# From ASA Statistical Consulting Section forum 2019-04

# Regarding T-Tests

Gurtej Bains11 days ago

Hello everyone, I have a question regarding T-Test. I am attempting to do a T-Test for two grou...

- 1. Regarding T-Tests

0 Recommend

Gurtej Bains

Actions Options Dropdown

Posted 11 days ago

Reply Inline Options Dropdown

Hello everyone,

I have a question regarding T-Test.

I am attempting to do a T-Test for two groups (test and control) to check if the means sales changed after a promotion was given to the test group.

In order to do the T-Test for comparing the mean sales of both groups, do I have to ensure that the sales in both groups are normally distributed? If yes, any prefered method of fixing the distribution (log sales)? Using JMP for this analysis.

Thanks for your help.

Gurtej

---
Gurtej S. Bains
Manager, Customer Insights
The Home Depot
---

- 2. RE: Regarding T-Tests

0 Recommend

Jon Shuster

Actions Options Dropdown

Posted 11 days ago

Reply Inline Options Dropdown

Not at all. But you should use a Satterthwaite correction (part of SAS Proc ttest). This corrected t-test is extremely robust, and much more so than the Wilcoxon test, which is a test for equal distributions, not equal means. The only issue that is a problem is a tendency to outliers. Money may be in that vogue. Perhaps logs could help assuming there are no zeros. The confidence interval would be transformed back via antilogs to a ratio of scales.

Provide a confidence interval for your results in any case.

In the 100-th year anniversary of the t-test (2008), Dr. Stephen Senn wrote on the amazing robustness of the t-test to lack of normality.

Try this: The exponential distribution is highly skewed. Even with 5 subjects per group, the t-test and confidence interval with a Satterhwaite correction work amazingly well.

Best,

Jon Shuster

See this additional support for the Satterhwaite corrected t-test :

Shuster JJ. Diagnostics for assumptions in moderate to large simple clinical trials: do they really help? Statistics in Medicine. 2005 Aug 30;24(16):2431-8. PubMed PMID:
15977289.

Shuster JJ. Student t-tests for potentially abnormal data. Stat Med. 2009 Jul 20;28(16):2170-84. doi: 10.1002/sim.3581. PubMed PMID: 19326398; PubMed Central PMCID: PMC3666168.

---
Jon Shuster
---

Original Message

- 3. RE: Regarding T-Tests

0 Recommend

John Dawson

Actions Options Dropdown

Posted 11 days ago

Reply Inline Options Dropdown

Assuming that you're working on the change scores (here, differences between pre- and post-intervention sales in each group) then there are a few things to consider:
1) Are your empirical distributions of sales change score values Normal-ish? Looking for unimodality (one-humpedness), symmetry and thin tails (few extreme values on either side of the empirical distributions).
2) How big are your sample sizes? With a big enough sample, a t-test will be valid because of the Central Limit Theorem. But "big enough" depends in part to your answer above -- the less "Normal" your populations are, the bigger your sample size needs to be before the CLT will kick in. As a rough rule of thumb, if your data are fairly Normal to begin with, t-test is OK with 20 in each group; less Normal but not too bad, 50 in each group; and otherwise you might want to use a nonparametric test or try a transformation (see 4 and 5 below).
3) If you do end up using a t-test, use the Welch t-test over the more commonly known Student's t-test. The plus of validity when there are unequal group variances more than offset the minus of slightly lessened power when the group variances are equal.
4) If you have very non-Normal change scores (e.g., very coarse data or very skewed data) then you might want to use a nonparametric test of different distributions. In this case, the corresponding NP test would be the two-sample Wilcoxon on the two sets of change scores (aka rank sum test, aka Mann-Whitney U-test).
5) Your change scores may also become Normal after transformation, allowing the use of a Welch t-test. Rather than eyeballing it or rigidly using a particular transformation out of hand, use the Box-Cox transformation procedure to determine which power you should raise your outcome values to in order to obtain data that best fit the assumptions of a linear model (the superclass to which the t-test belongs). The log is one possible decision outcome for this procedure, corresponding to lambda=0.

---
John Dawson
Assistant Professor
Texas Tech University
---

Original Message

- 4. RE: Regarding T-Tests

0 Recommend

Annette Gourgey

Actions Options Dropdown

Posted 11 days ago

Reply Inline Options Dropdown

This is about a different aspect of t tests, the equal variance assumption, but since people are suggesting alternate formulas: What do you think of the separate variance t test when the variances show a significant difference? I understand it's not actually t distributed. How widely is it actually used and how much confidence is there in it?

---
Annette Gourgey
CUNY
---

Original Message

- 5. RE: Regarding T-Tests

0 Recommend

John Dawson

Actions Options Dropdown

Posted 11 days ago

Reply Inline Options Dropdown

-- Student's t-test isn't "actually t distributed" when you don't have true underlying Normality either. They're both approximations.
-- Widely used enough and used confidently enough that it's the default in R when you call t.test().

---
John Dawson
Assistant Professor
Texas Tech University
---

Original Message

- 6. RE: Regarding T-Tests

0 Recommend

Eric Siegel

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

The unequal-variance t-test, a.k.a. Welch's t-test: how widely is it used and how much confidence is there in it? Back in the early days of gene-expression microarray analysis, when one would typically do 5,000 to 40,000 univariate t-tests on a data set, Terry Speed recommended using Welch's t-test in preference to the equal-variance (Student's) t-test. The computational reasons why were

obvious: one did not have to make 5,000 to 40,000 assumptions of equal variance. I believe that simulation studies from the time showed that Welch's t-test had better coverage performance in the microarray setting than Student's t-test. As a result, the overwhelming preference back then was for Welch's t-test. Let me hasten to add that, since that day, other kinds of null-hypothesis tests have been developed for the "-omics" setting that beat out Welch's t-test, but by extension, they also beat out Student's t-test.

Me myself, I have historically preferred Student's t-test, but I have used Welch's t-test when it was called for, and I have recently become interested in linear-models generalizations of Welch's t-test that use the Kenward-Roger method (rather than Satterthwaite's method) to calculate denominator degrees of freedom. The reason for my sudden interest is because a biologist colleague of mine told me within the past month that his university edition of JMP uses exclusively the Kenward-Roger method to determine denominator degrees of freedom for hypothesis tests in a linear mixed model. This claim indicates to me that there is at least one software developer out there that thinks unequal-variance methods are better than equal-variance methods when assembling a low-cost statistical software package for sale to grad students and beginning assistant professors.

---
Eric Siegel, MS
Biostatistics Project Manager
Department of Biostatistics
Univ. Arkansas Medical Sciences
---

Original Message

- 7. RE: Regarding T-Tests

0 Recommend

W. G. Konarasinghe

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Dear Gurtej,

Generally t-test is applied when; Normality is not met, sample sizes are small, population standard deviations are unknown etc.

Use the t-test for mean comparison with pooled standard deviation.

RGDS

---
W.G. Samanthi Konarasinghe( PhD)
Statistician /Senior Lecturer
Institute of Mathematics & Management,
Sri Lanka
---

Original Message

- 8. RE: Regarding T-Tests

0 Recommend

Stephen Smela

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

You might want to try a non-parametric approach like a Wilcoxon test. Using a log transform won't guarantee (approximate) normality.

---
Stephen Smela
Senior Data Scientist
United Health Group R&D
---

Original Message

- 9. RE: Regarding T-Tests

0 Recommend

Jon Shuster

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Not a good idea to use Wilcoxon. As I said, normality of the individual observations is not relevant, absent of a propensity for outliers. On Money related outcome, log dies very well unless there are zeros. The Satterthwaite corrected t-test is far more robust than the Wilcoxon, which does horribly when the means are the same but the distributions different, especially differential variability between the groups, a very common issue in $ outcomes. The central limit converges amazingly quickly and ultimately it is the approximate normality of the Satterthwaite standardized difference, not the individual data. The sample sizes both contribute to the convergence so if you have 2 samples of say 12 and 18, that is about as good as one sample of size 30, when it comes to the central limit.

We are very bad about not teaching the robustness of the T, as (1) There is no valid diagnostic test that can prove normality of your data beyond a reasonable doubt, or for that matter equal variance. (2) It is invalid to change horses in midstream. For example, testing for normality and then use a Wilcoxon if you reject and a t-test if you no not is unacceptable statistically. Your operating characteristics are only valid if they apply before you branch out. Such a process can grossly distort the study operating characteristics, and is therefore statistical miss-practice.

See the references I attached in my prior communication.

---
Jon Shuster
---

Original Message

- 10. RE: Regarding T-Tests

0 Recommend

Stephan Arndt

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Hi,
I have always found it odd to force fit the information (i.e., data) to fit the statistical assumptions rather than use a more tolerant method (e.g., Mann-Whitney) that is oftentimes more powerful.
Steve

---
Stephan Arndt
Professor
University of Iowa
---

Original Message

- 11. RE: Regarding T-Tests

0 Recommend

Gabriel Farkas

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Mann-Whitney is for 2 independent samples, while the original question pretty strongly leans toward the data being paired samples. I believe that's why others have suggested Wilcoxon as the nonparametric test to use (if one were to use a nonparametric test).

_____

Gabe Farkas
gfarkas@gmail.com

Original Message

- 12. RE: Regarding T-Tests

0 Recommend

Stephan Arndt

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Hi,

Sorry for the confusion.  I said Mann-Whitney instead of Wilcoxon, in "assumptions rather than use a more tolerant method (e.g., Mann-Whitney) that is oftentimes more powerful.".  However, e.g. stands for exempli gratia and means "for example."

Steve

---
Stephan Arndt
Professor
University of Iowa
---

Original Message

- 13.  RE: Regarding T-Tests

0 Recommend

Robert Kushler

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Gabriel,

This is not a paired data situation - it's a two group comparison of change scores.

Gurtej,

Some advice I haven't seen yet:  PLOT your data.  This will help you decide whether a transformation can help, whether there are outliers, etc.

Everyone,

I haven't seen any discussion of the (fairly recent) downgrading of p-values.  I recall one person suggesting use of a confidence interval, which I strongly endorse.

Regards,  Rob Kushler

---
Robert Kushler
---

Original Message

- 14. RE: Regarding T-Tests

0 Recommend

Frank Harrell

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Jon there is a wealth of information documenting the non-robustness of the t-test. So I wholeheartedly but respectfully disagree. Work by Rand Wilcox and others is recommended.

Frank

---
Frank Harrell
Vanderbilt University School of Medicine
---

Original Message

- 15. RE: Regarding T-Tests

0 Recommend

Jon Shuster

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Hi all:

I have looked at hundreds of cases where I put in my own non-normal distributions and ran the Satterthwaite-corrected t-test, the t-test, and the Wilcoxon (typically replicated 100,000 times each with common data--matched) They had common means and equal or unequal variance, equal or unequal sample sizes. By far the best overall was the Satterthwaite corrected t-test and by far the worst was the Wilcoxon test. You can try it too with my SAS macro on my website below, dedicated to the two-sample t-test. It handles continuous and discrete (or even mixed) distributions. In the real world, T is for terrific. The literature has criticized T for "Heavy-Tailed"

distributions, which are almost non-existent in practice.

You also can do great damage to your operating characteristics with any hybrid approach. (i.e. Some diagnostic test, which will likely be more harmful than helpful). It is always best to declare a primary plan for analysis up front, and stick to it, with sensitivity analyses fair game. But the primary sticks and the sensitivity is good for discussion in the paper.

Jon

Shuster, Jonathan

Ufl remove

Shuster, Jonathan
Professor Emeritus, Health Outcomes & Biomedical Informatics Phone:
692-0893 shusterj@ufl.edu Clinical and Translational Research Building
Mowry Road, Module 117-15 PO Box 100177 Gainesville, FL 32610-01
B.Sc. McGill University, Montreal M.Sc. McGill University, Montreal P
View this on Ufl >

---
Jon Shuster
---

Original Message

- 16. RE: Regarding T-Tests

0 Recommend

Frank Harrell

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Jon I think the generality of this finding is less than you think it is. In many cases (especially with a large amount of asymmetry), the standard deviation is not even an appropriate measure of dispersion, so the t-test (and the CLT) can't work. Witness the log-normal distribution where it is only appropriate to compute the SD on the log scale and not on the original scale.

Frank

---
Frank Harrell
Vanderbilt University School of Medicine
---

Original Message

- 17. RE: Regarding T-Tests

0 Recommend

Larisa Burke

Actions Options Dropdown

Posted 10 days ago

Reply Inline Options Dropdown

Jon - Thanks for the invaluable information of the t-test and normality! I will definitely use this in the future. I've always thought that the log(x+1) transformation is a good work around for when a the variable that needs to be transformed has zeros.

---
Larisa Burke
University of Illinois at Chicago
---

Original Message

- ## 18. RE: Regarding T-Tests

0 Recommend

Dennis Helsel

Actions Options Dropdown

Posted 9 days ago

Reply Inline Options Dropdown

I've had this discusson with engineers, biologists, chemists, hydrologists and geologists over the past 40 years. They are all taught in their 1 or 2 semester university course that the t-test is "robust", and therefore generally a great test. The folks deal with data that are largely lognormal, and usually have 20 or fewer observations per group. What they aren't told, and what Jon (whom I respectfully disagree with) is not mentioning, is that robustness only deals with Type 1 errors. The big problem with t-tests on non-normal, especially right-skewed data is its poor power. The folks I work with aren't really interested in a mean. They want to know if one group has higher values than the other. That is what the Wilcoxon type tests, paired or no, directly test. If someone is

actually interested in the mean rather then the previous more general statement, taking logs of their skewed data results in a test for difference in geometric means, and the confidence interval on the difference in logs maps to a CI on the ratio of geometric means. Neither is directly relevant to an arithmetic mean, but is often interpreted that way by the user because they were told that the transformation solves their problem and the t-test on logs is now a wonderful test for difference in means. Its quite a good test in that mode for differenece in medians (geometric means), but not arithmetic means. The bottom line is that if one is interested in an actual test for difference in means with skewed data, they either need a 'lot' of data to invoke the CLT or should run a permutation test. How much is 'a lot' is a function of the skewness of the data, and there was an American Statistician article with that formula back in the 1980s. When I last computed it back in the 80s for the data my clients dealt with, it was around 50-70 observations per group. There's a USEPA publication that recommends 100 observations to invoke the CLT, which I believe was based on skewed data of concentrations in soils. These details about the t-test are woefully lacking in the applied 1 and 2 semester courses taught in science and engineering departments. The result is that a lot of differences are missed, differences such as contamination in water, because a t-test was used for small data sets of skewed observations. And because variance is usually proportional to the mean in the data in these disciplines, and the two groups are often of the "one control site with low concentrations, one contaminated site with high concentrations" variety, variances are almost always very different. The Welch/Satterthwaite correct is a must in that situation, if using the t-test.

---
Dennis Helsel
Practical Stats LLC
---

Original Message

- 19. RE: Regarding T-Tests

0 Recommend

Rickey Carter

Actions Options Dropdown

Posted 9 days ago

Reply Inline Options Dropdown

This has been an interesting discussion. I do find some irony in what strikes me as a polarizing series of posts. The bright lines of right vs. wrong; $p<0.05$ vs not; etc. seem to remove much of the common ground and may lessen our ability to positively influence science.

I have often wondered about this problem. Through experience, I find the T-Test is rarely "the" definitive test but rather a starting point on the analysis. We are often faced with options where we can transform the data to fit our assumptions or alter the assumptions and work on the data from a more generalize perspective. If we find Y | X is skewed, for example, should we log it to make the T-Test assumptions more tenable? Or, could we ask the biologist about the skewness and perhaps learn if there are meaningful variables (e.g., sex- & age- effects) that are not included in the model that could explain this observation. Now we are in a regression framework with Y | X, C1, C2 etc where the adjustment covariates are working to also meet the distributional assumptions. Do we consider the robust variance estimator over degrees of freedom correction if the data set is "large"? Or more importantly, are the assumptions satisfied and are we actually starting to meaningfully model the system?

While p-value driven interpretation is going to be hard to break (perhaps this could be one of the hidden benefits of the AI efforts now), many true findings appear to be robust to the methods used for the analysis. I personally do not feel comfortable with test results that are only consistent with my alternative hypothesis if I do something a certain way. The potential for human bias is too great in these circumstances to have much faith in the findings.

These more editorial comments aside, I would love to see the community crowd source a robust application available out in the wild that users could explore the operating characteristics of these tests under a variety of data scenarios. I can visualize a user tracing a set of density curves and then having random data generated from those generalized tracings. The program could quickly analyze these simulated data using a variety of approaches and provide measures of statistical significance and effect sizes (point and interval estimates). As a bonus, we could use a source forge like approach to refine the interpretations of the results so that we can provide guidance on proper interpretations. The system could also store all of the scenarios tried. Some of these scenarios will be contrived. Some may match real data. As the system is used, we generate more and more generalizable knowledge about how these tests perform over many more scenarios than would ever be achievable by a single paper, or a century of papers. Visual representation of where the test performance diverges could be a powerful teaching tool. 3D modeling of these results to allow people explore at new depths the operating characteristics of

the tests would be fascinating. It would seem this would move us forward in the field and potentially highlight what I'm expecting is far more common ground than the recent series of comments would suggest.

If someone would be interested in co-developing such a system, I would enjoy that conversation.

I would also be interested in some actionable findings.  For example, suppose it is found the Wilcoxon test is approximately x% higher power in a certain scenario but we anticipate the need for parametric summaries to effectively communicate the results. How many additional experimental units are needed to achieve the same power if the analysis plan is designed with the T-Test instead of the WRS?

---
Rickey Carter
Professor of Biostatistics
Mayo Clinic
---

Original Message

- 20.  RE: Regarding T-Tests

0 Recommend

Andrew Ekstrom

Actions Options Dropdown

Posted 9 days ago

Reply Inline Options Dropdown

Something I did a few years ago was create an Excel sheet that took X1, S1 and N1 and generated 10,000 random values about X1, using S1/sqrt(N1) as the standard deviation. Then do the same thing for X2, S2 and N2. Then I set up a column to calculate the t-test given an X1 and X2 in a row of data. I then used that T-test result to determine how often someone would get "statistically significant results". The probability of getting "significant results" was about the same as the power of the original data set with X1, X2, S1, S2, N1 and N2. Say, P(Stat sig res) = Power +/- 2%.

From there, I looked at a Z-test. Z = (MOE - Difference)/Std Err gave a probability the results will not be statistically significant. So, 1 - P(Z) gave a value that was very close to the results of my simulations and very close to the power of T-test.

Now, when I give my stats students a question about the T-test, I ask them to calculate the probability others will get a "statistically significant result" and if that probability is good enough for them to trust the original results.

For that reason, I don't think reporting confidence intervals is any better than reporting a P-value. Both can be deceiving. (Sorry Dr Kushler.)

I'm going to have my Mgt Sci students run a simulation on their final exam to see how often Group A is greater than Group B given some set of X1, S1 and N1 vs X2, S2 and N2. Hopefully that will give some idea about a statistically significant result vs the probability one group is greater than the other...

---
Andrew Ekstrom

Statistician, Chemist, HPC Abuser;-)
---

Original Message

- 21. RE: Regarding T-Tests

0 Recommend

Laura Kapitula

Actions Options Dropdown

Posted 9 days ago

Reply Inline Options Dropdown

Interesting question about the t-test. I realize with questions such as this that I tend to want to take a modeling approach more than a testing approach (ideally).

So the question I have it not, "Is the independent t robust to assumption violations for testing for a difference in two means?". The question I have is, "Is the difference in population means the right parameter of interest for my applied problem?". Say, treatment A is better on average than treatment B for some outcome, but treatment A is way more variable, or perhaps treatment A patients have a distribution with a really long tail of people who were doing horribly. We might want to go with treatment B even if treatment A is better "on average". This is similar logic to using biased regression methods, we might take a bit of bias to have a smaller variance, if we want to make the best decisions/estimations/predictions.

So if I have a decent sample size and I plot the data and it looks really non-normal and/or there are big differences in variation. I tend to want to think about what model is best for the data so I can better understand the process and make better decisions.

Laura

Original Message

- 22. RE: Regarding T-Tests

0 Recommend

Frank Harrell

Actions Options Dropdown

Posted 9 days ago

Reply Inline Options Dropdown

I favor this approach: Bayesian Estimation Supersedes the t-test (BEST) - Online

Sumsar remove

Bayesian Estimation Supersedes the t-test (BEST) - Online
This page implements an online version of John Kruschke's Bayesian estimation supersedes the t-test (BEST) Bayesian model that can be used where you classically would use a two-sample t-test. BEST estimates the diff in means between two groups and yields a probability distribution over the difference.
View this on Sumsar >

Frank

---
Frank Harrell
Vanderbilt University School of Medicine
---

Original Message

- 23. RE: Regarding T-Tests

0 Recommend

Dewi Rahardja

Actions Options Dropdown

Posted 4 days ago

Reply Inline Options Dropdown

Dear All,

FYSA, perhaps the "**Roadmap**" in the following paper would be helpful:

https://digitalcommons.wayne.edu/jmasm/vol16/iss1/8/

Best Regards,

(Gaby)

---
Dewi Rahardja, PhD, MM, PhD
Statistician
U.S. Federal Agency
---

Original Message

- 24. RE: Regarding T-Tests

0 Recommend

Frank Harrell

Actions Options Dropdown

Posted 9 days ago

Reply Inline Options Dropdown

Amen to all that Dennis.  If you have a sample from a lognormal (0,1) distribution, the CLT is not accurate for computing a confidence interval for the mean until n > 50,000.

Power is really the issue, along with confidence interval coverage.  It is amazing how many people only look at type I error when considering the CLT.

A major issue for t- based methods is that if the data distribution is skewed or very heavy tailed, the standard deviation should not even be computed much less used in a test or confidence interval formula.

Frank

---
Frank Harrell
Vanderbilt University School of Medicine
---

Original Message