A variant of the student's *t*-test for data of varying reliability

Suril B. Sheth[*] and Bhavin R. Sheth[&]

[*]Carnegie Vanguard High School, Houston, TX, 77019

[&]Department of Electrical & Computer Engineering, University of Houston, Houston, TX 77204-4005

Number of pages: 12; Number of figures:0; Number of Tables: 5

Corresponding author: Suril B. Sheth

Carnegie Vanguard High School

Houston, TX 77019

email: surilbsheth@gmail.com

Abstract

The student's t-test has been a workhorse of statistical testing and is used to determine if two sets of sampled data are significantly different from one another, in a statistical sense. The samples of the data may be individual samples or the means – or some overall summary statistic – of independently acquired subsets of data (e.g. data from individual observers, neurons, or baseball games). The various subsets of data acquired that go into computing the t-statistic are likely to be of differing reliability on account of either different variances or of different numbers of subsamples corresponding to each subset; while all data are given equal weight in a standard t-test, the variation in data reliability across subsets of data needs to be accounted for. Solutions based on mixed model methods and Monte Carlo simulations exist, which do factor data reliability in computing statistics. However, no such extension exists for the ubiquitous student's t-test. Our proposal is a novel variant of the student's t-test that incorporates these issues and adopts a simple but effective alteration in the design that accounts for differing levels of data reliability. Specifically, we weighted each data subset by the inverse of the variance of the data contained therein, a measure that has been used in studies of Bayesian cue combination, or, in the absence of information about variance, by the relative proportion of the overall data contained in the subset. The changes proposed here extend the applicability of the student's t-test to a wider array of data sets.

Introduction

Data acquisition is a process in which the best-planned experiments can often get waylaid by random events, which affect the underlying reliability of the data. One way around this issue is to simply discard data that are not of the highest reliability and work with limited data that we know are of extremely high reliability. However, one might argue that data acquisition is a process that is inherently noise-ridden and unreliable, working with data that are not perfectly reliable is a fact of life, discarding data that are at least partly reliable reduces the effect size and statistical power, and data that are at least partly reliable also provide some useful information and must not be discarded.

Consider the following example from signal processing. A voltmeter measures the voltage in an electrical circuit, and the requirement is for the voltage to be above a prescribed value. The signal is measured multiple (albeit limited) number of times to get a more accurate value of signal voltage; each time the analog signal is digitized by one of a bank of analog to digital converters (ADC); the resolution, i.e. number of bits N, of each ADC differs. The signal to noise ratio (SNR) of the quantized digital signal varies linearly as a function of N. The average digital voltage computed needs to account for the differences in resolution among the different ADCs before being statistically compared with the prescribed threshold.

Moreover, some experiments are natural, "real world" experiments that cannot be controlled and inevitably generate data that have different amounts of reliability. Here are two examples from education and neuroscience (one can think of analogous examples from other areas).

Suppose the state introduces a new standardized test and a school has recently instructed its teachers to teach to the test, whereas a second neighboring school within the same district has done no such thing. In order to find out how well the teachers have succeeded in teaching to the test, one needs to compare mean test scores of the students in the teachers' classes in each of the two districts. Classes can be and are usually of different sizes – and this is not typically under the teacher's control; the reliability of the mean test score obtained from a class size of 10 students versus a class size of 50 students has to be accounted for.

A second example is from neuroscience. Researchers want to examine the relationship between consolidation of overnight learning and power contained in the delta band of frequencies during slow-wave sleep (SWS). In order to find out delta power in SWS, one computes the power contained in every single 30 second long epoch of SWS and then take the average across all such epochs (alternatively, one can combine all the 30 second epochs of SWS across all sleep cycles and then compute the power). Here, different individuals have different amounts of SWS and the variation in SWS amounts can be large and is not under the experimenter's control. This variation in SWS amounts across individuals is a factor that has to be accounted for in developing statistical tests. Thus, when acquired under different conditions with different levels of reliability or averaged over different numbers of repetitions, data are not equally reliable and the differing levels of reliability is an important factor that must be taken into consideration in conducting statistical tests.

Therefore, a reasonable alternative is to retain the partly reliable data and weigh the reliability of the data relative to the reliability of other data in statistical tests. There are several avenues available including mixed model statistics and Monte Carlo methods.

The student's t-test is arguably the most commonly used test of statistical significance, and clearly the first one taught in AP statistics classes in the country. However, as it stands, the t-test cannot account for the real world differences in reliability. Here, we offer a modification to the t-test that takes into account the reliability of disparate data.

Methods

The t-statistic is defined by t = Z /s where Z and s are functions of the data; Z is a measure of the difference in the means in units of standard deviation of the sample; s is the standard error of the mean. More generally, $t = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{N}}$ where $\bar{X}$ is the sample mean of a sample $X_1, X_2, \ldots . X_N$ of size N, μ is the population mean, and σ is the population standard deviation of the data. Samples $X_1, X_2, \ldots . X_N$ can be individual samples or may themselves be means of independent samples of the data, i.e. each X_i is the mean of samples $X_{i1}, X_{i2}, \ldots . X_{in_i}$, where n_i is the number of samples. The formula for the population standard deviation is as follows

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

a) In the latter case, the reliabilities of the individual samples are likely to differ, which is otherwise ignored in the conventional formula for the t-statistic. We can define reliability r_i of the ith sample as the inverse of the variance σ_i^2 of the ith sample data, i.e.

$$r_i = 1/\sigma_i^2.$$

Note that [1-3] have used a similar Bayesian formula for weighting the relative reliabilities of

sources. We define a new reweighted mean $\tilde{X}$, which is defined as the following:

$$\tilde{X} = \vec{W}.\vec{X} = \sum_{i=1}^{N} w_i X_i$$

where

$$\vec{W} = [w_1, w_2, \ldots, w_N],$$

$$\vec{X} = [X_1, X_2, \ldots, X_N],$$

$$w_i = \frac{r_i}{\sum_{j=1}^{N} r_j}.$$

Note that

$$\sum_{i=1}^{N} w_i = 1$$

The revised weighted population standard deviation $\tilde{\sigma}$ is as follows

$$\tilde{\sigma} = \sqrt{\frac{N}{N-1} \sum_{i=1}^{N} w_i (X_i - \tilde{X})^2}$$

The newly revised t- statistic $\tilde{t}$ is now given by

$$\tilde{t} = \frac{(\tilde{X} - \mu)}{\tilde{\sigma}/\sqrt{N}}$$

b) In cases where the individual data points $X_{i1}, X_{i2}, \ldots . X_{in_i}$ are not recorded or available for some reason, we designed a minor variant in which the reliability of each datum X_i is (proportional to) the number of the number of data points n_i that went into the calculaStion of X_i, i.e.

$$r_i = n_i$$

The remaining calculations for $w_i, \tilde{X}, \tilde{\sigma}, \tilde{t}$ are identical to a).

Results

Here, we analyze data inspired by research and demonstrate a numerical application of the t-test variant to the data. Liu and Sheth [4] studied the response of the brain to pure tones (1 kHz) of varying intensity (53/58/63 dB SPL) in wake and in stage II sleep. In particular, the study focused on the P200 component of the electroencephalography (EEG) response, and asked if the amplitude of the P200 differed between the two states for different sound intensities. Six subjects participated in the experiment. For each individual subject, the mean P200 amplitude was computed for each state by averaging data from 300 trials / state. The difference in mean P200 amplitudes in wake and SII sleep (P200$_{wake}$ − P200$_{SII}$) across all six subjects can be compared with zero difference in a classical paired t-test paradigm.

Table 1: P200 ERP amplitudes in response to different sound intensities

Sound intensity	53 dB			58 dB			63 dB	
Subject	Wake (µV)	SII (µV)		Wake (µV)	SII (µV)		Wake (µV)	SII (µV)
S1	6.67	2.81		6.57	7.38		11.87	7.38
S2	6.61	3.55		9.64	2.94		7.42	11.95
S3	5.06	2.85		5.20	5.15		11.17	5.82
S4	6.22	3.01		7.16	4.83		10.94	8.24
S5	4.09	5.22		6.88	6.29		6.36	4.67
S6	6.44	4.82		8.75	6.75		5.96	9.88

Using a classic paired t-test paradigm, for the 53 dB sound intensity, the calculations yielded the following results summarized in Table 2:

Table 2: Calculations in classic paired t-test paradigm

Sound intensity	53 dB	58 dB	63 dB
	$\bar{X}_{\text{Wake-SII}} = 2.14\ \mu V$	$\bar{X}_{\text{Wake-SII}} = 1.81\ \mu V$	$\bar{X}_{\text{Wake-SII}} = 0.96\ \mu V$
	$\sigma_{\text{Wake-SII}} = 1.78\ \mu V$	$\sigma_{\text{Wake-SII}} = 2.67\ \mu V$	$\sigma_{\text{Wake-SII}} = 4.22\ \mu V$
	$t(5) = \dfrac{2.14}{1.78/\sqrt{6}} = 2.94$	$t(5) = \dfrac{1.81}{2.67/\sqrt{6}} = 1.66$	$t(5) = \dfrac{0.96}{4.22/\sqrt{6}} = 0.56$

	$p = 0.032$	$p = 0.158$	$p = 0.602$

The above calculations did not take into account the variable reliabilities of the individual subject data. The numbers of trials during which the individual is awake and is in SII sleep are given below in Table 3. As can be seen, the numbers of trials vary across subjects.

Number of trials	53 dB			58 dB			63 dB	
Subject	Wake	SII		Wake	SII		SII	Wake
S1	620	715		611	565		309	323
S2	524	389		405	312		731	497
S3	320	509		529	416		420	334
S4	458	334		334	377		331	302
S5	371	409		312	459		571	402
S6	315	402		354	423		317	503

The variant proposed here takes the variation in numbers of trials (or variance, for that matter, not shown here) into account in a new t-test paradigm. The results of the calculations are given in Table 4 below.

Table 4: Calculations in new paired t-test paradigm

Sound intensity	53 dB	58 dB	63 dB
	$\tilde{X}_{\text{Wake-SII}} = 2.35\ \mu V$	$\tilde{X}_{\text{Wake-SII}} = 1.49\ \mu V$	$\tilde{X}_{\text{Wake-SII}} = 0.29\ \mu V$
	$\tilde{\sigma}_{\text{Wake-SII}} = 1.76\ \mu V$	$\tilde{\sigma}_{\text{Wake-SII}} = 2.61\ \mu V$	$\tilde{\sigma}_{\text{Wake-SII}} = 4.34\ \mu V$
	$\tilde{t}(5) = \dfrac{2.35}{1.76/\sqrt{6}} = 3.26$	$\tilde{t}(5) = \dfrac{1.49}{2.61/\sqrt{6}} = 1.39$	$\tilde{t}(5) = \dfrac{0.29}{4.34/\sqrt{6}} = 0.16$
	$p = 0.022$	$p = 0.222$	$p = 0.878$

Here, the re-weighting of data according to relative reliability changed the t-statistics and p-values slightly – t statistic changed from 2.94 to 3.26, 1.66 to 1.39, and 0.56 to 0.16 for 53 dB, 58 dB, and 63 dB sound intensities respectively. The changes were small but more in line with the underlying data and the data acquisition process that produced them.

Our proposed variant can also be used in an unpaired setting as well. For this, we assume, for illustrative purposes, that the data for wake and SII were acquired from different subjects,

which calls for an unpaired t-test for statistical testing. The t-statistic used for an independent two-samples test when both distributions have the same variance – here, the sample size is small enough that the equal variance assumption is not violated – is known to be given by the following formula $t = \dfrac{\overline{X_1} - \overline{X_2}}{s_p \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$, where s_p is the pooled standard deviation of the two samples

and $s_p = \sqrt{\dfrac{(N_1-1)\tilde{\sigma}_1{}^2 + (N_2-1)\tilde{\sigma}_2{}^2}{N_1 + N_2 - 2}}$, N_1 is the size of the first sample (wake data, in the present case), and N_2 is the size of the second sample (SII data, here)[5]. The calculations are given in Table 5 below.

Table 5: Calculations in new unpaired t-test paradigm

53 dB	58 dB	63 dB
$\overrightarrow{W_{Wake}}$ $= [.24, .20, .12, .18, .14, .12]$ $\overrightarrow{W_{SII}}$ $= [.26, .14, .18, .12, .15, .15]$	$\overrightarrow{W_{Wake}}$ $= [.24, .16, .21, .13, .12, .13]$ $\overrightarrow{W_{SII}}$ $= [.22, .12, .16, .15, .18, .17]$	$\overrightarrow{W_{Wake}}$ $= [.12, .27, .16, .12, .21, .12]$ $\overrightarrow{W_{SII}}$ $= [.14, .21, .14, .13, .17, .21]$
$\tilde{X}_{\text{Wake}} = 5.99\ \mu V$ $\tilde{X}_{\text{SII}} = 3.60\ \mu V$	$\tilde{X}_{\text{Wake}} = 7.19\ \mu V$ $\tilde{X}_{\text{SII}} = 5.80\ \mu V$	$\tilde{X}_{\text{Wake}} = 8.56\ \mu V$ $\tilde{X}_{\text{SII}} = 8.30\ \mu V$
$\tilde{\sigma}_{\text{Wake}} = 1.00\ \mu V$ $\tilde{\sigma}_{\text{SII}} = 1.05\ \mu V$ $s_p = \sqrt{\dfrac{(6-1)*1.00^2 + (6-1)*1.05^2}{6+6-2}}$ $= 1.03$	$\tilde{\sigma}_{\text{Wake}} = 1.63\ \mu V$ $\tilde{\sigma}_{\text{SII}} = 1.53\ \mu V$ $s_p = \sqrt{\dfrac{(6-1)*1.63^2 + (6-1)*1.53^2}{6+6-2}}$ $= 1.58$	$\tilde{\sigma}_{\text{Wake}} = 2.50\ \mu V$ $\tilde{\sigma}_{\text{SII}} = 2.81\ \mu V$ $s_p = \sqrt{\dfrac{(6-1)*2.50^2 + (6-1)*2.81^2}{6+6-2}}$ $= 2.66$
$\tilde{t}(5) = \dfrac{5.99 - 3.60}{1.03 * \sqrt{\frac{1}{6} + \frac{1}{6}}} = 4.04$	$\tilde{t}(5) = \dfrac{7.19 - 5.80}{1.58 * \sqrt{\frac{1}{6} + \frac{1}{6}}} = 1.53$	$\tilde{t}(5) = \dfrac{8.56 - 8.30}{2.66 * \sqrt{\frac{1}{6} + \frac{1}{6}}} = 0.16$
$p = 0.010$	$p = 0.187$	$p = 0.876$

Discussion

The student's t-test has had a rich history. Since being designed by William Sealy Gossett for determining if two sets of data are significantly different from each other [6], it has been extended to other important cases such as unpaired (independent) and paired (correlated) two-

sample t-tests, tests for when the size of the two samples is not equal, Welch's t-tests for when the two samples have unequal variance and/or unequal sample size [7], and a multivariate Hotelling's test for multiple, often correlated, measures within the same sample [8].

Here, we propose a heteroaxiopistíc (hetero = different, axiopistía = reliability) variant of the student's t-test so as to factor the relative reliabilities of the samples in the computation of the t-statistic. Reliability of data is an important variable and regularly factors into statistical testing and with the introduction of said variant here, data reliability can now factor into the popular students t-test as well. The methods described above take into account reliability of one set of samples, on a one-sample t-test. In the event of a comparison between two sets of samples, such as in a two sample t-test, the exact same methods as detailed above can be replicated for a second set of samples.

There are alternative powerful solutions to the problem of combining fixed and random effects in a mixed model [9]. Hierarchical linear effects models assume that data that are being analyzed are drawn from a hierarchy of different populations whose differences relate to that hierarchy [10]. Fixed and random effects typically refer to the population average and subject-specific effects, respectively and these effects are modeled using a classical matrix notation and fitted using an expectation maximization algorithm [11] where the variance components are treated as nuisance parameters that one does not care about but has to nonetheless account for. There are clear benefits to a mixed model approach, including the fact that it can deal with missed values or measurements in the data with remarkable ease – a likely possibility in data acquisition, and mixed models have justifiably formed the basis for a large amount of statistical research in recent

years. The present approach to designing a heteroaxiopistic variant of the student's t-test provides, under certain conditions, a simpler and yet effective alternative that is more accessible to the end-user than more sophisticated mixed model approaches.

References

1. Hillis JM, Ernst MO, Banks MS, Landy MS (2002) Combining sensory information: mandatory fusion within, but not between, senses. Science 298: 1627-1630.
2. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415: 429-433.
3. Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14: 257-262.
4. Liu S, Sheth BR (2009) Discrimination of behaviorally irrelevant auditory stimuli in stage II sleep. Neuroreport 20: 207-212.
5. Fadem B (2008) High-Yield Behavioral Science. Hagerstown, MD: Lippincott Williams & Wilkins.
6. Student (1908) The Probable Error of a Mean. Biometrika 6: 1-25.
7. Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. Biometrika 34: 28–35.
8. Hotelling H (1931) The Generalization of Student's Ratio. Annals of Mathematical Statistics 2: 360-378.
9. Laird NM, Ware JH (1982) Random-Effects Models for Longitudinal Data. Biometrics 38: 963–974.
10. McLean RA, Sanders WL, Stroup WW (1991) A Unified Approach to Mixed Linear Models. The American Statistician 45: 54–64.
11. Lindstrom ML, Bates DM (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. JASA 83: 1014–1021.