F_{0.975,14,14} = 2.98.$$

The null hypothesis of equal variances is not rejected because the null hypothesis is rejected only if the *F*-statistic is outside the interval formed by the two critical points.

This *F*-test assumes that both distributions are normally distributed. If that assumption is in doubt, one can utilize a test that is robust to assumptions. Such a test is the Levene Test, which is described in the next section.

3.3.2 Comparing variances from multiple instruments

The Levene test can be used to test if multiple samples come from distributions with the same variance [2]. Levene's test is more robust to assumptions than a more classical procedure like the Bartlett test, which assumes normal distributions for the data. This section describes a form of Levene's test that was proposed by Brown and Forsythe [9].

Suppose that there are *k* samples, each from its own distribution, with $k \geq 2$. For this test, the null and alternative hypotheses will be:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_A: \sigma_i^2 \neq \sigma_j^2 \text{ for some } 1 \leq i < j \leq k.$$

Suppose that from the *k* samples, sample i has n_i measurements $y_{i1}, \ldots, y_{in_i}$. Let $N = \sum_{i=1}^{k} n_i$ be the total number of observations. Define $\tilde{y}_i$ to be the median of the ith sample. Let

$$z_{ij} = |y_{ij} - \tilde{y}_i|.$$

Suppose $\bar{z}_{i\cdot}$ is the mean of the z_{ij} for the i^{th} group, and $\bar{z}_{\cdot\cdot}$ is the grand mean of all the z_{ij}. Then, the test statistic for Levene's test is

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i \ (\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i\cdot})^2}.$$

The null hypothesis of equal variances is rejected if $W > F_{1-\alpha\ ,k-1,n-k}$, which is the $100(1-\alpha\ /2)$ percentile point of the *F* distribution with *k-1* and *N-k* degrees of freedom.

Note that the original formulation of Levene's test used the sample mean instead of the median in the definition of the z_{ij}. Brown and Forsythe [9] proposed modifying the test to use the median or trimmed mean instead of the mean. This section uses the median form of the test because it should be robust to non-normality while still maintaining good efficiency [2].

Example

Let us return to the example from Section 3.2.1. Suppose for each of three instruments, a sample of 5 measurements:

The measurements for Instrument 1 $(y_{1,1}, \ldots, y_{1,5})$:

67 88 72 68 69

The measurements for Instrument 2 $(y_{2,1}, \ldots, y_{2,5})$:

99 83 102 101 76

The measurements for Instrument 3 $(y_{3,1}, \ldots, y_{3,5})$:

108 132 114 108 124

The respective medians of the three samples are 69, 99, and 114. Thus, we have

$(z_{1,1}, \ldots, z_{1,5}) = (2\ 19\ 3\ 1\ 0)$

$(z_{2,1}, \ldots, z_{2,5})$ = (0 16 3 2 23)

$(z_{3,1}, \ldots, z_{3,5})$ = (6 18 0 6 10)

Further calculation shows that the test statistic is $W = \frac{(15-3)(40.133)}{(3-1)(836.8)} = 0.288$. If the significance level is α=0.05, then the criticial value is $F_{0.95,2,12}$= 3.9. Since the test statistic W = 0.288 is far below the critical value, we cannot reject the null hypothesis of equal variances

4 Non-normal data

The *t*-test and its associated confidence interval as well as ANOVA are designed to be optimal when the data follow a normal distribution (also known as a Gaussian distribution). While the *t*-test is known to be robust to *Type I* errors when the data deviate from normality, it always pays to graph the data to check its shape, trends, and variability.

4.1 Diagnosis and repair

4.1.1 Graphical tools

It is recommended that experimenters always plot their data. Even the most rudimentary plots that could be done with paper and pencils in the past can provide insights into the data that can go beyond summary statistics.

A histogram can be useful for checking if the data approximately follows the bell-shaped curve of the normal distribution. If available, normal probability plots, also known as qq-norm plots (or normal Q-Q plots), should also be checked. The normal probability plot is a graph that plots the ordered measurements on the vertical axis versus the median or average of the corresponding order statistics of a standard normal sample of the same size. Samples that are normally distributed should produce a normal probability plot where the points appear to be scattered close to a straight line. A normal probability plots that is far from linearity is evidence of non-normality. Unfortunately, it is unfeasible to do a probability plot by hand; however, it is available in many statistical and worksheet programs. A histogram and a normal probability plot of a normally distributed sample is shown in Figure 4.

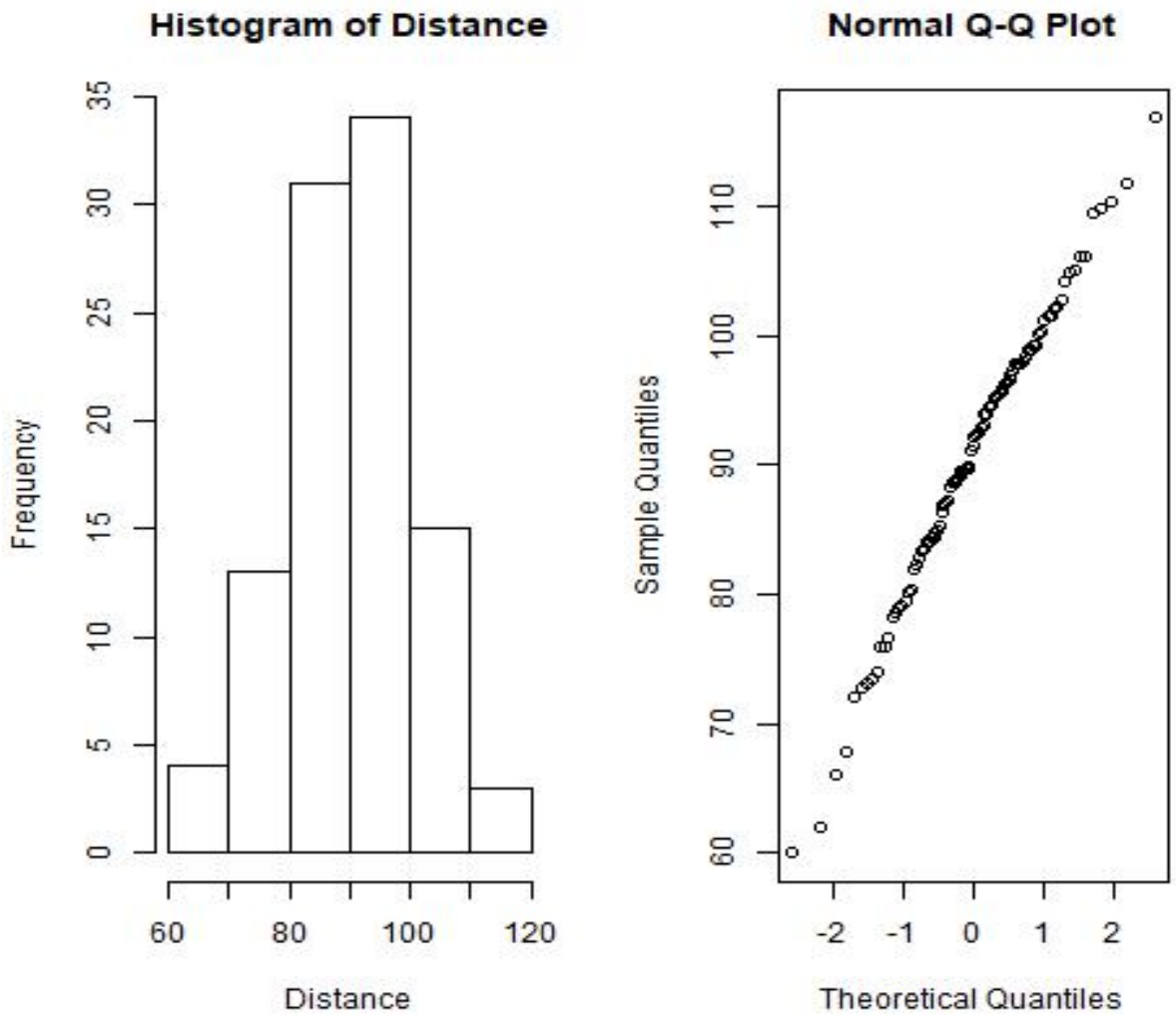

Figure 4: Histogram and Normal probability plot of a normally distributed sample of distance measurements

4.1.2 Transformations

It is possible that the data will be approximately normally distributed if transformed. Ideally, there will be a physical basis such a transformation. Otherwise, a particularly useful family of transformations is the Box-Cox transformation:

$$T(X) = \frac{X^{\lambda} - 1}{\lambda},$$

Here λ is a parameter chosen to make the data Gaussian. For $\lambda = 0$, $T(X) = \log(X)$.

The logarithmic transformation is the most commonly used transformation for positive measurement data that are skewed to the right.

Example

Testing of one instrument yielded 20 measurements of distance till detection (in cm):

68 45 477 60 48 143 39 109 238 217 563 291 110 137 278 353 19 75 262 764

The histogram and normal probability plots (Figure 5) of this data show that it is not normally distributed.

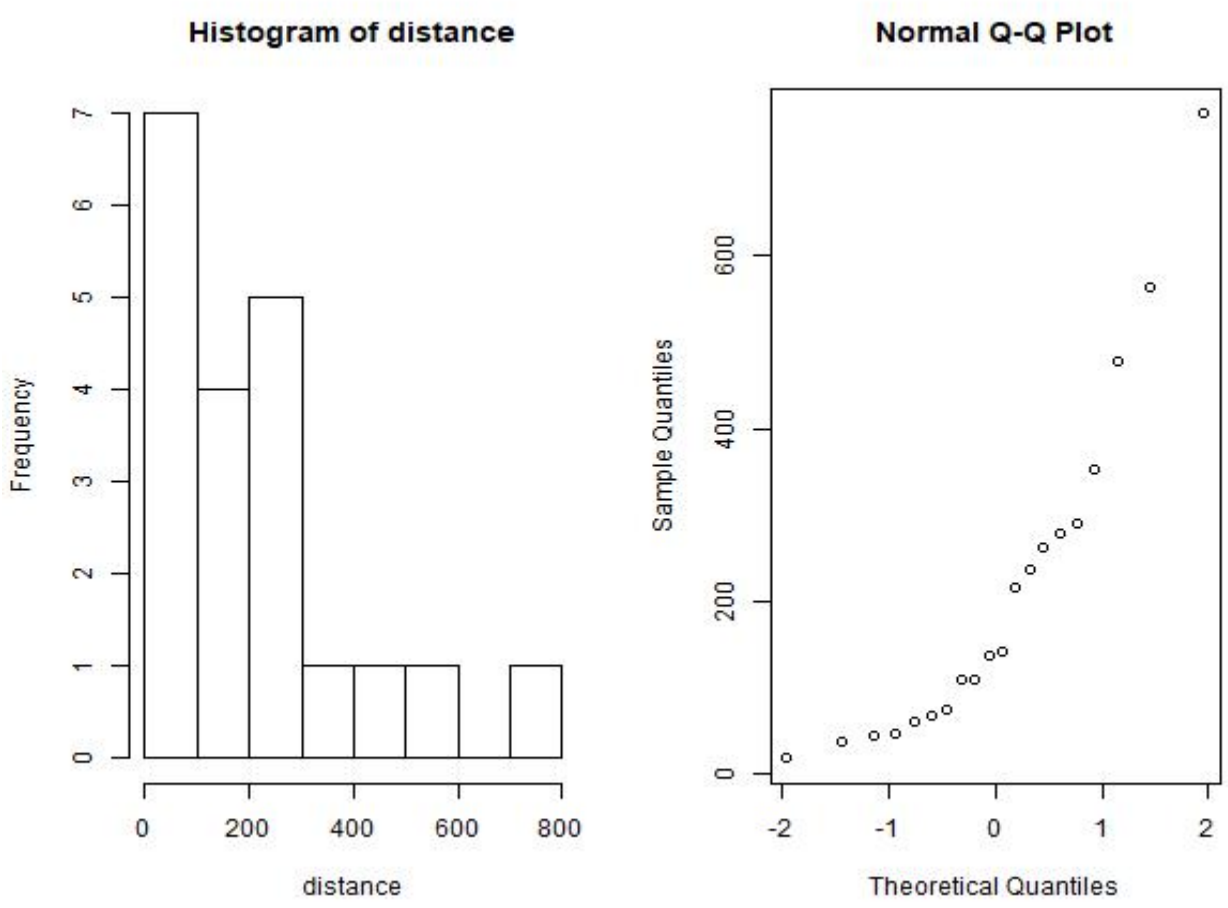

Figure 5: The histogram and normal probability plot show that this sample of distance measurements is not normally distributed.

Using a logarithmic transformation on the distance data yields:

4.22 3.81 6.17 4.09 3.87 4.96 3.66 4.69 5.47 5.38 6.33 5.67 4.70 4.92 5.63 5.87 2.94 4.32 5.57 6.64

The histogram and normal probability plot of the transformed data inFigure 6 show that the transformed data approaches approximate normality.

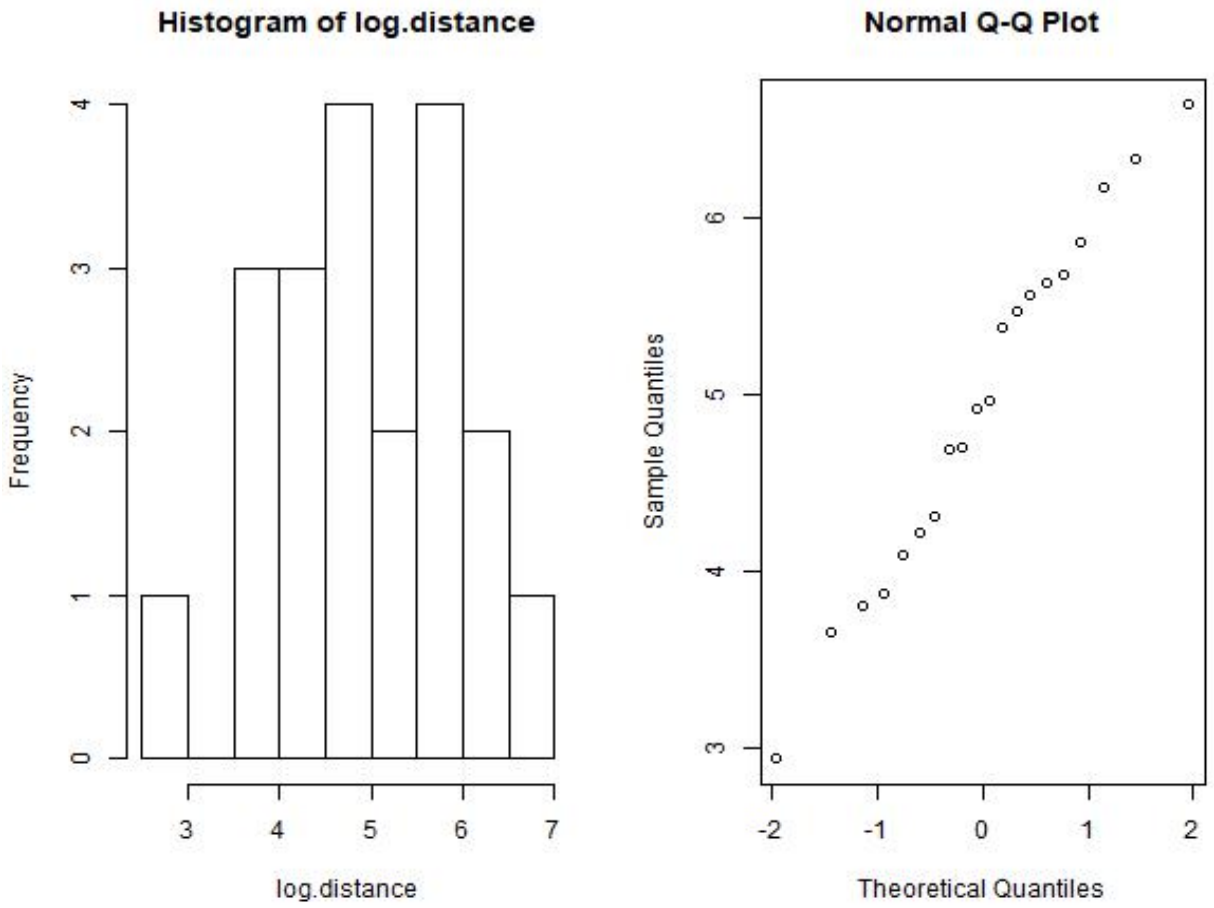

Figure 6 : The histogram and normal probability plot of the transformed data show that the transformed data approaches approximate normality.

One problem with transforming the data is the need to refer to the transformed data rather than the original data, as well as the possibility of no appropriate transformation existing for that particular data.

4.2 Nonparametric methods

It may be important to use methods not beholden to any, or at least many fewer, distributional assumptions. Nonparametric methods attempt to achieve some of the same purposes as more classical procedures such as Student's *t* test and the ANOVA *F*-test, but without the same required distributional assumptions. There is always a trade-off as classic methods are optimal for normally distributed data, but nonparametric methods can be much more powerful for non-normal data. Also, nonparametric statistics tend to be more cumbersome to compute than classical statistics, but the increasing prevalence of computational and statistical software has lessened that burden considerably.

4.2.1 Wilcoxon test for comparing two distributions

A robust efficient alternative to the *t*-test that is not based on normal data is the Wilcoxon Rank Sum test. Note that another popular nonparametric test known as the Mann-Whitney U-test has been shown to be equivalent to the Wilcoxon test, so it will not be covered here. For the Wilcoxon test, the null hypothesis is that the two distributions are equivalent, while the alternative hypothesis is that one of the distributions is shifted to the right or the left of the other distribution [3]. The test is designed for continuous numerical data, as the presence of many ties in the combined data set (as might happen in a situation where the same number occurs many times, as in a Likert scale sample) can be problematic.

Suppose that measurements for Instrument 1 are $x_1, \ldots, x_m$, and for Instrument 2 are $y_1, \ldots, y_n$. Then rank all $m + n$ measurements in order of magnitude, with the smallest one receiving rank one and the largest one receiving rank $m + n$. Tied observations are given ranks equal to the mean of their ranks. For example, if the seventh and eighth smallest measurements are the same, both measurements are assigned the rank of 7.5. Let T1 be the sum of the ranks from Instrument 1, and T2 be the sum of the ranks from Instrument 2. If the null hypothesis that the two distributions are the same is true, then one would expect that the average rank from each distribution would be roughly the same; for the equal sample size case of $m = n$, we would expect T1 and T2 to be similar; in that case, if either T1 or T2 were much larger than the other rank sum, that would be evidence against the null hypothesis. In practice take the rank sum of the sample with fewer observations (if $m = n$, then take either rank sum), and compare to the appropriate critical points given m and n in a table of Wilcoxon Rank Sum Test critical values, which should be in many statistics textbooks and online resources. If both sample sizes are large, say at least 10, then a normal approximation test can be used. Let T1 be the rank sum of the sample with fewer measurements (making $m \leq n$ in the following formula). Let

$$Z = (\mathrm{T1} - \left[\frac{mn + m(m+1)}{2}\right]) / \sqrt{\frac{mn(m+n+1)}{12}}.$$

Then it is known that under the null hypothesis of identical populations, the statistic Z can be approximated by a normal distribution with mean 0 and variance 1. Therefore, the null

hypothesis of same distributions is rejected if Z is beyond the critical point in a standard normal table, i.e. for significance level α, reject the equality of the instruments if $|Z| > z_{1-\frac{\alpha}{2}}$.

Example

Suppose for each of two instruments, a sample of 10 measurements of the distance till detection was measured in cm.

The measurements for Instrument 1 $(x_1, \dots, x_{10})$:

34 53 38 62 21 36 46 28 73 107

For Instrument 2, the measurements are $(y_1, \dots, y_{10})$:

72 57 70 54 127 128 61 56 99 55

It is needed to test whether the samples come from the same distribution. Neither sample appears normally distributed, so it is decided to use a Wilcoxon test. First, the two samples are combined, and each observation is assigned a rank. These ranks are listed in the following table.

x_i, y_i (cm)	Rank	Instrument
21	1	1
28	2	1
34	3	1
36	4	1
38	5	1
46	6	1
53	7	1
54	8	2
55	9	2
56	10	2
57	11	2
61	12	2
62	13	1
70	14	2
72	15	2
73	16	1
99	17	2
107	18	1
127	19	2
128	20	2

Table 3: List of instrument measurements with ranks.

The sum of ranks of the Instrument 1 measurements is T1 = 75, while the sum of ranks of the Instrument 2 measurements is T2 = 135. Since both sample sizes are at least 10, the normal approximation can be used, leading to the statistic:

$$Z = (75 - \left[\frac{10(10) + 10(10+1)}{2}\right] / \sqrt{\frac{10(10)(10+10+1)}{12}}$$

$$= -\frac{30}{13.23} = -2.27.$$

If a significance level of α = 0.05 is used, the critical point is $z_{.975} = 1.96 < |-2.27|$, so the null hypothesis of no instrument difference is rejected. Using statistical software on this example will yield a *p*-value of 0.02.

4.2.2 Sign test for paired comparisons

Let us go back to the situation of paired samples in Section 3.1.1.3, for which we previously used the paired *t*-test. Suppose that the assumptions of normally distributed differences do not hold. A simple nonparametric alternative to the paired *t*-test that is easy to calculate is the sign test, which tests if the median difference is significantly different from zero. For this paired situation, the sign test reduces each paired difference as either positive (+) or negative (-). Suppose we presume the two distributions were equal, then a (+) would be equally likely as a (-). Then under this null hypothesis, if S+ is the number out of *k* paired differences that are positive, and S- is the number out of *k* paired differences that are negative, then both S+ and S- would follow a Binomial distribution with *k* trials and probability ½, with average value being *k*/2. Let S=max(S+, S-). Then, the *p*-value associated with S is 2 P($X \geq$ S) if *X* is a Binomial random variable with *k* trials and probability ½. For a hypothesis test with significance level α, the null hypothesis is rejected if this *p*-value is less than α.

The Binomial distribution with *k* trials and p=1/2 can be approximated by a Normal distribution with mean *k*/2 and variance *k*/4 for large *k*, say $k \geq 10$ [3]. Thus, we can again transform the test into a *z* statistic:

$$z = \frac{S - \frac{k}{2}}{\sqrt{k/4}} .$$

The null hypothesis is rejected if *z* is greater than the 100 (1- α/2) th percentile point of the standard normal distribution.

Example

Refer to the dataset highlighted in the example in Section 3.1.3 on the paired *t*-test. The set of 20 paired differences between the two instrument measurements is:

13 7 8 1 -1 7 13 -6 23 1 6 2 12 12 -3 7 14 3 12 1.

Note that 17 differences are positive, and 3 are negative. Suppose the significance level is $\alpha = 0.05$. If X is a binomial random variate with 20 trials and probability 0.5, then the associated p-value is $2\ P(X \geq 17) = 0.0026$. Thus, the null hypothesis of equal distributions (or at least equal medians) is rejected.

It may be easier for some to use the normal approximation, which produces a z statistic of

$$Z = \frac{17 - \frac{20}{2}}{\sqrt{\frac{20}{4}}} = 3.13,$$

which also signifies rejection of the null hypothesis.

Since much of the information in the data is not taken into account, the sign test is much less powerful and efficient than the paired t-test if the data are normal or close to it. Conversely, the sign test can be more powerful than the t-test in non-normal cases.

4.2.3 Kruskal-Wallis procedure for testing multiple distributions

Suppose there are k instruments each yielding approximately continuous numerical measurements, with the i^{th} instrument having n_i measurements. The Kruskal-Wallis (KW) procedure tests the null hypothesis that all k samples come from the same distribution without making the distributional assumptions of ANOVA. The alternative hypothesis is that at least two of the distributions differ in location from each other. The samples do need to be random and independent. As with the Wilcoxon test, the measurements should be continuous numerical data to avoid repeated data values.

The KW procedure is similar to the Wilcoxon Test described in one of the previous sections in that it is based on ranks. Put all $N = \sum_{i=1}^{k} n_i$ measurements into one sample and rank them in order of magnitude, with the smallest one receiving rank one and the largest one receiving rank N. Tied observations are given ranks equal to the mean of their ranks. Let the sum of ranks for the i^{th} instrument be R_i. If the null hypothesis is true, and if each sample size is at least 5, then the test statistic

$$H = [\frac{12}{N(N+1)} \sum_{i=1}^{k} (R_i^2 / n_i)\ ] \ - \ 3(N+1)$$

should be approximated by a *chi-square* distribution with *k-1* degrees of freedom. For significance level α, the null hypothesis is rejected if H is greater than the 100(1-α) percentile point of the χ^2_{k-1} distribution.

Example

Refer back to the data set of the example in Section 3.2.1 on ANOVA. The three samples are pooled and then ranked according to magnitude as:

Distance (cm)	Rank	Instrument
67	1	1
68	2	1
69	3	1
72	4	1
76	5	2
83	6	2
88	7	1
99	8	2
101	9	2
102	10	2
108	11.5	3
108	11.5	3
114	13	3
124	14	3
132	15	3

The rank sums are: $R_1 = 17,\ R_2 = 38,\ R_3 = 65$. So the test statistic is

$$H = \frac{12}{15(15+1)}(1191.6) - 3(15+1) = 11.6.$$

For significance level α = 0.5, the null hypothesis is rejected because H=11.6 >$\chi^2_{0.95:\,2}$=5.99.

Statistical software shows the *p*-value is 0.003.

5 Proving Equivalence

The methods provided thus far have sought to show that the true mean performances of two (or more) instruments are different. In these traditional hypothesis test formulations, we began with the assumption that the instruments' true performances are the same and sought data-based evidence to prove that they are different. In this section we present the *equivalence test* (Wellek 2010), (Richter and Richter 2002) that allows for the experimenter to prove that the true mean performances of two (or more) instruments are the same.

The traditional hypothesis test formulation allows an experimenter to prove that the true mean performances of two (or more) instruments are different. The equivalence test allows an experimenter to prove that the true mean performances of two (or more) instruments are the same.

For a practical example where an experimenter may like to prove that the true mean performances of two instruments are equivalent, we look to the American National Standards Institute's, American National Standard Performance Criteria for Alarming Personal Radiation Detectors for Homeland Security, ANSI N42.32(2016). The functionality tests of ANSI N42.32 seek to demonstrate that an instrument is unaffected by an imposed environmental, electromagnetic, or mechanical shock. That is, the experimenter would like to prove that the performance of the instrument after the imposed shock is equivalent to the performance of the instrument prior to being exposed to the shock. Though we are not strictly comparing two instruments, we can view the pre-shock instrument and the post-shock instrument as two instruments.

5.1 Null Hypothesis and Indifference Zone

The equivalence test differs from the traditional hypothesis test formulation on two main accounts:

1. In equivalence testing we begin with the assumption (null hypothesis) that the instruments' performances are different. We then seek data-based evidence to prove that they are the same.
2. The equivalence test only considers differences in instruments' performances that are of practical significance to the experimenter.

Recall that the hypothesis test relies on the idea of proof by contradiction. That is, we state as the null hypothesis the conjecture that is opposite of what we would like to prove. In the traditional hypothesis test formulation where we seek to identify a meaningful difference, we define the null hypothesis as the instruments' true performances being the same. If we find sufficient evidence to reject the null hypothesis, then we can confidently state that the instruments' true performances are different. Failure to reject the null hypothesis does not prove that the instruments' true performances are the same, but rather that insufficient evidence was found to support the conclusion that the instruments' true performances are different.

In equivalence testing it is our goal to prove that the true mean performances of two (or more) instruments are the same. Thus, we begin with the assumption (null hypothesis) that the

instruments' performances are different and seek evidence to show that they are the same. When we find such evidence to support the rejection of the null hypothesis, we can then confidently state that the instruments' true performances are equivalent. Here again, failure to reject the null hypothesis is not proof the that null hypothesis is true, but simply indicates that insufficient evidence was found to support the conclusion that the instruments' true performances are equivalent.

Experimenters may find comfort in the fact that the equivalence test allows for deviations that are viewed as practically irrelevant in its definition of "the same". In developing the equivalence test, the experimenter must define the "indifference zone" where differences in the instruments' true mean performances are small enough to not be considered of practical importance in the context of the decision being made. We adopt the notation $[\delta_L, \delta_U]$ to represent the indifference zone where δ_L provides the lower bound of the indifference zone and δ_L provides the upper bound. Guidance for choosing the endpoints of the indifference zone can be found in Wellek (2010) and Anderson-Cook and Borror (2016).

For example, though not formulated strictly as an equivalence test, the functionality tests of ANSI N42.32 are written in such a way as to not penalize an instrument if its post-shock performance is within 15 % of its pre-shock performance. If we let μ_{pre} represent the instrument's true pre-shock performance, then the indifference zone is $[-0.15\mu_{pre}, 0.15\mu_{pre}]$. That is, a difference between the pre-shock and post-shock instrument performance that is within ±15 % of the pre-shock performance is considered practically irrelevant.

5.2 Hypothesis Test

In equivalence testing we begin with the null hypothesis that the instruments' performances are different and seek evidence that allows us to reject the null hypothesis and prove that they are the same. Let μ_A represent the true performance of instrument A and μ_B represent the true performance of instrument B. We define "different" such that the difference in the instruments' true performances, $\mu_A - \mu_B$ falls outside of the indifference zone $[\delta_L, \delta_U]$. That is the instruments are considered different if $\mu_A - \mu_B < \delta_L$ or $\mu_A - \mu_B > \delta_U$. We see that this equates to not a single hypothesis test, but two one-sided tests (TOST) as provided in Equations (5.1) and (5.2), where H_0 and H_1 represent the null and alternative hypothesis.

$$H_0\colon\ \mu_A - \mu_B < \delta_L \qquad versus \qquad H_1\colon\ \mu_A - \mu_B \geq \delta_L \tag{5.1}$$

and,

$$H_0\colon\ \mu_A - \mu_B > \delta_U \qquad versus \qquad H_1\colon\ \mu_A - \mu_B \leq \delta_U \tag{5.2}$$

Assuming that the instruments' performances are measured using a continuous measurement, the test statistics used to evaluate these two one-sided tests are based on Student's *t*-test (see Section 3) and are provided in Equations (5.3) and (5.4).

$$T_L = \frac{\left(\bar{X}_A - \bar{X}_B\right) - \delta_L}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \tag{5.3}$$

$$T_U = \frac{\left(\bar{X}_A - \bar{X}_B\right) - \delta_U}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \tag{5.4}$$

Where $\bar{X}_A$ and $\bar{X}_B$ are the means of the n_A and n_B performance measurements of instruments A and B. The pooled standard deviation, s_p is calculated according to Equation (5.5).

$$s_p = \sqrt{\frac{\left(n_A - 1\right)s_A^2 + \left(n_B - 1\right)s_B^2}{n_A + n_B - 2}} \tag{5.5}$$

Where s_A and s_B are the standard deviations of the of the n_A and n_B performance measurements of instruments A and B.

To ascertain whether we are to reject each of the null hypothesis in Equations (5.1) and (5.2), we compare the test statistics of Equations (5.3) and (5.4) to an associated rejection criteria. To define this rejection criteria, we must determine the largest probability of committing a type I error that we are willing to accept. Recall that a type I error occurs when we erroneously reject the null hypothesis. In the equivalence test setting, a type I error means that we conclude that instruments' performances are equivalent when in fact they are truly different. See Leber, Pibida, and Enders (2019) for a discussion on selecting an acceptable type I error probability in the context of radiation and nuclear detection systems. We denote the probability of committing a type I error with the Greek letter α.

With α identified, and T_L and T_U calculated according to Equations (5.3) and (5.4), we reject the null hypothesis of Equation (5.1) if:

$$T_L > t_{1-\alpha, n_A + n_B - 2} \tag{5.6}$$

and we reject the null hypothesis of Equation (5.2) if:

$$T_U < -t_{1-\alpha, n_A + n_B - 2} \tag{5.7}$$

Here, $t_{1-\alpha, n_A + n_B - 2}$ is the $\left(1-\alpha\right)\times 100th$ Student's t quantile with $n_A + n_B - 2$ degrees of freedom. Student's t quantile values are easily assessable through tables found in statistics textbooks, e.g., Montgomery and Runger (2018), or lookup functions in statistical software, spreadsheet tools, and online calculators.

If we are able to reject the null hypothesis in both Equations (5.1) and (5.2), then we reject the overall null hypothesis of the equivalence test, and conclude that the instruments' performances are the same. If we fail to reject either of the null hypotheses of Equations (5.1) and (5.2), then we cannot reject the overall null hypothesis of the equivalence test, and thus we are unable to deem that the instruments' performances are the same.

5.3 Example

In the spirit of using the ANSI N42.32 standard as an example, a personal radiation detection (PRD) system is subjected to an over-range shock for photons by subjecting the instrument to a gamma-ray field that exceeds its maximum measurable exposure rate by a factor of two for two minutes. The instrument's pre- and post-shock performance is captured by its ability to measure the radiation exposure rate of a small radioactive source that produces a stable radiation field. The goal is to prove that the instrument was not adversely affected by the over-range shock for photons. That is, we want to show that the instrument's pre-shock performance is equivalent to its post-shock performance. We define equivalence as a difference between the pre-shock and post-shock performance that does not exceed 15 % of the pre-shock performance. We will perform this test at the $\alpha = 0.05$ level.

In the test area a radiation field of 30 μR/h (7.74×10^{-9}(C/kg)/h) is produced by a small radioactive source at the reference point of the instrument, the instrument provided the following pre- and post-shock readings:

pre-shock (μR/h): 30, 30, 30, 30, 30, 31, 31, 31, 31, 31

post-shock (μR/h): 28, 30, 30, 30, 29, 29, 29, 30, 30, 30

The summary statistics for these observations are:

$$\bar{x}_{\text{pre}} = 30.5 \qquad s_{\text{pre}} = 0.527 \qquad n_{pre} = 10 \qquad \bar{x}_{\text{post}} = 29.5 \qquad s_{\text{post}} = 0.707 \qquad n_{post} = 10$$

We first define the lower and upper bounds of the indifference zone, $[\delta_L, \delta_U]$, as ±15 % of the pre-shock performance:

$$\delta_L = -0.15\bar{x}_{pre} = -4.575\,\mu R/h \quad \text{and} \quad \delta_U = 0.15\bar{x}_{pre} = 4.575\,\mu R/h\,.$$

Next, we calculated the pooled standard deviation, s_p, according to Equation (5.5) and the test statistics T_L and T_U according to Equations (5.3) and (5.4).

$$s_p = \sqrt{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2}} = \sqrt{\frac{(10-1)0.527^2 + (10-1)0.707^2}{10+10-2}} = 0.6236$$

$$T_L = \frac{(\bar{X}_A - \bar{X}_B) - \delta_L}{s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} = \frac{(30.5-29.5)-(-4.575)}{0.6236\sqrt{\frac{1}{10}+\frac{1}{10}}} = 19.990$$

$$T_U = \frac{(\bar{X}_A - \bar{X}_B) - \delta_U}{s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} = \frac{(30.5-29.5)-(4.575)}{0.6236\sqrt{\frac{1}{10}+\frac{1}{10}}} = -12.819$$

And finally, we formulate our conclusion by comparing the test statistics $T_L = 19.990$ and $T_U = -12.819$ with their appropriate rejection regions as defined by Equations (5.6) and (5.7). With $t_{1-\alpha,n_A+n_B-2} = t_{0.95,18} = 1.734$, we reject the null hypothesis of Equation (5.1) since $T_L > 1.734$ and we

reject the null hypothesis of Equation (5.2) since $T_U < -1.734$. Because we reject the null hypothesis in both Equations (5.1) and (5.2), then we reject the overall null hypothesis of the equivalence test, and conclude that the instrument's post-shock performance is equivalent to its pre-shock performance.

5.4 Confidence Interval Representation

In general, a hypothesis test can be formulated and presented in terms of a confidence interval that is compared to some threshold. For the equivalence test presented in this section, we reject the overall null hypothesis of the equivalence test and conclude that the instruments' performances are the same at the α-level of significance if the confidence interval provided in Equation (5.8) is entirely contained within the indifference zone $[\delta_L, \delta_U]$.

$$\left(\bar{X}_A - \bar{X}_B\right) \pm t_{1-\alpha, n_A+n_B-2} s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \tag{5.8}$$

The confidence interval of Equation (5.8) is formulated by combining the two one-sided confidence intervals that correspond to the tests of Equations (5.1) and (5.2). The result is a $100(1-2\alpha)\%$ two-sided confidence interval with an expansion factor based on the $100(1-\alpha)th$ Student's t quantile. Note that this is unlike the $100(1-\alpha)\%$ confidence intervals presented in Section 3 that are used to detect a difference between two instruments and rely on an expansion factor based on the $100\left(1-\frac{\alpha}{2}\right)th$ Student's t quantile.

In continuing with the above example, the confidence interval for equivalence is:

$$\left(\bar{X}_A - \bar{X}_B\right) \pm t_{1-\alpha, n_A+n_B-2} s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = (30.5 - 29.5) - 1.734 \cdot 0.6236 \sqrt{\frac{1}{10} + \frac{1}{10}} = 1 \pm 0.484 = [0.516, 1.484]$$

Since this interval is entirely contained within the indifference zone $[\delta_L, \delta_U] = [-4.575, 4.575]$ we reject the overall null hypothesis of the equivalence test, and conclude that the instrument's post-shock performance is equivalent to its pre-shock performance. An illustration of this confidence interval approach is provided in Figure 7.

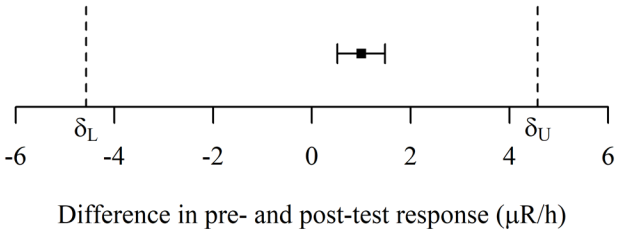

Figure 7: Combined one-sided 95 % confidence intervals for the difference in the pre- and post-test response. The indifference zone [δL,δU] is displayed by the dashed lines.

5.5 Power and Sample Size

As with the traditional formulation of a hypothesis test, two errors are possible in an equivalence test. A *type I error* occurs when the equivalence test leads us to conclude that the performances of

the instruments are the same when in fact, they are truly different. While a *type II error* is when we fail to conclude that the performances of the instruments are equivalent when they are truly the same. We can control the probabilities of these errors through our choice of sample size and can use a *power curve* to evaluate the expected impact of sample size on the equivalence test's type I and type II errors.

In the case of the equivalence test, the power curve provides the probability of concluding that the instruments are equivalent (i.e., rejecting the null hypothesis) as a function of the true difference in the instruments' performances. Ideally, when the true difference in the instruments' performances is within the indifference zone we would always conclude (probability of one) that the instruments are equivalent and otherwise always fail to conclude (probability of zero) that the instruments are equivalent. This ideal power curve, for the above example with the indifference zone $[\delta_L, \delta_U] = [-4.575, 4.575]$, is displayed in Figure.

Unfortunately, a test with no risk (no type I and type II errors), such as the ideal test illustrated with the power profile displayed in Figure 8, requires an infinite number of samples. Therefore, common practice is to state a maximum acceptable type I error probability and construct a suitable acceptance criterion and sample size. The resulting power curve is examined, and the sample size adjusted to satisfy the desired type II error probability. See Leber, Pibida, & Enders(2002) for further discussion on setting maximum acceptable errors in the context of homeland security applications.

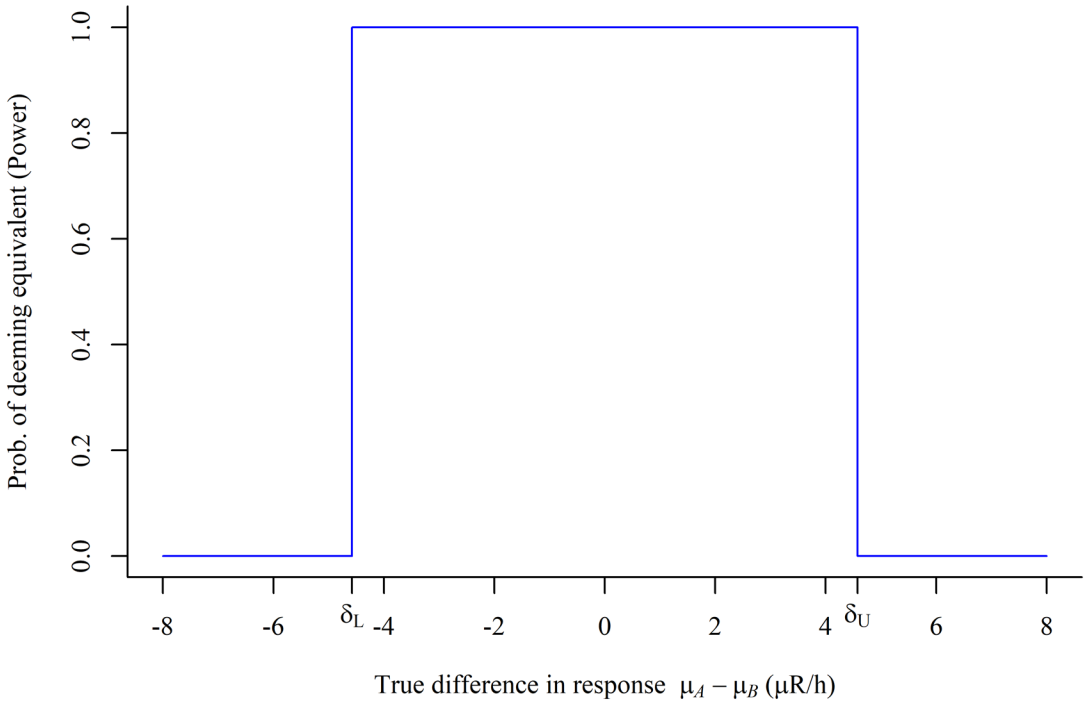

Figure 8: Power curve for an ideal equivalence test with indifference zone $[\delta_L, \delta_U] = [-4.575, 4.575]$.

To calculate the power of the equivalence test for a given finite sample size, $\{n_A, n_B\}$, we must calculate the probability of rejecting the null hypotheses of both Equations (5.1) and (5.2). Since we reject the null hypothesis of Equation (5.1) when $T_L > t_{1-\alpha, n_A+n_B-2}$ and we reject the null hypothesis of Equation (5.2) when $T_U < -t_{1-\alpha, n_A+n_B-2}$, then it follows that the power of the equivalence test is provided by Equation (5.9).

$$Power = P\left(T_L > t_{1-\alpha, n_A+n_B-2} \text{ and } T_U < -t_{1-\alpha, n_A+n_B-2}\right) \tag{5.9}$$

where T_L and T_U are the test statistics provided by Equations (5.3) and (5.4) that are distributed according to a non-central t-distribution with non-centrality parameters $\left(\mu_A - \mu_B - \delta_L\right) \Big/ \left(\sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right)$ and $\left(\mu_A - \mu_B - \delta_U\right) \Big/ \left(\sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right)$, respectively (Leber, Pibida et al. 2019). Using Equation (5.9) and s_p as an estimate for σ, we calculate the power for the test from the above example with $n_A = n_B = 10$ and several alternative sample sizes. These power curves are presented, along with the ideal power curve $\left(n_A = n_B = \infty\right)$ in Figure 9.

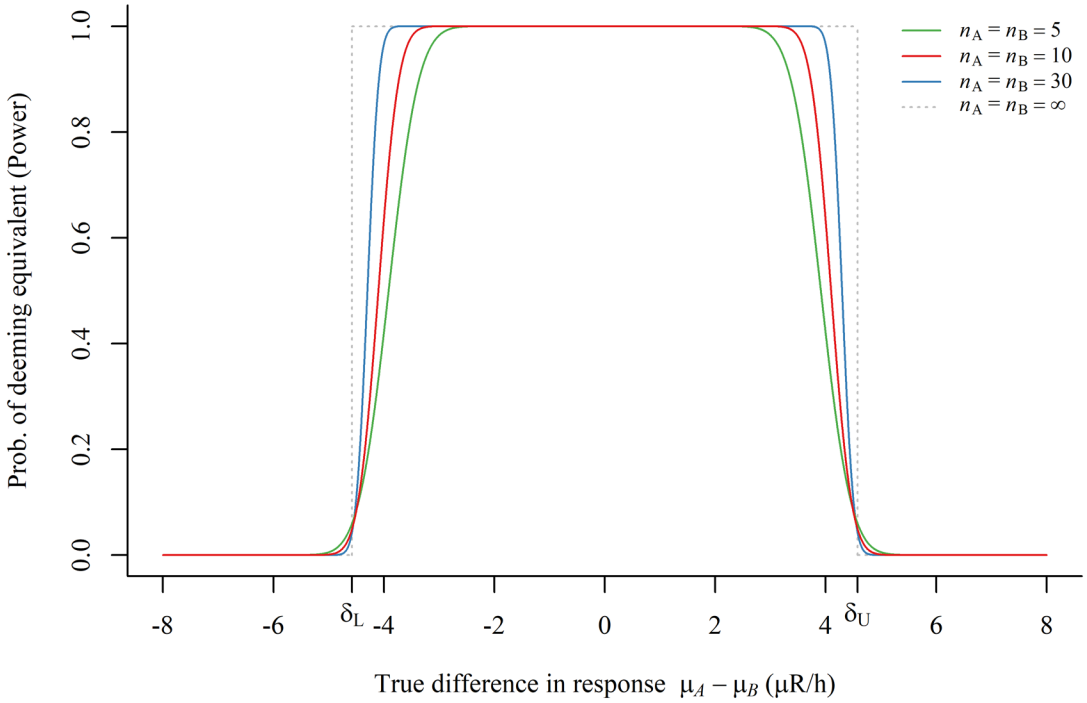

Figure 9: Power curves for several equivalence tests of varying sample sizes, $n_A = n_B$, each with a maximum type I error probability $\alpha = 0.05$, indifference zone $[\delta_L, \delta_U] = [-4.575, 4.575]$, and s estimated as 0.624.

Among the finite tests displayed in Figure 9, we observe that the power curve of the $n_A = n_B = 30$ test most closest represents that of the ideal $n_A = n_B = \infty$ test. Further, when the true difference in

the instruments' responses is of practical significance to the experimenter, i.e., falls outside of the indifference zone $[\delta_L, \delta_U] = [-4.575, 4.575]$, the probability of erroneously deeming the instruments equivalent does not exceed 0.05 for all of the finite tests. We also note that for all tests displayed in Figure 9 the power curves reach their maximum value of one within the indifference zone. The difference between the tests is the rate at which they reach the maximum power. Consider, for example, two "equivalent" instruments whose true difference in response is 4 μR/h (1.03×10^{-9} (C/kg)/h), then, from the curves provided in Figure 9, we observe:

1. When $n_A = n_B = 5$, the probability of deeming these truly equivalent instruments as such is 0.42 (probability of a type II error = 0.58);
2. When $n_A = n_B = 10$, the probability of deeming these truly equivalent instruments as such is 0.63 (probability of a type II error = 0.37); and,
3. When $n_A = n_B = 30$, the probability of deeming these truly equivalent instruments as such is 0.97 (probability of a type II error = 0.03).

Thus, amongst the three finite tests illustrated in Figure 9, we are most likely to draw the correct conclusion in the $n_A = n_B = 30$ test. The experimenter must consider whether this increase in test performance (power) is worth the increased cost associated with the additional required samples.

In Figure 10 we illustrate how variability in the test measurements, in addition to the sample size, effects the power of the equivalence test. In Figure 10 we again calculate the power for the test from the above example, however, here we have increased s_p, the estimate for σ, by a factor of four.

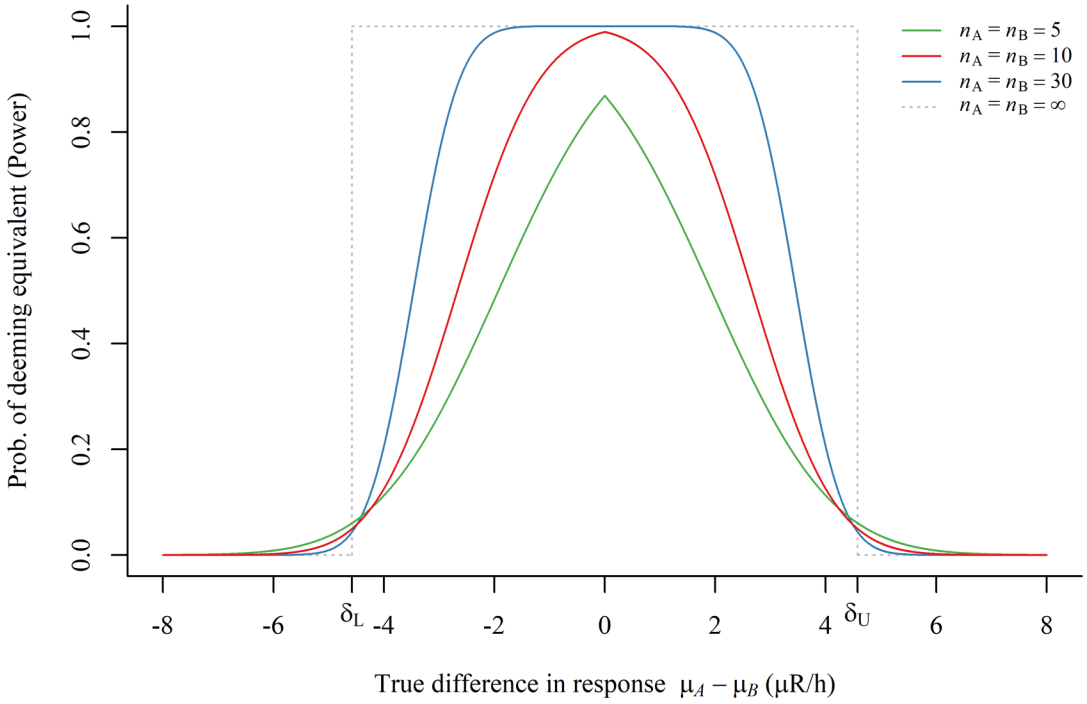

Figure 10: Power curves for several equivalence tests of varying sample sizes, $n_A = n_B$, each with a maximum type I error probability a = 0.05, indifference zone $[\delta_L,\delta_U] = [-4.575,4.575]$, and s estimated as 2.494.

By comparing the power curves in Figure 10 to those in Figure 9, we see that the increased variability in the measurements has reduced our ability to deem truly equivalent systems as such. As an extreme case, consider two instruments whose responses are exactly the same, i.e., the true difference in response is 0 μR/h (0 (C/kg)/h). With the lower measurement variability of Figure 9, the probability of deeming these two systems equivalent is one, even with the small sample size of $n_A = n_B = 5$. However, when the measurement variability is larger (Figure 10), we see that for the small sample size test of $n_A = n_B = 5$, the probability of deeming these two systems equivalent is 0.87. That is, there is a 13 % chance that the test will fail to deem these two systems equivalent.

5.6 Other Applications

This section provided an overview of an equivalence test that allows for experimenter to prove that the true mean performances of two (or more) instruments are the same. Pardo (2002) provides practical applications of equivalence tests that expand beyond this application to include the equivalence of proportions and variances.

6 References

[1] Leber, D. D., Pibida, L., & Enders, A. L. (2019, June). Confirming a Performance Threshold with a Binary Experimental Response. *NIST Technical Note*. Retrieved from https://doi.org/10.6028/NIST.TN.2045

[2] NIST/SEMATECH e-Handbook of Statistical Methods, available at: http://www.itl.nist.gov/div898/handbook/

[3] Mendenhall, W, & Sincich, T. (1992). *Statistics for Engineering and the Sciences*, Third Edition, Dellen - MacMillan, New York.

[4] Hahn, G. & Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*, John Wiley & Sons, Inc., New York.

[5] Ross, S. (1976). *A First Course in Probability*, Macmillan, New York.

[6] Agresti, A., & Caffo, B., (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures, *The American Statistician*, *54*(4), 280-288.

[7] Fleiss, J. L., (1981). *Statistical Methods for Rates and Proportions*, Second Edition, John Wiley & Sons, New York.

[8] Snedecor, G.W. & Cochran, W.G. (1989). *Statistical Methods,* Eighth Edition, Iowa State University Press, Ames, IA.

[9] Brown, M.B. & Forysthe, A.B. (1974), *Journal of the American Statistical Association, 69*, 364-367.

[10] American National Standards Institute. (2016). *American National Standard Performance Criteria for Alarming Personal Radiation Detectors for Homeland Security (N42.32),* Institute of Electrical and Electronics Engineers, Inc., New York.

[11] Anderson-Cook, C. M., & Borror, C. M. (2016). The difference between "equivalent" and "not different". *Quality Engineering, 28*(3), 249-262.

[12] Casella, G., & Berger, R. L. (2002). *Statistical Inference,* Second Edition, Duxbury, Pacific Grove, CA.

[13] Pardo, S. (2014). *Equivalence and Noninferiority Tests for Quality, Manufacturing and Test Engineers,* CRC Press, Boca Raton, FL.

[14] Richter, S. J., & Richter, C. (2002). A Method for Determining Equivalence in Industrial Applications. *Quality Engineering, 14*(3), 375-380.

[15] Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, Second Edition, Chapman & Hall/CRC, Boca Raton, FL.

[16] Montgomery, D.C.. & Runger, G.C. (2018). *Applied Statistics and Probability for Engineers*, Seventh Edition, John Wiley & Sons, Inc., New York.