The Stata Journal (2024)
24, Number 3, pp. 446–477 DOI: 10.1177/1536867X241276111

Tests and confidence bands for multiple one-sided comparisons

David M. Drukker
Clemson University
Clemson, SC
ddrukke@clemson.edu

Kevin S. S. Henning
Sam Houston State University
Huntsville, TX
henning@shsu.edu

Christian Raschke
Sam Houston State University
Huntsville, TX
raschke@shsu.edu

Abstract. One-sided inference should be used in some applications, but Stata has limited support for one-sided tests and confidence intervals for a single comparison and almost no support for one-sided tests and confidence bands for multiple comparisons. In this article, we provide an introduction to one-sided tests and confidence intervals for a single hypothesis and to one-sided tests and confidence intervals for multiple comparisons. We also discuss extensions of the `sotable` command introduced in Drukker (2023, *Stata Journal* 23: 518–544) to cover one-sided tests and confidence bands for a single comparison and for multiple comparisons. We also provide examples of how to use `sotable` to perform multiple tests against values other than zero and how to perform multiple tests after commands like `margins` and `nlcom` that support the `post` option.

Keywords: st0718_1, sotable, one-sided testing, multiple comparisons, confidence bands, simultaneous inference, multiple testing, max-t, simultaneous tests

1 Introduction

Standard Stata output tables for frequentist estimators provide almost no support for one-sided inference. They also fail to correct for the multiple simultaneous inferences implicitly performed in many empirical studies. In fact, the default p-values and confidence intervals produced in Stata do not account for the fact that using two or more of the reported p-values or confidence intervals will incur an error rate higher than the specified level. This well-known multiple-testing problem was addressed by Drukker (2023), who presented the `sotable` command as a solution when the hypotheses are two sided.[1] We now tackle the issue of one-sided tests. We argue that practical one-sided testing situations arise naturally in empirical work, and therefore, researchers should use one-sided tests in applications that require them. We discuss extensions to

1. Because of the direct relationship between hypothesis tests and confidence intervals, when we refer to "multiple testing", we are referring to the general problem of the increased error rate that occurs when making multiple statistical inferences simultaneously. These inferences could be made using a test statistic or a confidence interval.

the `sotable` command that provide p-values and confidence intervals for one or more one-sided inferences.

Section 2 argues that one-sided tests should be used in some applications and addresses several myths and misconceptions about one-sided tests. Section 3 provides a real-data application that illustrates using `sotable` for one-sided tests about one parameter. Section 4 discusses multiple testing in a real-data application and provides a simulated-data application that illustrates the importance of correcting for multiple testing. Section 5 presents the methods and formulas that are implemented in the `sotable` command. Section 6 presents the syntax of the `sotable` command. Section 7 presents examples that illustrate how to use `sotable` with parameters other than a simple output table and how to test against values other than zero. Section 8 presents some simulations that illustrate that the `sotable` command performs as expected. Section 9 provides our conclusions and thoughts for future research.

2 When one-sided tests are useful

One-sided tests are useful when we are interested in a joint hypothesis about the size and the direction of the effect. Consider the case in which a treatment is useful only if it has a strictly positive effect, such as what might occur in a pharmaceutical trial when testing whether a drug performs better than an existing treatment. In this case, the null hypothesis of no useful effect includes zero and all negative values. The alternative hypothesis is that the effect is strictly greater than zero. For concreteness, we discuss this upper-tailed case in this section.[2]

Clarity is a clear advantage of using `sotable` to perform one or more one-sided tests. It is possible to manipulate the p-value and the confidence interval reported by an estimation command to perform a single-comparison one-sided inference. It is also possible to manipulate the p-values and the confidence intervals reported by `sotable` to perform multiple-comparison one-sided inferences. But these manipulations require that the user know the conditions and the formulas required. Some of these manipulations are trivial, and some are more involved. The prevalence of misconceptions about one-sided tests discussed in this section indicates that not all researchers are comfortable making the correct adjustments.

Arguments against using one-sided tests even in applications that require them usually stem from one of three common misconceptions.

The first misconception is that by employing a one-sided test, the researcher "assumes away" the possibility of a test statistic falling into the other tail. This is not true. The fact that a one-sided test will not reject an alternative in the other tail respects what the researcher defined to scientifically interesting. Consider an example in which a strictly positive treatment effect is the alternative of interest. If we use an upper-tailed test and the effect is truly negative, we will never reject the null hypothesis (with the

2. Discussing the general case of one-sided tests leads to including too many qualifying clauses. Using a lower-tailed case has all the same points, with the signs flipped.

p-value approaching one as $n \to \infty$). This result is a benefit of the one-tailed test, not an assumption limiting its usefulness. One-sided tests empower researchers to perform inference about the kinds of outcomes they decide to be scientifically interesting.

The second misconception is that a one-sided test makes it easier to reject the null hypothesis. This misconception arises because, in practice, the p-value of a one-sided test is one half the p-value of the corresponding two-sided test. This argument is not exactly true. The p-value is divided by two only when the estimated coefficient has the correct sign, which occurs only in one-half of the repeated samples when the null hypothesis is true with equality.[3] In fact, the additional requirement that the coefficient have the correct sign is why the p-value must be divided by two.

A related argument against one-sided testing is that it provides a temptation for researchers to switch hypotheses post hoc from two-sided to one-sided after finding that an insignificant two-sided test could become a significant one-sided test. This objection ignores the fact that making the data fit the hypotheses is not a unique pitfall of one-sided testing. All significance levels are ad hoc and the result of conventions in the various disciplines, and this problem is easily disposed of by using a significance level that is divided by two. Instead of using the conventional significance levels of 0.1, 0.05, and 0.01, use the significance levels of 0.05, 0.025, and 0.005 for one-sided tests. The convention that we propose here in fact mirrors the existing convention in economics of using a significance level of $\alpha = 0.05$ in two-sided tests. We follow the convention of rejecting only when the estimate has the desired sign when using a two-sided test yields the same result.

To illustrate this point, let's consider figure 1. Each panel shows the critical value, the rejection probability (in black) and the nonrejection probability (in gray) for a test statistic that has a standard normal distribution. The values under a black area lead to rejection, and the values under a gray area lead to failing to reject the null.[4] Panel A shows these regions for a two-sided test with a significance level of $\alpha = 0.05$. Panel B shows these regions for an upper-tailed, one-sided test with a significance level of $\alpha = 0.05$. Panel C shows these regions for an upper-tailed, one-sided test with a significance level of $\alpha = 0.025$.

3. With one-sided tests where the null hypothesis is not a single number but a set of points in the parameter space, the level of significance is more precisely the maximum probability of making a type I error. In the tests we consider, this maximum is obtained when the null hypothesis is true with equality.
4. We note that it is necessary to use the phrase "fail to reject" rather than "accept" because not rejecting a null hypothesis does not mean that the null is true.

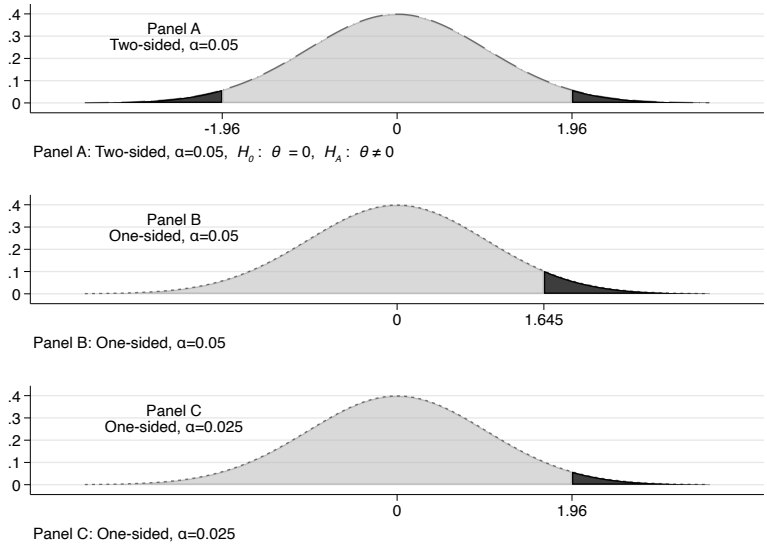

Figure 1. Rejection regions

Two-sided tests like the one in panel A can reject when the estimated effect has the wrong sign. This is known as a type III error, or directional error. Such an error will occur with, at most, a probability of $\alpha/2$. If the alternative hypothesis is one sided, it makes no sense to reject the null hypothesis when the estimated effect has the wrong sign. One solution to this problem is to use a two-sided test and to reject the null hypothesis only when the p-value is below the significance level and the estimated coefficient is of the correct sign. This approach yields the same inference as the one-sided test with one half the significance level. To see this, note that the critical values in panels A and C are the same.

Using an upper-tailed test means that we will fail to reject the null of no effect when the estimator is randomly too low in repeated samples. This makes sense. We cannot distinguish between the case in which we observe a negative estimate because the true effect is negative and the case in which we observe a negative estimate by chance. We are not assuming that the random variation in the estimator is inherently positive; rather, negative random values of the test statistic simply do not provide evidence of a positive effect. Only allowing for one-sided variation to cause a rejection may also support the practice of dividing the conventional significance by 2.

Finally, there is a myth that one-sided tests have twice the type I error as two-sided tests (see, for example, Brown et al. [2019, 1533]). This is false. The derivation of the critical value in a one-sided test makes this clear; see Bickel and Doksum (1977, ex. 5.2.2) and Wooldridge (2020, 735–737) for exemplary derivations. We also illustrate this point by simulation in appendix A.1. While we believe that it makes sense to

divide the conventional significance levels by 2 to account for the fact that only one-sided random variation leads to a rejection under the null hypothesis, the probability of a type I error is equal to the significance level that we specify, when the null hypothesis is true with equality.

3 Real data application

We use an application from Elfenbein and McManus (2010) as a guiding example to illustrate the statistical interpretation and to introduce `sotable`. Data and do-files are available at https://www.openicpsr.org/openicpsr/project/114735/version/V1/view.[5]

Elfenbein and McManus (2010) investigate whether consumers are willing to spend more for products that generate charitable donations. The outcome of interest in this regression is `log_ps`, which contains the log of the price paid in an eBay auction, including shipping fees. The covariates of interest are `donpct_10`, `donpct_100`, and `donpct_middle`, which are indicators for a small donation level, a large donation level, and a middle donation level, respectively. The control variables and their definitions are given in table 1. The observations have a panel-data structure, and `gw` is the ID variable for the auctions in which the observations are clustered.

Table 1. Control variables

`log_sellerrating`	natural log of seller rating
`dseller100`	seller percentage = 100
`dseller995`	seller percentage in $[99.5, 100)$
`dseller990`	seller percentage in $[99, 99.5)$
`dseller98`	seller percentage in $[98, 99)$
`pwrsell`	`1` if power seller, `0` otherwise
`d_len3`	$1 <$ length $<= 3$
`d_len5`	$3 <$ length $<= 5$
`d_len7`	$5 <$ length $<= 7$
`d_len10`	$7 <$ length $<= 10$
`bin`	`1` if buy-it-now auction, `0` otherwise

`sgw`
`chicetc`
`sfgw`
`blackbier`
`bookusa`
	dummy variables for large-volume sellers

NOTE: "Length" is the auction length in days, and "seller percentage" is the positive rating percentage for the seller.

5. This section repeats some of the descriptions from Drukker (2023), because we are reconsidering the same example.

Results from column (6) in table 3 of Elfenbein and McManus (2010) are reproduced here:

```
. use prepped_aucdata_aejpol-2008-0041
. // outcome
. local outcome log_ps
. // covariates of interest
. local interest donpct_10 donpct_100 donpct_middle
. // control variables
. local controls log_sellerrating dseller100 dseller995 dseller990 dseller98
. local controls `controls' pwrsell d_len3 d_len5 d_len7 d_len10 bin
. local controls `controls' sgw chicetc sfgw blackbier bookusa
. xtreg `outcome' `interest' `controls', i(gw) fe cluster(gw)
Fixed-effects (within) regression               Number of obs     =      2,433
Group variable: gw                              Number of groups  =        723

R-squared:                                      Obs per group:
     Within  = 0.0846                                         min =          2
     Between = 0.0010                                         avg =        3.4
     Overall = 0.0003                                         max =          6

                                                F(19, 722)        =       7.67
corr(u_i, Xb) = -0.0685                         Prob > F          =     0.0000

                                    (Std. err. adjusted for 723 clusters in gw)
```

log_ps	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
donpct_10	.0385109	.015243	2.53	0.012	.008585	.0684367
donpct_100	.0824094	.01448	5.69	0.000	.0539815	.1108373
donpct_middle	.027038	.0236634	1.14	0.254	-.0194194	.0734953
log_sellerrat~g	.0050287	.0029762	1.69	0.092	-.0008144	.0108717
dseller100	.0104112	.0186338	0.56	0.577	-.0261717	.0469941
dseller995	.0035016	.0203058	0.17	0.863	-.0363638	.0433671
dseller990	-.0021439	.0220303	-0.10	0.923	-.0453951	.0411072
dseller98	-.0044067	.0231118	-0.19	0.849	-.049781	.0409676
pwrsell	.0160304	.0113956	1.41	0.160	-.0063421	.0384029
d_len3	.0011487	.0152913	0.08	0.940	-.028872	.0311694
d_len5	.0094176	.016738	0.56	0.574	-.0234434	.0422787
d_len7	.0093424	.0145587	0.64	0.521	-.0192401	.0379248
d_len10	.0594335	.0267447	2.22	0.027	.0069268	.1119403
bin	.042998	.0106808	4.03	0.000	.0220289	.0639672
sgw	-.0069474	.026346	-0.26	0.792	-.0586714	.0447766
chicetc	.2242664	.0522418	4.29	0.000	.1217024	.3268303
sfgw	-.1777547	.044254	-4.02	0.000	-.2646365	-.0908728
blackbier	-.0345159	.0313844	-1.10	0.272	-.0961314	.0270996
bookusa	.0068463	.0948352	0.07	0.942	-.1793393	.1930319
_cons	3.961137	.0260659	151.97	0.000	3.909963	4.012312
sigma_u	.96227252					
sigma_e	.15799629					
rho	.97374907	(fraction of variance due to u_i)				

3.1 Single one-sided test

The research question of whether consumers are willing to spend *more* for products that generate charitable donations implies a one-sided hypothesis. For now, consider only one of the variables of interest, `donpct_10`, indicating a lower donation level of the seller. Suppose we want to conduct only a single test of the coefficient on the variable `donpct_10`. The `xtreg` output reports a two-sided test and two-sided confidence interval for the effect of `donpct_10`. This is of limited use because the alternative of interest is that the effect is strictly greater than zero. Therefore, we use `sotable` to compute an upper-tailed test and an upper-tailed confidence interval for the effect.

```
. sotable, pnames(donpct_10) alternative(upper)
Single-comparison results
       p-value = 0.006
Critical value = 1.963
```

log_ps	Coef.	Std. Err.	t	P>|t|	[97.5% Conf.	Band]
donpct_10	.0385109	.015243	2.526	0.006	.008585	.

We used the `pnames()` option to specify the parameter of interest, and we used the `alternative(upper)` option to specify that we want upper-tailed results. As expected, the p-value of the test of $H_0 : \beta_{\texttt{donpct\_10}} \leq 0$ versus the alternative hypothesis $H_a : \beta_{\texttt{donpct\_10}} > 0$ is one half the p-value for the two-sided test. Following the convention of using one half the two-sided significance level that we set to 0.05, we see that the reported p-value is less than 0.025, so we reject the null hypothesis of no effect.

Note that the default confidence level is 97.5% instead of 95%. This default deflects any allegation of using a one-sided confidence interval to fish for significant results.[6] The estimated endpoints of a 97.5% confidence level are 0.008585 and positive infinity. The confidence interval is the do-not-reject region; we cannot reject the null hypothesis $H_0 : \beta_{\texttt{donpct\_10}} \leq \beta_0$ for the alternative hypothesis $H_a : \beta_{\texttt{donpct\_10}} > \beta_0$ for any $\beta_0 \in [0.008585, \infty)$. The complement of the confidence interval is the rejection region. We can reject the null hypothesis for any $\beta_0 \leq 0.008585$. In particular, we reject the null hypothesis of no effect, because 0 is outside the confidence interval.

We can obtain this confidence interval by inverting the test statistic. Let $\hat{\beta}$ be an estimator that produces a t statistic with a t distribution; let β_0 be the value of the coefficient when the null hypothesis is true with equality; let se be the standard error of $\hat{\beta}$; let n be the sample size; and let $t(1-\alpha)$ be the inverse function of the t distribution with $n-2$ degrees of freedom at $1-\alpha$.

6. As discussed in section 6, `sotable` uses the default level of `c(level)`+(100−`c(level)`)/2 for one-sided confidence intervals, where `c(level)` is the default level specified by `set`.

$$\Pr\left(\frac{\hat{\beta}-\beta_0}{\text{se}} > t(1-\alpha)\right) = \alpha$$
$$\Pr\left(\frac{\hat{\beta}-\beta_0}{\text{se}} \leq t(1-\alpha)\right) = 1-\alpha$$
$$\Pr\left(\hat{\beta} - \text{se}\ t(1-\alpha) \leq \beta_0\right) = 1-\alpha$$

Looking at the p-value of the test statistic provides some intuition as to why the confidence interval is open on the right. The p-value of a test of $H_0\colon \beta_{\texttt{donpct\_10}} \leq \beta_0$ for the alternative hypothesis $H_0\colon \beta_{\texttt{donpct\_10}} > \beta_0$ is $1-\tau\{n-2, (\hat{\beta}-\beta_0)/\text{se}\}$,where $\tau(\nu, t_0)$ is the cumulative distribution function of a t distribution with ν degrees of freedom at the value t_0. The p-value increases as we increase β_0, holding everything else constant, meaning that a rejection becomes less likely as we increase β_0.

The width of a two-sided confidence interval is the width of the fail-to-reject region. Because one-sided confidence intervals are open on one side, the width of the interval is infinite, and this width is not informative about the width of the fail-to-reject region. But when zero is the no-effect value, the distance between an endpoint of a two-sided confidence interval and zero provides a measure of the scientific significance. When the significance level for the one-sided confidence interval is one half the significance level of the two-sided confidence interval, the finite endpoint of the one-sided confidence interval equals one of the endpoints of the two-sided confidence interval. This equality implies that the one-sided confidence interval and the two-sided confidence interval provide the same measure of scientific significance, but note that the one-sided confidence interval has the benefit of not indicating rejection for very large effects.

In a two-sided alternative there is a "no-effect" value. In a one-sided alternative, there is a bound for the no-effect region. In the previous paragraph, we set this bound "cutoff" value to zero, which is common in applied work. The same logic applies when the cutoff no-effect value is not zero. For `sotable` examples with nonzero values, see section 7.

4 Multiple one-sided tests

The Elfenbein and McManus (2010) application actually involves several coefficients, not just $\beta_{\texttt{donpct\_10}}$. The coefficients of interest are $\beta_{\texttt{donpct\_10}}$, $\beta_{\texttt{donpct\_100}}$, and $\beta_{\texttt{donpct\_middle}}$. Because several coefficients are tested to investigate the research question, we need the concepts and the tools from the literature on multiple testing. Drukker (2023) provides an introduction to these tools and concepts and an introduction to using the `sotable` command for multiple two-sided tests. In this section, we provide a brief review of these tools and concepts before we discuss using the `sotable` command for multiple one-sided tests.

In this application, we would like to test each of the following individual hypotheses:

$$\begin{aligned} \widetilde{H}_{0,1} &: \beta_{\texttt{donpct\_10}} \leq 0 \quad \text{versus} \quad \widetilde{H}_{a,1} : \beta_{\texttt{donpct\_10}} > 0 \\ \widetilde{H}_{0,2} &: \beta_{\texttt{donpct\_100}} \leq 0 \quad \text{versus} \quad \widetilde{H}_{a,2} : \beta_{\texttt{donpct\_100}} > 0 \\ \widetilde{H}_{0,3} &: \beta_{\texttt{donpct\_middle}} \leq 0 \quad \text{versus} \quad \widetilde{H}_{a,3} : \beta_{\texttt{donpct\_middle}} > 0 \end{aligned} \tag{1}$$

We are also interested in testing the overall null hypothesis that

$$\begin{aligned} &H_0 : \widetilde{H}_{0,j} \text{ is true for all } j \in \{1,2,3\} \\ &\text{versus} \\ &H_a : \widetilde{H}_{0,j} \text{ is not true, for at least one } j \in \{1,2,3\} \end{aligned} \tag{2}$$

The tests of the individual hypotheses tell us for which coefficients we can reject the null hypothesis of no effect, and the test of the overall hypothesis tells us whether we reject the null hypothesis that none of the coefficients imply an effect.

Formally, we want a testing procedure that will tell us whether we can reject the overall null hypothesis in (2) and which, if any, of the individual null hypotheses in (1) we can reject. For the moment, let's concentrate on the overall null hypothesis. Note that we reject the overall null hypothesis if any of the individual null hypotheses are rejected. So we want a procedure that performs each of the three individual tests and rejects the overall null hypothesis if any of the three individual hypotheses are rejected.

The key to correctly doing multiple tests is to use adjusted p-values that account for the multiple tests performed. Adjusted p-values are calculated to control the probability of falsely rejecting one or more true individual null hypotheses. This probability is called the familywise error rate (FWER). When there is only one test, the significance level controls the probability of falsely rejecting a true null hypothesis. The FWER plays the role of the significance level when there is more than one test.[7]

We use three standard terms when discussing rejection rates. First, a multiple-testing procedure that rejects at least one true null hypothesis more frequently than the specified FWER is said to *overreject*. Note that we cannot use a testing procedure that overrejects. Second, a multiple-testing procedure that rejects at least one true null hypothesis less frequently than the specified FWER is said to *underreject*. While we could use a testing procedure that underrejects, we avoid them because they sacrifice power.[8] Third, a multiple-testing procedure that rejects at least one true null hypothesis at the specified FWER is said to *reject at the nominal rate*. We prefer to use a testing procedure that rejects at the nominal rate.

We can now use these terms to characterize some multiple-testing procedures. A testing procedure that uses unadjusted p-values overrejects and thus should not be

7. The FWER depends on which subset of the family of hypotheses is true, which is unknown. Therefore, we need procedures that control the maximal FWER over all possible configurations of true and false null hypotheses.
8. The power of a test is the probability that it will correctly reject a false null hypothesis. The higher the power of a test, the better.

used. There are different methods for adjusting the p-values for multiple tests. As discussed in Drukker (2023), `sotable` uses the max-t method. Using the max-t–adjusted p-values rejects at the nominal rate. In contrast, the Bonferroni method, discussed in many introductory statistics classes, underrejects.[9] We discuss these properties in the examples in section 4.2 and in the simulations in section 8.

4.1 Multiple one-sided tests in an application

As discussed in section 3, the application uses `xtreg` to compute the results. Unfortunately, `xtreg` reports only unadjusted p-values for two-sided tests. We can use the unadjusted p-values reported by `xtreg` to perform one two-sided test. We should not use the `xtreg` results to perform multiple tests using unadjusted p-values because these procedures will overreject at least one of the true null hypotheses. However, using the max-t–adjusted p-values computed by `sotable` will reject at least one of the true null hypotheses at the nominal rate. We give the formulas used by `sotable` in section 5.

Now let's use `sotable` to compute the max-t–adjusted p-values for each of the individual null hypotheses tested.

```
. sotable, pnames(donpct_10 donpct_100 donpct_middle) alternative(upper)
Max-t results
       p-value = 0.000
Critical value = 2.390
```

log_ps	Coef.	Std. Err.	t	P>|t|	[97.5% Conf.	Band]
donpct_10	.0385109	.015243	2.526	0.017	.0020877	.
donpct_100	.0824094	.01448	5.691	0.000	.0478094	.
donpct_middle	.027038	.0236634	1.143	0.330	-.0295059	.

We reject two of the three individual hypotheses and therefore reject the overall hypothesis. Note that the adjusted p-values are greater than the unadjusted p-values as a result of accounting for multiple testing.

We could also use the joint confidence intervals reported by `sotable` to reach the same conclusion. Each of the confidence intervals reported by `sotable` has been adjusted for the multiple tests so that the group of them form a confidence band. The confidence band is the joint fail-to-reject region. Any value outside the reported confidence band can be rejected at the 0.025 level. Zero is outside the fail-to-reject region for two of the coefficients, and therefore at least one of the null hypotheses is rejected.

In this example, using the unadjusted p-values happened to lead us to the correct conclusion. But the increase in p-values as a result of the adjustment illustrates the properties discussed above. Using the unadjusted p-values will overreject, but using the max-t–adjusted p-values will reject at the nominal rate.

9. The major reason for this is that the Bonferroni method does not incorporate the dependence structure in the test statistics, whereas the max-t method and some other methods do.

4.2 Simulated example

We illustrate the gravity of the issue by using data from a fictional study that mimics a randomized treatment in an online teaching environment in the file `olcdata3.dta`. Suppose there are three new versions for an online class. The treatment consists of randomly assigning students into one of the three new versions of the class (denoted version 1, version 2, or version 3) or into the original version of the class (denoted version 0).

Let μ_j be the mean class score for version $j \in \{0, 1, 2, 3\}$. We have three individual hypotheses of no effect versus some effect, and getting an answer from the data is going to require more than one hypothesis test,

$$\begin{aligned} H_{0,j}&: \mu_j - \mu_0 \leq 0 \\ H_{a,j}&: \mu_j - \mu_0 > 0 \end{aligned} \quad j \in \{1, 2, 3\}$$

We begin by reading in the simulated data and using `regress` with `vce(robust)` to estimate the difference in means.[10]

```
. use olcdata3, clear
. regress score i.version, vce(robust)
Linear regression                               Number of obs     =        344
                                                F(3, 340)         =       5.21
                                                Prob > F          =     0.0016
                                                R-squared         =     0.0453
                                                Root MSE          =     2.3833
```

score	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
version						
1	-.7168975	.3309735	-2.17	0.031	-1.367911	-.0658841
2	.3396586	.377246	0.90	0.369	-.4023714	1.081689
3	.6650583	.3362483	1.98	0.049	.0036693	1.326447
_cons	70.32257	.2127093	330.60	0.000	69.90418	70.74097

For the estimated coefficient that is negative, the unadjusted one-sided p-value is greater than 0.5. For the estimated coefficients that are positive, the unadjusted one-sided p-values are the reported p-values divided by 2. Below, we compute the unadjusted one-sided p-values using the textbook formulas for one-sided tests and find that they are 0.984, 0.184, and 0.024, respectively.

```
. local df = e(df_r)
. display "unadjusted one-sided p-value for j=1 is "
> 1 - t(`df',_b[1.version]/_se[1.version])
unadjusted one-sided p-value for j=1 is .98449816
```

10. The classical approach to multiple comparisons of several treatment means with a control is the Dunnett procedure, which is given as an option in the `pwcompare` command. However, the approach discussed here using `regress` with robust standard errors is more flexible.

```
. display "unadjusted one-sided p-value for j=2 is "
> 1 - t(`df',_b[2.version]/_se[2.version])
unadjusted one-sided p-value for j=2 is .18428203
. display "unadjusted one-sided p-value for j=3 is "
> 1 - t(`df',_b[3.version]/_se[3.version])
unadjusted one-sided p-value for j=3 is .02437476
```

If we used these unadjusted p-values, we would reject at least one of the individual tests, because the smallest unadjusted p-value is less than 0.025. Because we rejected at least one of the individual tests, we would reject the overall null hypothesis.

We now use `sotable` to compute the max-t–adjusted p-values.

```
. sotable, pnames(1.version 2.version 3.version) alternative(upper)
Max-t results
       p-value = 0.065
Critical value = 2.377
```

score	Coef.	Std. Err.	t	P>|t|	[97.5% Conf.	Band]
1bn.version	-.7168975	.3309735	-2.166	1.000	-1.503558	.
2.version	.3396586	.377246	0.900	0.388	-.5569823	.
3.version	.6650583	.3362483	1.978	0.065	-.1341391	.

We see that the smallest-adjusted one-sided p-value is 0.065, so we do not reject any of the individual null hypotheses. Because we do not reject any of the individual null hypotheses, we do not reject the overall null hypothesis.

We could also use the joint confidence intervals reported by `sotable` to reach the same conclusion. Each of the confidence intervals reported by `sotable` has been adjusted for the multiple tests so that the group of them form a confidence band. The confidence band is the joint fail-to-reject region. Any value outside the reported confidence band can be rejected at the 0.025 level. Zero is inside the region for each coefficient, so none of the individual hypotheses are rejected.

In this example, using the unadjusted p-values would cause us to reject an individual hypothesis and therefore the overall null hypothesis. In contrast, using the max-t–adjusted p-values would cause us not to reject any individual null hypothesis or the overall null hypothesis. This example also illustrates the rejection properties discussed above. First, using the unadjusted p-values will overreject at least one of the true individual null hypotheses and therefore the overall null hypothesis. Second, using the max-t–adjusted p-values will reject null hypotheses at the nominal rate.

The problem of using unadjusted p-values is not limited to this specific example. Even with only three tests, using the unadjusted p-values instead of the adjusted p-values can cause the rejection rate to be more than twice the specified error rate. See section 8 for additional simulations illustrating this issue.

5 Details and formulas for the general case

In the previous sections, we have argued that one-sided tests should be used to test joint hypotheses of size and direction. We have also shown that we need to use testing procedures based on adjusted p-values when we test more than one parameter. This section provides the details on how the `sotable` postestimation command computes max-t–adjusted p-values and joint confidence bands for one-sided tests. Hothorn, Bretz, and Westfall (2008) derived the formulas for two-sided tests, and Drukker (2023) gave a simulation algorithm to approximate the critical value and the adjusted p-values in Stata. The one-sided formulas given in this section use the standard technique of inverting the test statistic; see Casella and Berger (2002, ex. 9.2.4) and Wooldridge (2020, 735–737) for example derivations.

Let's begin by supposing that there are q parameters of interest $\theta_1, \theta_2, \ldots, \theta_q$ and that estimators for the these parameters are a subset of the parameters estimated by a frequentist estimation command and stored in `e(b)`.

Sections 5.1 and 5.2 provide the individual hypotheses when the alternative is upper or lower, respectively. Regardless of the alternative, for each of the individual hypotheses, we have a t statistic

$$t_j = \hat{\theta}_j / \hat{s}_j \tag{3}$$

where $\hat{\theta}_j$ is the estimator of θ_j and $\hat{s}_j$ is a standard error of $\hat{\theta}_j$. We require that the individual t statistics have either an asymptotic normal distribution or a finite-sample t distribution. More formally, let $\mathbf{w} = (w_1, w_2, \ldots, w_q)$ be a multivariate random variable. For estimators whose t statistics have a multivariate normal distribution, $\mathbf{w} \sim N(\mathbf{0}, \widehat{\mathbf{C}})$, where $\widehat{\mathbf{C}} = \texttt{corr}(\widehat{\mathbf{V}})$ and $\widehat{\mathbf{V}}$ is the estimated variance–covariance of the estimator extracted from the `e(V)` stored by the estimation command. For estimators whose t statistics have a multivariate t distribution, $\mathbf{w}$ has a multivariate t distribution with scale matrix $\widehat{\mathbf{C}}$.

Recall from above that the probability of falsely rejecting at least one individual null hypothesis in a family of tests is known as the FWER. The FWER is the analog of the significance level in a single hypothesis test. We denote the FWER by α_F.

5.1 Details for alternative(upper)

In this section, we provide details and methods for the case of an upper-tailed hypothesis that is covered when option `alternative(upper)` is specified.

For the `alternative(upper)` case, the q individual hypotheses are

$$H_{0,j}\colon \theta_j \leq 0 \quad \text{versus} \quad H_{a,j}\colon \theta_j > 0 \tag{4}$$

for each $j \in \{1, 2, \ldots, q\}$. The overall null hypothesis and alternative hypothesis are

$$\begin{aligned} &H_0\colon \theta_j \leq 0 \text{ for all } j \in \{1, 2, \ldots, q\} \\ &\quad \text{versus} \\ &H_a\colon \theta_j > 0 \text{ for at least one } j \in \{1, 2, \ldots, q\} \end{aligned} \tag{5}$$

The critical value for the overall test in (5) is c_{upper}, and it solves

$$\Pr\left(\max_{j\in\{1,2,\ldots,q\}} \{w_j\} \geq c_{\text{upper}}\right) = \alpha_F \tag{6}$$

`sotable` uses algorithm 1 given in appendix A.2 to approximate c_{upper}. The adjusted p-value (p_j) for the jth test in the family is

$$p_j = 1 - \Pr\begin{pmatrix} w_1 & \leq & t_j \\ w_2 & \leq & t_j \\ \vdots & & \\ w_q & \leq & t_j \end{pmatrix} \tag{7}$$

where t_j is the t statistic for the jth parameter defined in (3). `sotable` uses algorithm 1 given in appendix A.2 to approximate p_j.

We reject the jth individual hypothesis in (4) if $p_j \leq \alpha_F$.

The p-value for the overall null hypothesis (p) is given by

$$p = \min(p_1, p_2, \ldots, p_q)$$

We reject the overall null hypothesis in (5) if $p \leq \alpha_F$.

Each of the upper-tailed confidence intervals in the confidence band reported by `sotable` is

$$(\widehat{\theta}_j - c_{\text{upper}}\ \widehat{s}_j, .)$$

where c_{upper} is computed from (6).

5.2 Details for alternative(lower)

In this section, we provide details and methods for the case of a lower-tailed hypothesis that is covered when option `alternative(lower)` is specified.

For the `alternative(lower)` case, the q individual hypotheses are

$$H_{0,j}\colon\ \theta_j \geq 0 \quad \text{versus} \quad H_{a,j}\colon \theta_j < 0 \tag{8}$$

for each $j \in \{1, 2, \ldots, q\}$. The overall null hypothesis and alternative hypothesis are

$$\begin{aligned} &H_0\colon \theta_j \geq 0 \text{ for all } j \in \{1, 2, \ldots, q\} \\ &\quad \text{versus} \\ &H_a\colon \theta_j < 0 \text{ for at least one } j \in \{1, 2, \ldots, q\} \end{aligned} \tag{9}$$

The critical value for the overall test in (9) is c_{lower}, and it solves

$$\Pr\left(\max_{j\in\{1,2,\ldots,q\}} \{w_j\} \leq c_{\text{lower}}\right) = \alpha_F \tag{10}$$

`sotable` uses algorithm 2 given in appendix A.2 to approximate c_{lower}. The adjusted p-value (p_j) for the jth test in the family is

$$p_j = \Pr\begin{pmatrix} w_1 & \leq & t_j \\ w_2 & \leq & t_j \\ \vdots & & \\ w_q & \leq & t_j \end{pmatrix} \tag{11}$$

where t_j is the t statistic for the jth parameter defined in (3). `sotable` uses algorithm 2 given in appendix A.2 to approximate p_j.

We reject the jth individual hypothesis in (8) if $p_j \leq \alpha_F$. The p-value for the overall null hypothesis (p) is given by

$$p = \min(p_1, p_2, \ldots, p_q)$$

We reject the overall null hypothesis in (9) if $p \leq \alpha_F$. Note that once adjusted p-values are computed, this rejection rule is the same as for upper tailed tests.

Each of the lower-tailed confidence intervals in the confidence band reported by `sotable` is

$$(., \hat{\theta}_j - c_{\text{lower}}\ \hat{s}_j) \tag{12}$$

where c_{lower} is computed from (10).

Please note that the expression for the lower-tailed confidence intervals in (12) does not contradict the standard way of writing the single-comparison lower-tailed confidence interval. When the t statistics have a standard normal distribution, the usual expression for a single-comparison lower tail for a parameter θ with standard error $\hat{s}$ confidence interval is

$$\hat{\theta} + z_{1-\alpha}\hat{s}$$

where $z_{1-\alpha}$ is the inverse of the standard normal distribution at quantile $1-\alpha$. Because the normal distribution is symmetric, $z_{1-\alpha} = -z_\alpha$. So the confidence interval could also be written as

$$\hat{\theta} - z_\alpha \hat{s}$$

which is the form that we use.

6 Syntax for one-sided tests

This article extends the `sotable` postestimation command to handle upper-tailed and lower-tailed tests. This section inevitably repeats some of the syntax discussion in Drukker (2023), which introduced the `sotable` command.

Note that `sotable` works only after frequentist estimation commands and that it uses the stored `e(b)` and the `e(V)`.

6.1 Syntax

```
sotable [ , [pnames(pnames)|pelements(numlist)]
    alternative(two|upper|lower) normal draws(#) level(#)]
```

6.2 Options

`pnames(`*pnames*`)` or `pelements(`*numlist*`)` specifies which parameters will be in the output table. Only one of `pnames(`*pnames*`)` or `pelements(`*numlist*`)` may be specified.

`pnames(`*pnames*`)` specifies a list of parameter names to include.

`pelements(`*numlist*`)` specifies a *numlist* of which elements in the parameter vector to include.

The `pnames()` option deserves further explanation. Each of the parameters in `e(b)` has a complete name of the form *eqn*`:`*pname*, where *eqn* is the equation name and *pname* is the parameter name. In the simplest case, `pnames()` contains a subset of the complete names that are displayed by the `coeflegend` option or by listing the matrix `e(b)`.

`pnames()` supports two standard abbreviations. To help explain these abbreviations, let's consider an example that uses `sureg` to estimate the parameters,

```
. sureg (y1 x1 x2) (y2 x1 x3)
```

The list of complete parameter names would be `y1:x1`, `y1:x2`, `y1:_cons`, `y2:x1`, `y2:x3`, and `y2:_cons`.

The first abbreviation indicates that specifying an equation name followed by a colon includes all the parameters in that equation. For example,

```
. sotable, pnames(y1:)
```

after the above `sureg` would produce results for `y1:x1`, `y1:x2`, and `y1:_cons`.

The second abbreviation indicates that the equation name can be omitted. When the equation name is omitted, the first equation is assumed. For example,

```
. sotable, pnames(x1 x2)
```

after the above `sureg` would produce results for `y1:x1` and `y1:x2`.

`alternative(two | upper | lower)` specifies the alternative hypotheses of the tests.

`alternative(two)`, the default, specifies that the tests have a two-sided alternative.

`alternative(upper)` specifies that the tests have an upper-tailed alternative. Note that the default confidence level is `c(level)` + (100 − `c(level)`)/2 when the option `alternative(upper)` is specified.

`alternative(lower)` specifies that the tests have a lower-tailed alternative. Note that the default confidence level is `c(level)` + (100 − `c(level)`)/2 when the option `alternative(lower)` is specified.

Note that you can use `nlcom` with the `post` option to handle tests with a mixture of upper-tail and lower-tail alternatives. See section 7 for an example.

`normal` specifies to use a multivariate normal distribution to calculate the adjusted p-values, the overall critical value, and the overall p-value.

By default, `sotable` uses the distribution used by the command that produced the estimates and the variance–covariance of the estimator.

`normal` specifies that `sotable` use a multivariate normal distribution instead of a multivariate t distribution after estimators that use a multivariate t distribution.

`draws(#)` specifies the number of Monte Carlo draws to use in estimating the max-t critical value. The default is `draws(1000000)`. More draws will reduce the variance of the estimated critical value.

`level(#)` specifies the level ℓ for the confidence band. The FWER is $1 - \ell/100$. The default is `level(95)`. The default FWER is $0.05 = 1 - 95/100$.

6.3 Stored results

`sotable` stores the following results in `r()`:

Scalars

`r(df_r)`	degrees of freedom for t distribution (when t statistics have a t distribution)
`r(c)`	critical value
`r(draws)`	number of draws used in simulation approximation
`r(p)`	p-value of overall test

Macros

`r(alternative)`	`two`, `upper`, or `lower`
`r(level)`	level
`r(dist)`	`t` or `z`
`r(nmethod)`	`simulation`
`r(method)`	`maxt` or `scomparison`

Matrices

`r(results)`	estimates, standard errors, adjusted p-values, t statistics, and confidence interval for each parameter

7 Other examples

`sotable` computes results for whatever is stored in `e()`. We can use the `post` option on some postestimation commands to make `sotable` work on the cases covered by these commands. The postestimation commands `margins` and `nlcom` are clearly important. In this section, we provide illustrations of this use of `sotable`. Note that any frequentist postestimation command that will post `e(b)` and `e(V)` into `e()` will work in the manner discussed below. For example, `pwcompare` and `contrast` could also be used with `sotable` because they offer the `post` option.

7.1 margins

`margins` estimates average partial effects and contrasts thereof, among other estimands. By default, `margins` stores its results in `r()`, but it has a `post` option to store its results in `e()`. This section illustrates how to use `sotable` after `margins`.

Cattaneo (2010) included an empirical application that tested for effects of different levels of pregnant women smoking on the probability of a baby being born at a low birthweight. `cattaneo2.dta` is a nonrepresentative extract of the data used in Cattaneo (2010). Because the extract is not representative, the discussion below illustrates the statistical methods, but it should not be interpreted as evidence about the actual effects.

We begin by estimating the coefficients in a logit model of the low-birthweight indicator `lbweight` on indicators created from the treatment factor `msmoke` and some covariates to control for selection. The covariates are an indicator for the mother being married (`mmarried`), the mother's age (`mage`), the number of prenatal visits (`nprenatal`), and an indicator for whether it was the first baby born to the mother (`fbaby`).

```
. use cattaneo2, clear
(Excerpt from Cattaneo (2010) Journal of Econometrics 155: 138-154)
. logit lbweight i.msmoke mmarried mage nprenatal fbaby, vce(robust)
Iteration 0:  Log pseudolikelihood = -1057.6513
Iteration 1:  Log pseudolikelihood = -990.01763
Iteration 2:  Log pseudolikelihood = -966.00817
Iteration 3:  Log pseudolikelihood = -965.97905
Iteration 4:  Log pseudolikelihood = -965.97905
Logistic regression                                     Number of obs =  4,642
                                                        Wald chi2(7)  = 184.74
                                                        Prob > chi2   = 0.0000
Log pseudolikelihood = -965.97905                       Pseudo R2     = 0.0867
```

lbweight	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
msmoke						
1-5 cigarett..	-.2779935	.3397676	-0.82	0.413	-.9439258	.3879388
6-10 cigaret..	.8605174	.1912178	4.50	0.000	.4857374	1.235297
11 or more c..	.6593036	.1991424	3.31	0.001	.2689916	1.049616
mmarried	-.4325707	.1603298	-2.70	0.007	-.7468114	-.1183301
mage	.01334	.0135297	0.99	0.324	-.0131777	.0398576
nprenatal	-.1662089	.0194022	-8.57	0.000	-.2042364	-.1281813
fbaby	.0880387	.1391384	0.63	0.527	-.1846674	.3607449
_cons	-1.410956	.3673209	-3.84	0.000	-2.130892	-.6910207

Now we use `margins` to estimate the average treatment effect of moving each woman from "0 cigarettes daily" to "1–5 cigarettes daily", "6–10 cigarettes daily", and "11 or more cigarettes daily".

```
. margins, at(msmoke=(0,1,2,3)) contrast(atcontrast(r) nowald) vce(unconditional)
> post
Contrasts of predictive margins                         Number of obs = 4,642
Expression: Pr(lbweight), predict()
1._at: msmoke = 0
2._at: msmoke = 1
3._at: msmoke = 2
4._at: msmoke = 3
```

	Contrast	Unconditional std. err.	[95% conf.	interval]
_at				
(2 vs 1)	-.0118427	.0130329	-.0373868	.0137014
(3 vs 1)	.0590894	.0167301	.026299	.0918798
(4 vs 1)	.0416265	.0152798	.0116787	.0715743

Note that we specified the `post` option to post the `margins` results in `e()`. The p-values and confidence intervals reported by `margins` do not account for the multiple tests implied.

For each treatment level, our null hypothesis is that the average treatment effect is less than or equal to zero versus the alternative hypothesis that the average treatment effect is greater than zero. Below, we use `sotable` to calculate adjusted p-values and a confidence band for these parameters.

```
. sotable, alternative(upper)
Max-t results
       p-value = 0.001
Critical value = 2.389
```

depvar	Coef.	Std. Err.	z	P>|z|	[97.5% Conf.	Band]
r2vs1._at	-.0118427	.0130329	-0.909	0.988	-.0429744	.
r3vs1._at	.0590894	.0167301	3.532	0.001	.0191264	.
r4vs1._at	.0416265	.0152798	2.724	0.010	.0051278	.

At the FWER of 0.025, we can reject the null hypotheses for the levels of "6–10 cigarettes daily" and "11 or more cigarettes daily" but not for the level of "1–5 cigarettes daily".

7.2 nlcom, exponentiated coefficients

By default, the adjusted p-values reported by `sotable` are for tests against zero. The example in this subsection illustrates how to perform tests against a nonzero value.

Continuing with the previous example, instead of an average treatment effect, some researchers might be interested in the conditional log odds of going from the treatment level of "0 cigarettes daily" to each of the other treatment levels. For each level, the expression of interest is $\exp(\hat{\beta}_j)-1$, where $\hat{\beta}_j$ is the estimated coefficient on the indicator for level j. The $\exp(\hat{\beta}_j)$ is the estimated conditional odds ratio for level j. When the conditional odds ratio for level j is less than or equal to one, there is no effect. When the conditional odds ratio for level j is greater than one, there is an effect. Below, we use `nlcom` to estimate each of these expressions. Note that we repeat the `logit` command because we overwrote its results with the `post` option on `margins`. We also used the `post` option on `nlcom`.

```
. quietly logit lbweight i.msmoke mmarried mage nprenatal fbaby, vce(robust)
> coeflegend
. nlcom (exp(_b[1.msmoke])-1) (exp(_b[2.msmoke])-1) (exp(_b[3.msmoke])-1), post
       _nl_1: exp(_b[1.msmoke])-1
       _nl_2: exp(_b[2.msmoke])-1
       _nl_3: exp(_b[3.msmoke])-1
```

lbweight	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	-.2426983	.2573066	-0.94	0.346	-.74701	.2616134
_nl_2	1.364384	.4521123	3.02	0.003	.4782599	2.250508
_nl_3	.9334454	.385031	2.42	0.015	.1787984	1.688092

Now we use `sotable` to compute adjusted p-values for each of the individual hypotheses,

$$\begin{aligned} H_{0,j}&: \ \exp(\beta_j) - 1 \leq 0 \\ H_{a,j}&: \ \exp(\beta_j) - 1 > 0 \end{aligned} \quad \text{for } j \in \{1, 2, 3\}$$

```
. sotable, alternative(upper)
Max-t results
       p-value = 0.004
Critical value = 2.386
```

lbweight	Coef.	Std. Err.	z	P>|z|	[97.5% Conf.	Band]
_nl_1	-.2426983	.2573066	-0.943	0.987	-.8565446	.
_nl_2	1.364384	.4521123	3.018	0.004	.285797	.
_nl_3	.9334454	.385031	2.424	0.023	.0148919	.

We reject $H_{0,2}$ and $H_{0,3}$ at the FWER of 0.025, but we do not reject $H_{0,1}$.

7.3 Multiple tests against a nonzero value

Sometimes, we want to test whether multiple parameters are equal to values other than zero. This section shows how to do such a test using `nlcom` and `sotable`.

`bdsianesi5.dta` is an extract of the data used by Blundell, Dearden, and Sianesi (2005) in their study of the effects of education on wages. This extract was used in example 1 in the `teffects multivalued` entry in the *Stata Causal Inference and Treatment-Effects Estimation Reference Manual*. In the dataset, `wages` contains the hourly wages in pounds; `ed` contains the highest educational degree obtained; `paed` contains the highest educational level obtained by each individual's father; `math7` contains a score obtained on a standardized math test when the individual was seven; `read7` contains a score obtained on a standardized reading test when the individual was seven; and `london` and `eastern` are indicators for whether an individual lives in London or the East.

As in example 1 in the `teffects multivalued` entry in the *Stata Causal Inference and Treatment-Effects Estimation Reference Manual*, we use `teffects ra` to estimate the treatment effects. We specify the option `coeflegend` because we want to know the names of the parameters.

```
. use https://www.stata-press.com/data/r18/bdsianesi5, clear
(Excerpt from Blundell, Dearden, & Sianesi (2005) JRSSA 168: 473)
. teffects ra (wage london eastern paed math7) (ed), coeflegend
Iteration 0:  EE criterion =  1.041e-28
Iteration 1:  EE criterion =  1.032e-30
Treatment-effects estimation                    Number of obs     =      1,693
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
```

wage	Coefficient	Legend
ATE		
ed		
(O vs none)	1.191423	_b[ATE:r1vs0.ed]
(A vs none)	1.758726	_b[ATE:r2vs0.ed]
(H vs none)	3.986172	_b[ATE:r3vs0.ed]
POmean		
ed		
none	6.501982	_b[POmean:0.ed]

We are interested in testing the hypotheses, expressed in Stata notation, that

$$
\begin{aligned}
H_{0,1}&: \texttt{\_b[ATE:r2vs0.ed]} - \texttt{\_b[ATE:r1vs0.ed]} = 1\\
H_{a,1}&: \texttt{\_b[ATE:r2vs0.ed]} - \texttt{\_b[ATE:r1vs0.ed]} \neq 1\\
H_{0,2}&: \texttt{\_b[ATE:r3vs0.ed]} - \texttt{\_b[ATE:r2vs0.ed]} = 1.5\\
H_{a,2}&: \texttt{\_b[ATE:r3vs0.ed]} - \texttt{\_b[ATE:r2vs0.ed]} \neq 1.5
\end{aligned}
$$

We start by using `nlcom` to estimate the expressions implied by the tests; note that we use the option `post`.

```
. nlcom (_b[ATE:r2vs0.ed] - _b[ATE:r1vs0.ed] - 1)
>       (_b[ATE:r3vs0.ed] - _b[ATE:r2vs0.ed] - 1.5), post
       _nl_1: _b[ATE:r2vs0.ed] - _b[ATE:r1vs0.ed] - 1
       _nl_2: _b[ATE:r3vs0.ed] - _b[ATE:r2vs0.ed] - 1.5
```

wage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	-.4326966	.2394872	-1.81	0.071	-.9020829	.0366896
_nl_2	.7274454	.2895162	2.51	0.012	.1600042	1.294887

Now we can use `sotable` to test the hypotheses.

```
. sotable
Max-t results
       p-value = 0.023
Critical value = 2.228
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Band]
_nl_1	-.4326966	.2394872	-1.807	0.133	-.9661669	.1007737
_nl_2	.7274454	.2895162	2.513	0.023	.082533	1.372358

At the 0.05 significance level, we do not reject $H_{0,1}$, but we do reject $H_{0,2}$.

8 Simulations

The methods discussed in section 5 are simple adaptations of the formulas in Hothorn, Bretz, and Westfall (2008). This section illustrates that our simulation algorithm to approximate the critical values and the adjusted p-values works well. The simulations also illustrate some properties of one-sided tests.

The cases to be simulated are t distribution, lower tail; t distribution, upper tail; normal distribution, lower tail; and normal distribution, upper tail. We use `regress` to estimate the coefficients of a linear regression with normally distributed errors for the t distribution cases. We use `logit` to estimate the coefficients in a logit model for the normal distribution cases.

We use the same method to draw from the covariates for both the linear regression model and the logit model. We drew 6 covariates (`x1`, `x2`, `x3`, `x4`, `x5`, and `x6`), using the following process. First, we defined a covariance matrix,

$$\mathbf{W} = \begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{pmatrix}$$

Then, we let $\mathbf{C}$ be the Cholesky of $\mathbf{W}$ and let $\widetilde{\mathbf{C}}$ be in the inverse of $\mathbf{C}$. Next, we obtained a (1×6) vector $\mathbf{v}$. Each element in $\mathbf{v}$ is an independent draw from a χ^2 distribution with 25 degrees of freedom that has been normalized to have mean zero and variance one. Each element of $\mathbf{v}$ has been standardized, but its distribution is not close to a standard normal. Finally, we obtained the vector of covariates $\mathbf{x}$ by $\mathbf{x} = \mathbf{v}\widetilde{\mathbf{C}}$. This process produces a vector of correlated, nonnormally distributed covariates.

For the linear regression model, the outcome y_i is given by

$$y_i = \mathbf{x}_i\beta' + \epsilon_i$$

where the values in β vary over the designs, as discussed below, and ϵ_i is independently and identically normally distributed over the observations.

For the logit model, the outcome $\tilde{y}_i$ is given by

$$\tilde{y}_i = (\mathbf{x}_i\alpha' + \nu_i > 0)$$

where the values in α vary over the designs, as discussed below, and ν_i is independent and identically standard logistically distributed over the observations.

In all the designs, the coefficients on `x1`, `x2`, and `x3` were the coefficients of interest. The covariates `x4`, `x5`, and `x6` were treated as controls. The values of the coefficients on `x1`, `x2`, and `x3` varied over the designs.

We used three designs. The first design was "null", in which the vector of parameter values is set so that the null hypothesis is true with equality. When this is true, the rejection rate should be very close to the FWER. The second design was "alt", in which the vector of parameter values lies in the space of the alternative hypotheses. When the alternative hypothesis is true, the rejection rate is the power of the test; the closer the power is to one, the better. The third design was "null interior", in which the vector of parameter values lies in the interior of the space of the null hypothesis. For values in the interior of this space, the ideal rejection rate is zero. Table 2 gives the values for β and α for each design.

Table 2. Coefficient values

Linear regression model

Case	Design	β
upper	null	$J(1,3,0)$, $J(1,3,.3)$
upper	alt	$J(1,3,.25)$, $J(1,3,.3)$
upper	null interior	$J(1,3,-.25)$, $J(1,3,.3)$
lower	null	$J(1,3,0)$, $J(1,3,.3)$
lower	alt	$J(1,3,-.25)$,$J(1,3,.3)$
lower	null interior	$J(1,3,.25)$, $J(1,3,.3)$

Logit model

Case	Design	α
upper	null	$J(1,3,0)$, $J(1,3,.3)$
upper	alt	$J(1,3,.07)$, $J(1,3,.3)$
upper	null interior	$J(1,3,-.07)$, $J(1,3,.3)$
lower	null	$J(1,3,0)$, $J(1,3,.3)$
lower	alt	$J(1,3,-.07)$, $J(1,3,.3)$
lower	null interior	$J(1,3,.07)$, $J(1,3,.3)$

In each design, there were three coefficients to test. In the upper-tailed cases, the overall null hypothesis was that all three were less than or equal to zero, and the overall alternative hypothesis was that at least one of the three was greater than zero. In the lower-tailed cases, the overall null hypothesis was that all three were greater than or equal to zero, and the overall alternative hypothesis was that at least one of the three was less than zero.

Table 3 gives the rejection rates for the multiple-single-tests method and the max-t method for the null designs. The multiple-single-tests method computes the unadjusted p-value from the `regress` or `logit` results for each of the three coefficients of interest, and it rejects the overall null hypothesis if any of them are less than or equal to 0.025. The max-t method rejects if the max-t–adjusted p-value for the overall test is less than or equal to 0.025. In table 3, the ideal rejection rate is 0.025. The results for the multiple-single-tests method reveal that it is not usable because it does not control the FWER. The multiple-single-tests method rejects the true overall null hypothesis about 2.5 times too frequently. The max-t method, on the other hand, has a rejection rate that is very close to the nominal 0.025 in each case.

Table 3. Null design results

Case		Design	Multiple single	Max-t
t distribution	upper tail	null	0.0625	0.0255
t distribution	lower tail	null	0.0634	0.0234
normal distribution	upper tail	null	0.0683	0.0271
normal distribution	lower tail	null	0.0684	0.0254

It does not make sense to look at the results of the multiple-single-tests method for cases of alt or null interior, because the null results for this method show that it is not usable because it does not control the FWER. The alt and null interior cases measure how well a usable test procedure performs under cases besides the case in which the null hypothesis is true with equality.

Table 4 presents the rejection rates (power) for the max-t method for the cases of alt and null interior.

Table 4. Alt and null interior results

Case		Design	Max-t
t distribution	upper tail	alt	0.4049
t distribution	upper tail	null interior	0.0001
t distribution	lower tail	alt	0.4145
t distribution	lower tail	null interior	0.0001
normal distribution	upper tail	alt	0.2134
normal distribution	upper tail	null interior	0.0012
normal distribution	lower tail	alt	0.1949
normal distribution	lower tail	null interior	0.001

The ideal number for the alt designs is 1, but the distance from the null hypothesis values is not large enough to always reject. The ideal number of the null interior designs is zero, and the max-t results are close to the ideal.

In short, the simulations illustrate that the max-t method implemented in `sotable` performs well for the cases and designs considered. The simulations also illustrate that one should not use the multiple-single-tests method because it leads to rejection of the overall null hypothesis at a rate higher than the specified FWER.

9 Conclusion

Applied researchers should not ignore the problems of multiple testing. It has never been easier to perform simultaneous inference, but without methods to account for higher rejection rates, separating real results from noise becomes difficult. The `sotable` command discussed in this article and in Drukker (2023) fills a major hole in Stata's features for frequentist hypothesis testing. It enables researchers to perform multiple hypothesis tests and to compute the associated confidence bands for multiple hypotheses.

Many applied studies have one-sided alternatives. We have discussed two applications in detail, one real and the other simulated but realistic. This article has argued that researchers should be able to easily perform one-sided inference that addresses their questions, without fear of "fishing" or "p-hacking" accusations, and it has shown how to do so using the `sotable` command.

This article and Drukker (2023) have discussed overall hypotheses that have a null hypothesis in which all the individual null hypothesis are true. This form for the overall null hypothesis is handled by max-t tests. Min-t tests can handle an overall null hypothesis in which at least one of the individual null hypotheses is true. But min-t tests suffer from a tremendous lack of power. Future work could find and implement a more powerful test for the at-least-one form for the overall null hypothesis.

10 Acknowledgments

We thank an anonymous referee and the editor for comments and feedback.

11 Programs and supplemental material

To install the software files as they existed at the time of publication of this article, type

```
. net sj 24-3
. net install st0718_1     (to install program files, if available)
. net get st0718_1         (to install ancillary files, if available)
```

12 References

Bickel, P. J., and K. A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden–Day.

Blundell, R., L. Dearden, and B. Sianesi. 2005. Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society*, A ser., 168: 473–512. https://doi.org/10.1111/j.1467-985X.2004.00360.x.

Brown, A. W., D. G. Altman, T. Baranowski, J. M. Bland, J. A. Dawson, N. V. Dhurandhar, S. Dowla, K. R. Fontaine, A. Gelman, S. B. Heymsfield, W. Jayawardene, S. W. Keith, T. K. Kyle, E. Loken, J. M. Oakes, J. Stevens, D. M. Thomas, and D. B. Alliso. 2019. Childhood obesity intervention studies: A narrative review and guide for investigators, authors, editors, reviewers, journalists, and readers to guard against exaggerated effectiveness claims. *Obesity Reviews* 20: 1523–1541. https://doi.org/10.1111/obr.12923.

Casella, G., and R. L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.

Cattaneo, M. D. 2010. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155: 138–154. https://doi.org/10.1016/j.jeconom.2009.09.023.

Drukker, D. M. 2023. Simultaneous tests and confidence bands for Stata estimation commands. *Stata Journal* 23: 518–544. https://doi.org/10.1177/1536867X231175333.

Elfenbein, D. W., and B. McManus. 2010. A greater price for a greater good? Evidence that consumers pay more for charity-linked products. *American Economic Journal: Economic Policy* 2(2): 28–60. https://doi.org/10.1257/pol.2.2.28.

Hothorn, T., F. Bretz, and P. Westfall. 2008. Simultaneous inference in general parametric models. *Biometrical Journal* 50: 346–363. https://doi.org/10.1002/bimj.200810425.

Wooldridge, J. M. 2020. *Introductory Econometrics: A Modern Approach*. 7th ed. Boston: Cengage Learning.

About the authors

David M. Drukker is an associate professor of economics at Clemson University.

Kevin S. S. Henning is a clinical associate professor of business analysis at Sam Houston State University.

Christian Raschke is an associate professor of economics at Sam Houston State University.

A Appendixes

A.1 Simulation of the type I error myth

Our simulation mimics a two-sample study for a treatment effect when subjects are randomly allocated to group 1 or group 2 and only a treatment effect greater than zero is of interest. We implement three designs. In all the designs, we are interested only in an effect that is strictly positive. In the first design, the effect is exactly zero, and we expect the rejection rate to equal the specified significance level. In the second case, the effect is negative, and we expect the rejection rate to be less than the specified significance level. In the third case, the true effect is positive, and the rejection rate is the power of the test. The closer the power is to one, the better.

In all the designs, the data for each group are drawn from a normal distribution. Let $\mu_{d,g}$ be the mean for the group $g \in \{1, 2\}$ data in design $d \in \{1, 2, 3\}$. The standard deviations differ over the groups but not over the designs. We let $\sigma_1 = 2$ be the standard deviation for group 1 and let $\sigma_2 = 2.5$ be the standard deviation for group 2. The sample size for each group is 60. For each design, the effect is $\alpha_d = \mu_{d,1} - \mu_{d,2}$, and effects that are less than or equal to zero are not useful. For each design d we specify

$$\begin{aligned} H_0&: \alpha_d \leq 0 \\ H_a&: \alpha_d > 0 \end{aligned} \tag{A1}$$

The following table gives the group means and the effects for each design.

Design	$\mu_{1,i}$	$\mu_{2,i}$	True effect
1	2	2	$2 - 2 = 0$
2	1.5	2	$1.5 - 2 = -0.5$
3	2.5	2	$2.5 - 2 = 0.5$

To test H_0 versus H_a, we use a two-sample t test with unequal variances and the Satterthwaite method for approximating the degrees of freedom.[11] For each design, we drew sample data from this data-generating process 10,000 times. In each sample, we calculated the p-value for the two-sample t test. For each design, we then used the 10,000 p-values to calculate the rejection rate when the significance level is 0.05 and the rejection rate when the significance level is 0.025. The results are in the following table.

Design	Sig. level 0.05	Sig. level 0.025
1	0.0497	0.0252
2	0.003	0.0009
3	0.3232	0.219

In design 1, we expect the rejection rates to equal the significance level. The design 1 simulation results illustrate that the rejection rate of this upper-tailed test is approximately equal to the specified significance level. This debunks the myth that the rejection rate is twice the specified significance level.

In design 2, the ideal rejection rate is 0 because the true effect is negative. Both significance levels yield rejection rates that are close to 0. The rejection rate for the 0.025 significance level is closer to 0 than the rejection rate for the 0.05 significance level.

In design 3, the rejection rate is measuring the power of the test, because the effect is strictly positive. The higher the power, the better the test. Here we see the cost of using a significance level of 0.025 instead of 0.05. Using the significance level of 0.025 significantly reduces the power of the test.

Here is one final justification of the one-sided p-value for a joint test of size and direction. Suppose that we are going to use the results presented by `regress` to test the hypothesis in (A1). A way to adjust the p-value that `regress` gives for a one-sided test is to use one half the reported p-value if the estimated coefficient is positive. If the estimated coefficient is negative, use 1. This method logically accounts for the joint test of size and direction. It never rejects when the estimated coefficient is negative, and it divides the reported p-value by 2 to account for throwing away the negative-valued rejections. This one-sided-adjustment method provides identical inference to the well-known, one-sided p-value formula of `1 - t(`df`,` t`)`, where `t()` calculates the cumulative

11. `ttest` with option `unequal` does these calculations. The methods and formulas section for `ttest` provides the details.

distribution function of the t distribution with df degrees of freedom and t is the t statistic. When the estimated coefficient is positive, the one-sided-adjusted p-value and the one-sided p-value are numerically identical. When the estimated coefficient is negative, neither the one-sided adjustment nor the one-sided p-value will reject.[12]

A.2 Algorithms

Algorithm 1 describes how `sotable` approximates the critical value c_{upper} and the adjusted p-values in (7). Algorithm 2 describes how `sotable` approximates the critical value c_{lower} and the adjusted p-values in (11). Algorithm 1 and algorithm 2 are simple variations on algorithm 1 in Drukker (2023).

12. When the estimated coefficient is negative, the one-sided-adjustment p-value is 1, which cannot lead to a rejection. When the estimated coefficient is negative, the one-sided p-value is greater than 0.5, which cannot lead to a rejection at any of the conventional significance levels.

Algorithm 1 Algorithm for approximating c_{upper} and upper-adjusted p_j

Initial values:

$\widehat{\mathbf{V}}$ is the estimated variance of the parameters of interest, and it is extracted from `e(V)`. This algorithm avoids constructing $\widehat{\mathbf{C}}$ from $\widehat{\mathbf{V}}$.

Let R be a large number of simulation repetitions. `sotable` uses $R = 1,000,000$ by default.

1. Draw $r = \{1, 2, \ldots, R\}$ independent copies of $(z_1, z_2, \ldots, z_q) \sim N(0, \widehat{\mathbf{V}})$. Let $(z_{r,1}, z_{r,2}, \ldots, z_{r,q})$ be the elements of $(z_1, z_2, \ldots, z_q)$ in draw r.

2. For the normal case:

   Let $(\check{t}_{r,1}, \check{t}_{r,2}, \ldots, \check{t}_{r,q}) = \left(z_{r,1}/\sqrt{\widehat{\mathbf{V}}[1,1]}, z_{r,2}/\sqrt{\widehat{\mathbf{V}}[2,2]}, \ldots, z_{r,q}/\sqrt{\widehat{\mathbf{V}}[q,q]}\right)$.

   For the t case:

   a. Let $g_r = \sqrt{\texttt{df}/\texttt{rchi2(df)}}$ for each r, where `rchi2()` is the Stata function that generates χ^2 variates. The formula does not vary over the repetitions, but there is a separate draw for each r. In step 2b, I use these g_r variates to convert standardized normal variates to t-distributed variates.

   b. Let $(\check{t}_{r,1}, \check{t}_{r,2}, \ldots, \check{t}_{r,q}) =$
   $g_r \left(z_{r,1}/\sqrt{\widehat{\mathbf{V}}[1,1]}, z_{r,2}/\sqrt{\widehat{\mathbf{V}}[2,2]}, \ldots, z_{r,q}/\sqrt{\widehat{\mathbf{V}}[q,q]}\right)$.

3. For each draw r, let $\tilde{t}_r = \max\{(\check{t}_{r,1}, \check{t}_{r,2}, \ldots, \check{t}_{r,q})\}$.

4. Then c_{upper} is approximated by the sample quantile of the R draws of $\tilde{t}_r$.

5. The upper-adjusted p-value p_j is approximated by the fraction of the R draws for which $\tilde{t}_r$ is greater than the t statistic t_j.

Algorithm 2 Algorithm for approximating c_{upper} and upper-adjusted p_j

The initial values and steps 1 and 2 are the same as in algorithm 1.
Steps 3, 4, and 5 below replace the corresponding steps in algorithm 1.

3. For each draw r, let $\tilde{t}_r = \min\{(\check{t}_{r,1}, \check{t}_{r,2}, \ldots, \check{t}_{r,q})\}$.
4. Then c_{lower} is approximated by the sample quantile of the R draws of $\tilde{t}_r$.
5. The lower-adjusted p-value p_j is approximated by the fraction of the R draws for which $\tilde{t}_r$ is less than the t statistic t_j.