How can I do a t-test with survey data?

Authored by

stats writer

July 1, 2024

A t-test is a statistical method used to compare the means of two groups and determine if there is a significant difference between them. When working with survey data, a t-test can be performed by first organizing the data into two groups based on a specific characteristic, such as gender or age. Then, the mean scores for each group can be calculated. Finally, the t-test can be conducted to determine if there is a significant difference between the means of the two groups, providing valuable insights into the relationship between the characteristic and the survey responses. This method is commonly used in social sciences and market research to analyze survey data and draw valid conclusions.

How can I do a t-test with survey data? | Stata FAQ

There is no svy: ttest command in Stata; however, svy: mean is an estimation command and allows for the use of both the test and lincom post-estimation commands. It is also easy to do a t-test using the svy: regress command.

We will show each of these three ways of conducting a t-test with survey data below.

We will illustrate this using the hsb2 dataset pretending that the variable socst is the sampling weight (pweight) and that the sample is stratified on ses. Let's say that we wish to do a t-test for write by gender.

In our dataset, the variable female is coded 1 for

females and 0 for
males.

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

svyset , strata(ses)

pweight: socst
VCE: linearized
Strata 1: ses
SU 1:
FPC 1:

Method 1: Using the test command

First, we use the svy: mean command with the over option to get
the means for each gender. Next, we use the test command to test
the null hypothesis that these two means are equal.

svy: mean write, over(female)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10481
Design df = 197

male: female = male
female: female = female

| Linearized
Over | Mean Std. Err.
---+---
write |
male | 51.65351 1.041066 49.60045 53.70658
female | 55.81467 .721354 54.3921 57.23723

To use the test command, we need to know the labels that Stata has assigned to the values in the output. We can see these labels by using the coeflegend option on the svy: mean command.

svy: mean write, over(female) coeflegend
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10,481
Design df = 197

| Mean Legend
---+---
-
c.write@female |
male | 51.65351 _b
female | 55.81467 _b

Now that we know what the labels are, we can use them in the test command.

test _b = _b

Adjusted Wald test

(1) c.write@0bn.female - c.write@1.female = 0

F(1, 197) = 10.45
Prob > F = 0.0014

Method 2: Using the lincom command

We could also use the lincom command to test the two means. This command should be run after the svy: means command shown above. The lincom command gives us the difference between the means (51.65351 - 55.81467 = -4.161156), the standard error of the difference, as well as the t-value and the p-value. Notice that the p-value is the same as above, and that squaring the t-value yields the F-value shown above ((-3.23)^2 = 10.45).

svy: mean write, over(female)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10481
Design df = 197

male: female = male
female: female = female

--

| Linearized

Over | Mean Std. Err.

----------+----------

write |

male | 51.65351 1.041066 49.60045 53.70658

female | 55.81467 .721354 54.3921 57.23723

To use the licom command, we need to know the labels that Stata has assigned to the values in the output. We can see these labels by using the coeflegend option on the svy: mean command.

svy: mean write, over(female) coeflegend

(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 3 Number of obs = 200

Number of PSUs = 200 Population size = 10,481

Design df = 197

| Mean Legend

----------+----------

-

c.write@female |
male | 51.65351 _b
female | 55.81467 _b
---

lincom _b - _b

(1) c.write@0bn.female - c.write@1.female = 0

Mean | Coef. Std. Err. t P>|t|
---+---
(1) | -4.161156 1.2871 -3.23 0.001 -6.699419 -1.622892
---

* The precise value of the t statistic can be obtained from the list of values
* stored by Stata after running the estimation command svy: mean.

return list

scalars:
r(df) = 197
r(ub) = -1.622892488144128
r(lb) = -6.699418642311276

r(p) = .0014363375306614
r(t) = -3.232969710887891
r(level) = 95
r(se) = 1.287100077434656
r(estimate) = -4.161155565227702

display (-3.232969710887892)^2
10.452093

We can see from the output above that the means are not statistically
equivalent.

Method 3: Using the regress command

The svy: regress command can also be used to compute the t-test.
To do this, simply include the single dichotomous predictor variable. The
coefficient for female is the t-test. As you can see, you get the
same coefficient and p-value that we did when we used the lincom command.
The sign of the coefficient is different because above, the mean of the females

was subtracted from the mean of males. Below, the mean of males was
subtracted from the mean of the females.

svy: regress write female

(running regress on estimation sample)

Survey: Linear regression

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10481
Design df = 197
F(1, 197) = 10.45
Prob > F = 0.0014
R-squared = 0.0519

--
| Linearized
write | Coef. Std. Err. t P>|t|
-------------+--
female | 4.161156 1.2871 3.23 0.001 1.622892 6.699419
_cons | 51.65351 1.041066 49.62 0.000 49.60045 53.70658
--

We can use the test command after the svy: regress if we would like to get the F-ratio.

test female

Adjusted Wald test

(1) female = 0

F(1, 197) = 10.45
Prob > F = 0.0014

Regardless of the method that we use, we obtain an F-ratio of 10.45 or a t-value
of 3.23 with a p-value of 0.0014.

Note: This FAQ was inspired by several responses to a question on the Statalist.