Keysight Technologies
How Much Indentation Testing is Enough Testing?

Application Note

Abstract

The Student's t-test is rearranged in order to predict the sampling (N) required to conclude significant difference between two observation sets at a particular confidence level. The expression is appropriate for small or large observation sets. When the difference in means is small relative to the standard deviation, more tests are required in order to reliably detect that difference.

Introduction

In 2012, Keysight Technologies introduced the Express Test option for the G200 NanoIndenter platform. This option implements traditional indentation testing in a revolutionary way in order to achieve unprecedented testing speeds [1,2]. Express Test performs one complete indentation cycle per second, including approach, contact detection, load, unload, and movement to the next indentation site. One hundred indentations can be performed at one hundred different sites in less than 100 seconds. Given that indentations can be performed so quickly, a new question arises: How much testing is enough testing? The purpose of this note is to answer this question by applying the Student's t-test in an uncommon way.

The Student's t-test is a statistical test used to determine, to a reasonable degree of confidence, whether two observation sets obtain from different populations. Implementation of the Student's t-test always begins with the assumption that the two observation sets come from the same population. This is called the "null-hypothesis". If the difference between the two averages is sufficiently large relative to measurement scatter, then we reject the null hypothesis and conclude that the two observation sets do in fact come from different populations. If we assume that each set contains the same number of observations (N), then the Student's t-criteria for concluding significant difference is expressed as

$$\frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{N}}} > z_{critical}, \qquad (1)$$

where $(\overline{x_i})$ and σ_i^2 represent the average and standard deviation of each observation set. The left-hand side of the above inequality is called the "test statistic". The test statistic is compared to a value, $z_{critical}$, which is the threshold for concluding significant difference at a particular confidence level. Typically, values for $z_{critical}$ are obtained from a table, organized according to the number of independent measurements and the desired level of confidence. Such a table can be found in any textbook on statistics and is provided in this note as Table 1.

N	$z_{critical}$ Confidence (2-sided)		
	95%	99%	99.90%
2	4.303	9.925	31.600
3	2.776	4.604	8.610
4	2.447	3.707	5.959
5	2.306	3.355	5.041
6	2.228	3.169	4.587
7	2.179	3.055	4.318
8	2.145	2.977	4.140
9	2.120	2.921	4.015
10	2.101	2.878	3.922
11	2.086	2.845	3.850
12	2.074	2.819	3.792
13	2.064	2.797	3.745
14	2.056	2.779	3.707
15	2.048	2.763	3.674
16	2.042	2.750	3.646
21	2.021	2.704	3.551
26	2.009	2.678	3.496
31	2.000	2.660	3.460
41	1.990	2.639	3.416
51	1.984	2.626	3.390
61	1.980	2.617	3.373

Table 1. Critical values of the test parameter for the Student's t-test required to conclude significant difference[3]. These values are for 2-sided comparisons which make no presumption about the value of the second mean relative to the first.

KEYSIGHT TECHNOLOGIES

The following example illustrates how the Student's t-test may be used to interpret indentation measurements. Let us suppose that we perform 10 indentations on each of two materials (A & B) with the following results: For material A, the average hardness is 4.91 GPa with a standard deviation of 0.23 GPa. For material B, the average hardness is 5.11 GPa with a standard deviation of 0.21 GPa. The test-statistic is calculated as

$$\frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{N}}} = \frac{|4.91 - 5.11|}{\sqrt{\frac{(0.23)^2 + (0.21)^2}{10}}} = 2.031,$$

and this value is compared to the critical values for significant difference. For $N = 10$, we find that $z_{critical}$ is 2.101 at the level of 95% confidence and even larger for greater confidence levels. Since the value of the test statistic (2.031) is less than the value of $z_{critical}$, we accept the null hypothesis. In other words, even though we obtain an average hardness for material B which is 0.20 GPa greater than that obtained for material A, we have no justification for concluding that material B is actually harder than material A. The fact that the test statistic is less than $z_{critical}$ at the level of 95% confidence tells us that if two observations sets (N=10 for each) were in fact drawn from the same population, we would expect this degree of variation in more than 5% of cases. Thus, because there is at least a 5% chance that these two observation sets could have come from the same population, we continue in our presumption that material A and material B have the same hardness.

Now if the hardness of material B were 5.17 GPa with the same standard deviation (0.21 GPa), then the value of the test statistic would be 2.640. Thus, we conclude that material B is harder than material A at the level of 95% confidence, but not at the level of 99% confidence. In order to conclude significant difference at the level of 99.0%, the difference in average hardness between material A and material B would have to be more than 0.28 GPa. At the level of 99.9% confidence, the difference would have to be 0.39 GPa.

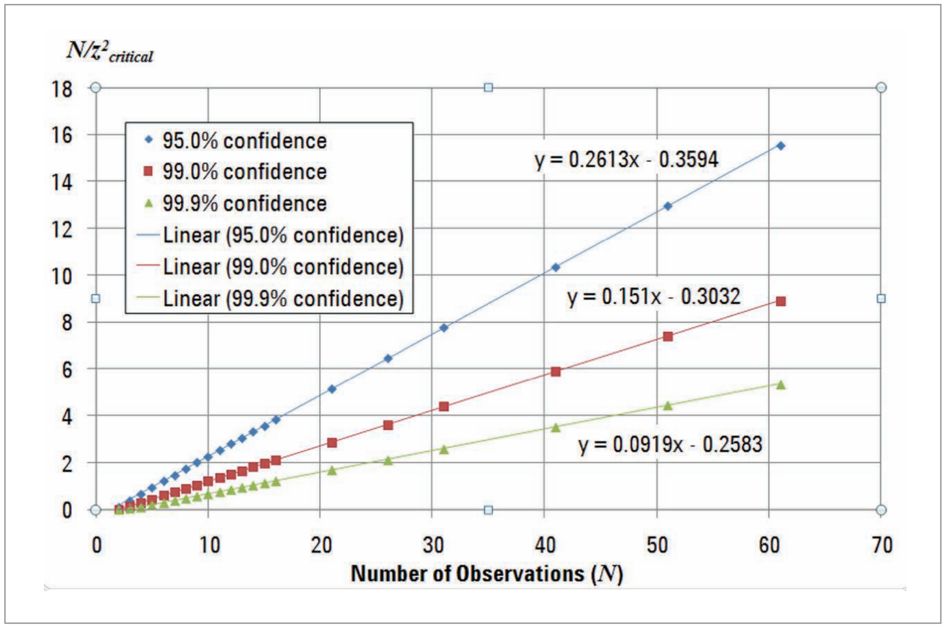

Figure 1. Approximate linear relationships between $N/z_{critical}$ and N for three different confidence levels, used to solve the Student's t-test for N. Values for this plot are calculated using values from Table 1.

This is the regular use of the Student's t-test: Given two sets of observations, the experimenter uses the Student's t-test in order to determine whether the two sets are significantly different, presumably as the result of some controlled (independent) variable.

However, the Student's t-test can also be used in the experimental design phase in order to predict the number of observations which must be made in order to be sensitive to a given difference at a given confidence level. First, we note that both the test statistic and $z_{critical}$ depend on N. The test statistic increases with N, and $z_{critical}$ decreases with N. Thus, more observations increase the sensitivity to significant difference. Given two observation sets from slightly different populations, one may be able to conclude significant difference at a particular confidence level if N= 20, but not if N=10. Thus, if we solve the Student's t-test for N, then the resulting expression would tell us the number of observations we must make in order to be sensitive to a given difference at a given confidence level. This is the motivation behind this application note.

Theory

In order to use the Student's t-test to predict the number of necessary observations, we must solve it for N. The solution is not as trivial as it might seem, because $z_{critical}$ depends on N. We begin by making some useful simplifications. First, we assume that $\overline{x_1}$ is greater than $\overline{x_2}$, and we express the ratio of $\overline{x_2}/\overline{x_1}$ as the factor F:

$$F = \overline{x_2}/\overline{x_1};\ F < 1. \tag{2}$$

The requirement of $F < 1$ is no real restriction. Because we assume two-sided comparisons[1], the two observation sets can always be ordered so that the set with the larger average is identified as "Set 1". Further, we assume that for both observation sets, the standard deviation is a constant fraction, q, of the mean:

$$q = \sigma_1/\overline{x_1} = \sigma_2/\overline{x_2}. \tag{3}$$

With these simplifications in mind, we square both sides of the t-test inequality[2] and isolate the parameters which depend on N on the left-hand side:

$$\frac{N}{z^2_{critical}} > \frac{\sigma_1^2 + \sigma_2^2}{(\overline{x_1} - \overline{x_2})^2}. \tag{4}$$

1. We are using the more conservative "two-sided" version of the Student's t-test which makes no presumption about the value of the second mean relative to the first.
2. Squaring both sides holds no ambiguity, because both sides of the inequality are positive.

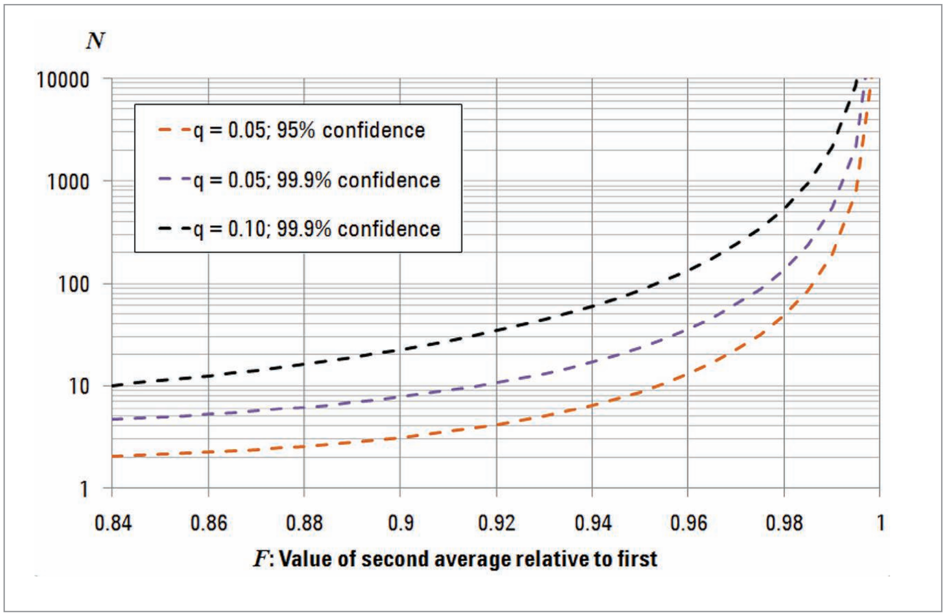

Figure 2. Student's t-test, solved for N, for three exemplary situations. Lines represent the equalities which limit inequality (Equation 8). Thus, N must be greater than the relevant curve.

Then, we apply the aforementioned simplifications to get

$$\frac{N}{z^2_{critical}} > \frac{q^2\bar{x}_1^2 + F^2 q^2 \bar{x}_1^2}{\bar{x}_1^2(1-F)^2} \qquad (5)$$

or more simply,

$$\frac{N}{z^2_{critical}} > \frac{q^2(1+F^2)}{(1-F)^2}. \qquad (6)$$

The above expression reveals the value of our simplifications: the right-hand side of the inequality is independent of the absolute values of the averages and the standard deviations. The left-hand side of the inequality is handled by elucidating the relationship between ($N/z^2_{critical}$) and N. Figure 1 shows a plot of the parameter ($N/z^2_{critical}$) vs. N for three common confidence levels. The values for this plot are calculated from those provided in Table 1. Figure 1 clearly reveals an approximately linear relationship between ($N/z^2_{critical}$) and N. Table 2 summarizes the values for

Confidence Level	m	b
95.0%	0.2613	-0.3594
99.0%	0.1510	-0.3032
99.9%	0.0919	-0.2583

Table 2. Linear best-fit constants for $N/z^2_{critical}$ vs. N (Figure 1).

slope and intercept (m and b, respectively) for the three relevant confidence levels.[3] Values for m and b for other confidence levels can be determined easily in the same way. Thus, the left-hand side of the inequality can be expressed as a linear function of N, making the inequality

$$mN + b > \frac{q^2(1+F^2)}{(1-F)^2} \qquad (7)$$

from which we derive our criteria for N:

$$N > \frac{1}{m}\left[\frac{q^2(1+F^2)}{(1-F)^2} - b\right]. \qquad (8)$$

Figure 2 illustrates the functionality of the criteria for a few exemplary situations. The plotted curves in Figure 2 are the equalities for expression (Equation 8); for the inequality to be met, N must lie in the space above the relevant curve.

Discussion

The form of our expression (Equation 8) for N confirms intuition: the number of observations required to adequately compare two normal populations ought to depend merely on the difference in the means (quantified by F), the variance (quantified by q^2), and the confidence level (quantified by m and b). If the variance is large, then N must be correspondingly large. Further, N increases as F approaches unity, which is as it should be: more observations are required in order to distinguish means which are very close together. Finally, if the two populations are in fact identical, then there is no value of N which is large enough to distinguish them.

There are two ways to use Figure 2 (or a similar plot generated with the appropriate confidence level and value for q). First, we can use this plot to predict how many observations we must make in order to be sensitive to a particular difference in means. For example, let us assume that we wish to work at the level of 99.9% confidence and we expect the standard deviation to be 5% of the mean (q = 0.05). Under these conditions, if we wish to discern significant difference in two observations sets for which the means differ by 2%, then we must make more than 138 independent observations, because this is the value of the middle curve (representing our conditions) at F = 0.98. Another way to use this plot is to determine the maximum value of F for which significant difference can be discerned for a given number of observations. Again, using the middle curve (q = 0.05; 99.9% confidence), we can see that if we make 10 observations of each population, we will be unable to conclude significant difference if the second average is more than 91% of the first average, because this is the point at which the middle curve crosses the threshold N = 10.

This analysis is indifferent to the physical cause of significant difference in the observed parameter. In experimentation, the independent variable is that parameter which is purposely and systematically varied order to understand its effect on the observed dependent variable. However, other variables which are not deliberately controlled may also influence the dependent variable. The independent variable of a hardness test might be something like tempering time. The dependent variable is hardness. Other variables which may influence the measured hardness might include measurement temperature or the rigidity of the test frame. The well designed experiment minimizes the influence of all physical variables other than the independent

3. Although clearly N must be zero when $N/z_{critical}$ is zero, we have a non-zero intercept, because the relationship is not linear (or meaningful) for N< 3.

variable, but no experiment is perfect. The Student's t-test will discern significant difference (or not) regardless of whether that difference is due to the independent variable or other uncontrolled variables. The degree to which the influence of these other variables can be minimized sets practical limits on F. One cannot blindly increase sensitivity by increasing N, because eventually, one may become sensitive to significant differences which are caused by variables other than the independent variable. Thus, the experimenter should wisely choose F – the expected ratio of means – to include sensitivity to the independent variable, but exclude sensitivity to other variables. For example, if normal variations in testing temperature may cause a 1% variation in the measured hardness, then one must be content with $F < 0.99$. One way to establish reasonable limits on F is to compare the means from two large observation sets acquired under conditions which the experimenter believes to be identical (or as much so as possible).

By virtue of its speed, Express Test dramatically improves the ability to detect significant differences in Young's modulus and hardness, relative to typical nanoindentation technology. Let us say that it takes 10 minutes to perform 10 nanoindentation tests at 10 different sites (1 minute per site) on each of two materials. If the standard deviation is 5% of the mean ($q = 0.05$) and we employ a confidence level of 99.9%, then we can detect significant difference if the two means differ by more than 9%. In the same amount of time (10 minutes or 600 seconds) we can perform 600 measurements with Express Test on each material, which implies that with the same standard deviation and confidence level, we can detect significant difference if the two means differ by only 1%. Thus, for a given testing time, Express Test dramatically improves sensitivity to significant difference.

Conclusions

The Student's t-test is used in an uncommon way to predict the number of observations (N) which must be made in order to be sensitive to a given difference at a given confidence level. Subject to a few simplifications, N depends on three things: the difference in means one wishes to sense (F), the normalized variance (q^2), and the desired confidence level. This analysis is appropriate for any kind of experimentation to which the Student's t-test might apply. With respect to nanoindentation, this analysis illuminates the benefits of the ultra-fast testing afforded by the Express Test option for Keysight's G200 NanoIndenter. Because it allows many more independent observations in a given time frame, Express Test dramatically improves sensitivity to significant difference.

References

1. Hay, J., "Rapid Characterization of Elastic Modulus and Hardness via Express Test," Keysight Technologies, Inc., 2014, Document No: 5991-0123EN, Date Accessed: December 17, 2012; Available from: http://cp.literature.keysight.com/litweb/pdf/5991-0123EN.pdf.
2. Hay, J., "Revolutionary New Keysight Express Test Option for G200 NanoIndenters," Keysight Technologies, 2014, Document No: 5990-9948EN, Date Accessed: April 9, 2012; Available from: http://cp.literature.keysight.com/litweb/pdf/5990-9948EN.pdf.
3. "Student's t-distribution," Wikipedia, December 17, 2012; Available from: http://en.wikipedia.org/wiki/Student's_t-distribution