Two-sample t test

Carol Howald

2019-04-22'

Activities

- Comparing two groups
- Comparing confidence intervals
- Two sample t-test
- Instructor customizations:
  - Ambika Silva: Comparing two samples with confidence intervals activity & assessment

Learning objectives

Typical course objectives relating to the t-based test and confidence interval are:

- Confidence interval:
  - Compute from data the confidence interval on the difference of means.
  - Interpret the confidence interval in the context of the question intended to be addressed with the data.
- p-value:
  - State an appropriate null hypothesis
  - Draw a sketch relating the sampling distribution under the null to the observed difference between sample means, and marking the region of the distribution corresponding to the p-value. [NOTE IN DRAFT: TO POINT OUT AT WORKSHOP: We can add pedagogical displays as needed.]
  - FOR THE NEW DIAGRAM: Parametric diagram. Show the t-distribution. Variable versus probability density.
  - Compute the numerical p-value
  - Appropriately frame the result as "reject the null" or "fail to reject the null" (the only two valid outcomes of the test).
  - State a valid interpretation of result in the context of the question intended to be addressed with the data.
- Additional objectives:
  - Be able to translate a confidence interval into a simple, approximate statement about the p-value (e.g. $p > 0.05$ or $p < 0.05$).
  - Identify situations such as outliers that may call into question whether the results can be taken at face value. Know how to deal with such situtations.

NOTE: One-tailed versus two-tailed tests. ARTICULATION AGREEMENT: State of MD affinity group but no mechanism for organizing them or making sure they are consistent.

Additional resources

Orientation for instructors

The two-sample t-test is a staple of introductory statistics. In many courses, it is the most advanced topic taught. In others, it is followed by linear regression and/or ANOVA. At Howard Community College, the t-test is taught in the last 2 weeks of our course.

A t-test is part of the apparatus of *statistical inference*. The purpose of statistical inference is to guide valid conclusions about the *population* from a *sample*. The specific question underlying the t-test is how the mean value of some quantitative variable differs between two groups . . . in the population. Using samples from each of the two groups, the apparatus provides a way to calculate two closely related quantities:

1. A *confidence interval* on the difference between the two sample means. The purpose of a confidence interval is to provide a reasonable statement of how the difference in means in the population relates to the difference of means of the two samples.
2. A *p-value* which quantifies how plausible is a claim that in the population the two groups are the same and that the observed difference in sample means comes about by the chance variability stemming from random sampling.

For simplicity, we'll refer to both of these as the "t-test."

Role in statistical practice

The t-test is one of the most time-honored statistical methods. But it is also incapable of handling research problems of typical complexity in the modern era. Let's focus on where the t-test fails to apply:

1. There are two groups and a quantitative response variable, but there are also covariates. For instance, suppose we compare a drug and a placebo for the effect on lowering high blood pressure. In medical studies, sex and age are *always* covariates. There are likely others as well, e.g. smoking, race, the type of illness creating high blood pressure, and so on. To apply a t-test one needs to arrange to nullify such covariates. This could be done, for example, by specifying a particular, narrow group, such as white men in their 50s who do not smoke. But this may be much narrower than the population of interest in applying the results of the study, and we lose the opportunity to examine the role of covariates.

   Another way in which covariates can in principle be nullified is the random assignment of subject to each of two experimental groups. But relying on equivalence produced by random assignment is not robust. Even if you arranged an experiment in this way, you would still want to record values of covariates and adjust for them. In randomized clinical trials there are often people dropping out of the study or violating the experimental protocol (e.g., taking aspirin to deal with headaches). Randomization refers to an intent rather than necessarily an outcome.

2. There are multiple time points at which measurements are made. The classic case of before-and-after measurements can – neglecting covariates – be handled with a one-sample test, but this cannot be extended to multiple measurements over time as with before-during-after studies.

3. There are multiple tests. A conceptually simple situation might involve looking at the differential expression of a gene in two different groups. Again neglecting covariates, that might seem like a good setting for a t-test. But such genetic expression measurements are often done using micro-array or similar technology, which might involve hundreds or thousands of simultaneous measurements. Such situations render meaningless confidence intervals or p-values generated from a t-test.

So while there are simple situations in which a t-test is appropriate, it's a grave error to suggest that the t-test is representative of the range of concerns in contemporary research.

Conceptual pitfalls

A two-sample t-test is a special case of one-way analysis of variance (ANOVA) and produces the same results as would be optained from ANOVA.

There are three forms of two-sample t-test:

1. The equal-variance t-test, which is mathematically identical to ANOVA.
2. The paired t-test, which is really a one-sample test on differences.
3. The unequal-variance t-test which involves more intricate formulas and conceptual challenges such as non-integer degrees of freedom.

There's hardly ever a good reason to carry out an unequal-variance t-test. For one, it offers hardly any advantage over the equal-variance t-test. Such an advantage would be expressed in terms of the "power" of the test. Insofar as introductory courses do not introduce the concept of power, there's not even a way to explain why one might prefer one test to another. For another, insofar as the variance of the two groups differ, haven't you already established that the groups are different? Why worry about comparing the means – the distributions are clearly different.

The two-sample t-test is completely equivalent to simple regression. Just recode the two-level categorical grouping variable as zero and one, then treat the grouping variable quantitatively. But whereas simple regression is naturally seen as a special case of the more general methods of multiple regression, there is no path from the t-test to multiple variables (for instance the covariates mentioned in the previous section).

The "t" component of a t-test is relevant only for small data sets, say $n < 20$.

The t statistic is the square root of the more general F statistic, but applies only for situations where the degree of freedom in the denominator is 1.

One place where statistics instructors make use of t differently from F is that t can be handled as either a one-tailed or a two-tailed test, while F is always the equivalent of the two-tailed test. But keep in mind that one should always be suspicious of one-tailed tests. The *only* justification for a one-tailed test is to increase power, but since power is not usually a subject in intro stats, there is no meaningful way to explain what the potential benefit of doing one would be. And, there are large potential costs. Often one-tailed tests are used as a form of p-hacking. The New England Journal of Medicine has a nice explanation of why to avoid one-tailed test:

> Expectation of a difference in a particular direction is not adequate justification. In medicine, things do not always work out as expected, and researchers may be surprised by their results. For example, Galloe et al found that oral magnesium significantly increased the risk of cardiac events, rather than decreasing it as they had hoped. If a new treatment kills a lot of patients we should not simply abandon it; we should ask why this happened.
>
> Two sided tests should be used unless there is a very good reason for doing otherwise. If one sided tests are to be used the direction of the test must be specified in advance. One sided tests should never be used simply as a device to make a conventionally non-significant difference significant.

It's irresponsible to teach one-tailed tests as a purely mathematical topic without engaging their negative impact on research integrity. And the one-tailed test is of so little benefit even in legitimate settings that a much more reliable instruction would be to always use two-tailed tests.

Student pre-requisites

Students will need some background statistical knowledge to be able to follow lessons on the t-test.

- Basic:
  - Know the difference between a quantitative variable and a categorical variable. For a categorical variable, know the number of *levels* of the variable. Resources: Little App on jitter plots and the lessons on point plots and variable types
  - Be comfortable with graphical presentations showing a quantitative variable versus a two-level categorical variable. In this lesson, we use jitter plots. Helpful resources: Little App on jitter plots
  - Understand the distinction between "center" and "spread" of a distribution of values. Resources: Little App on center and spread and lessons on describing spread and the standard deviation.
  - Understand the process of sampling and the distinction between a population and a sample, and, correspondingly, a "parameter" and a "statistic".
  - Understand how a descriptive statistic is a summary of a *group* and combines many individual observations.
- Intermediate
  - Be aware central purpose of *statistical inference*, namely to draw valid conclusions about the *population* from a *sample*.
  - Understand that confidence intervals describe the uncertainty in a sample statistic due to sampling variation. Resources: Little App on resampling
  - Be familiar with the basic nomenclature and logic of hypothesis testing: null-hypothesis, test-statistic, sampling distribution under the null, observed value from the sample, p-value.

Creating an active classroom

See the document on general tips for creating an active classroom.

Some specific discussion topics/themes for t-tests.

1. A think/pair/share activity. Looking, say, at `income_poverty` versus `home_type`, there the two confidence intervals on the mean do not overlap. Respond to this prompt as best you can: Suppose a friend claimed that

a decent prediction of a person's income would be to say that it almost always falls within the confidence interval for the group the person belongs to. Is your friend right? Explain why or why not in terms that would make sense to a fellow student.

2. Have students discover for themselves the correspondence between the elements of the graphic in the little app and the statistical report in the "statistics" tab. See Carol's tasks 3 and 4 for wording.
3. After completing each lesson, form students into small groups to explore a new set of variables. This gives me the chance to circulate among the groups to provide feedback. After letting each group explore and analyze 15-20 minutes, I will give the groups a few minutes each to present their results. Their goal will be to incorporate the language correctly as they present their results.
    - Students often want to spend too much time just choosing variables. I need to give them a signal when it is time to commit and move on! I also need to assure them that finding out that the variables do not relate in the way they thought is still a valid investigation.

Assessment Items

1. Ask students to explore to use the t-test Little App to find variables that show a difference at p < 0.05.
2. Ask students to explore to find variables that produce as low a p-value as they can.
3. Figure out, for the variables selected in (1) and (2) whether larger sample size is associated with larger or smaller p values.
4. Use static graphics showing the data, confidence intervals on the means, and the t-interval, but where sometimes one or another of the intervals doesn't match with the data or where a mis-matched t-interval suggests a very different conclusion than the confidence intervals. Ask which graphs are self consistent. [NOTE IN DRAFT: We should create a set of these.]
5. Give students a picture of the graph in the Little App and the corresponding t-test report (in the "statistics" tab.) Draw arrows from each element of the report to the corresponding glyph in the graphic.

Looking forward

- Useful approximations
    - Checking whether the 95% confidence intervals on the individual means overlap with each other is a valid equivalent. When the intervals don't overlap, the p-value will be $p < 0.03$.
    - For data with, say, $n_1 > 5$ and $n_2 > 5$, the t distribution doesn't add much to the test.

- Pedagogical innovations:
  - The t-test is a special case of “one-way” ANOVA. The equal-variance t-test is mathematically identical to ANOVA. The unequal-variance t-test generally gives a result very similar to ANOVA.
  - The t-test and ANOVA are forms of regression, so it may be more effective to start with regression and then move on to ANOVA and the t-test.
- Streamlining the curriculum:
  - Focus on the confidence interval.
  - Forget about one-tailed tests.
  - For several reasons, there’s never much reason to use an unequal-variance t-test: mathematical complexity, failure to add much power to the test, alternatives (such as rank transforms), philosophical quandries (if you know the variances are unequal, why do you need to look at the means to see if the groups are different).
  - Using regression to set up the t-test

Author Info

Carol Howald is an Associate Professor of mathematics at Howard Community College. She is also a StatPREP Hub Leader.

Contact info:

- Email: chowald@howardcc.edu
- Location: Howard Community College, Columbia, Maryland, USA