The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–16

distcomp: Comparing distributions

David M. Kaplan
Department of Economics, University of Missouri
Columbia, MO, USA
kaplandm@missouri.edu

Abstract. The `distcomp` command is introduced and illustrated. The command assesses whether or not two distributions differ at each possible value while controlling the probability of any false positive, even in finite samples. Syntax and the underlying methodology (from Goldman and Kaplan 2018) are discussed. Multiple examples illustrate the `distcomp` command, including revisiting the experimental data of Gneezy and List (2006) and the regression discontinuity design of Cattaneo, Frandsen, and Titiunik (2015).

Keywords: st0001, distcomp, familywise error rate, ksmirnov, regression discontinuity, treatment effects

1 Introduction

The new `distcomp` command implements a new statistical procedure for comparing distributions, introduced in Goldman and Kaplan (2018). The usage is similar to a two-sample *t*-test or two-sample Kolmogorov–Smirnov test, i.e., `ttest` or `ksmirnov` (respectively) with the `by` option (see [R] **ttest** or [R] **ksmirnov**). However, instead of only comparing the distributions' means (like `ttest`) or only testing a single hypothesis of distributional equality (like `ksmirnov`), `distcomp` assesses equality of the distribution functions point by point. Thus, instead of a single rejection or non-rejection, `distcomp` displays ranges of values in which the distributions' difference is statistically significant. (Goodness-of-fit test results like from `ksmirnov` can also be shown.)

The new procedure controls false positives with a property called strong control of the familywise error rate (described later). As a special case, if the distributions are truly identical, then no ranges will be deemed statistically significant 95% of the time if a 5% level is used. This familywise error rate is controlled in finite samples (not just asymptotically).

Even for goodness-of-fit testing, `distcomp` may be preferred to `ksmirnov` since the new method's sensitivity to deviations is more evenly spread across the distribution. This new goodness-of-fit test was proposed by Goldman and Kaplan (2018), refining an idea from Buja and Rolke (2006). Specifically, the Kolmogorov–Smirnov test has long been known to lack sensitivity to deviations in the tails of a distribution (e.g., Eicker 1979, p. 117). For example, if one sample has observed values $0.02, 0.04, \ldots, 0.98$, like a standard uniform distribution, the second sample may have even six out of 21 values exceeding one million without a two-sided Kolmogorov–Smirnov test rejecting at a 10% level. In contrast, `distcomp` rejects equality at even a 1% level. The following Stata code shows such a result.

```
. set obs 69
number of observations (_N) was 0, now 69
. gen grp = (_n>49)
. gen y = (_n<=49)*(_n/50) + (_n>49)*(_n-49)/21
. replace y = 1000000+_n if _n>63
(6 real changes made)
. ksmirnov y , by(grp) exact

Two-sample Kolmogorov-Smirnov test for equality of distribution functions
 Smaller group       D       P-value      Exact
 ----------------------------------------------
 0:                  0.3000    0.078
 1:                 -0.0378    0.960
 Combined K-S:       0.3000    0.155      0.121
. distcomp y , by(grp) alpha(0.01) noplot p
Comparing distribution of y when grp=0 vs. grp=1

Global test of equality of two CDFs:
    Simulated p-value = .0055889
    At a 10% level: reject
    At a  5% level: reject
    At a  1% level: reject
```

Section 2 discusses the methodology at a relatively intuitive level. Section 3 describes syntax and usage of `distcomp`. Section 4 provides empirical examples that can be replicated with the provided do-file. Section 5 shows some of the theoretical foundations before concluding. Readers interested in related methods (one-sided, one-sample, or uniform confidence bands) are referred to Goldman and Kaplan (2018) and the corresponding R code. Abbreviations are used for cumulative distribution function (CDF), goodness-of-fit (GOF), Kolmogorov–Smirnov (KS), and familywise error rate (FWER).

2 A gentle introduction to methodology

This section discusses methodology at a relatively non-technical level (compared to Section 5). Interest is in the distribution of some outcome variable (like wage) for two groups (like union members and non-members). More specifically, the question is whether (and where) the two cumulative distribution functions (CDFs) are different. Let $F(\cdot)$ be the first group's CDF (like the CDF of wage for union members); $G(\cdot)$ is the second group's CDF. Estimated CDFs $\widehat{F}(\cdot)$ and $\widehat{G}(\cdot)$ are computed from iid samples. These are the stair-step functions like in the graphs in Section 4.

Consider the Kolmogorov–Smirnov (KS) goodness-of-fit (GOF) test. The null hypothesis is

$$\text{(GOF)} \quad H_0\colon F(r) = G(r) \text{ for all } r, \tag{1}$$

identical CDFs. If H_0 is true, then $\widehat{F}(\cdot)$ and $\widehat{G}(\cdot)$ should be "close" to each other; if not, the test rejects.

KS defines "close" with vertical distance. At point r, this distance is

$$\widehat{D}(r) \equiv |\widehat{F}(r) - \widehat{G}(r)|. \tag{2}$$

To compare the entire functions, KS looks across all r to find the biggest gap, $\max_r \widehat{D}(r)$. Happily, under H_0, the sampling distribution of the "biggest gap" does not depend on the true distribution, so it can be simulated. Thus, finite-sample p-values can be simulated, without any asymptotic approximation, and the KS test can control the type I error rate at level α.

Instead of (1), null hypotheses can be defined to show *where* two distributions differ. For each possible value r of the outcome variable, define

$$H_{0r}\colon F(r) = G(r). \tag{3}$$

The GOF null hypothesis could be rewritten as

$$\text{(GOF again)} \quad H_0\colon \text{all } H_{0r} \text{ are true.} \tag{4}$$

Whereas the GOF test only distinguishes whether all H_{0r} are true or at least one is false, now we care about specifically which H_{0r} are true and which are false. This is summarized by `distcomp` by the ranges of r for which H_{0r} is rejected.

It does not work to run a level α t-test on all H_{0r}. Imagine just two values $r = 1, 2$, and $\alpha = 0.1 = 10\%$. Let both H_{0r} be true. Then, each test's probability of not rejecting is 0.9. Assuming independence for simplicity, the probability that neither rejects is $(0.9)(0.9) = 0.81$, so the probability that one or both rejects is $1 - 0.81 = 0.19$. That is, the probability of at least one type I error (false positive) is 19%, much larger than the desired $\alpha = 10\%$. This is called the "multiple testing problem."

The following is one way to define type I error control for multiple testing. A "familywise error" is made if at least one true H_{0r} is rejected. For example, if H_{0r} is true only for $r \le 0$, then a familywise error occurs if any H_{0r} with $r \le 0$ is rejected. The probability of such an error is the familywise error rate (FWER):

$$\text{FWER} \equiv \Pr(\text{reject any true } H_{0r}). \tag{5}$$

"Weak control of FWER at level α" guarantees FWER $\le \alpha$ when all H_{0r} are true. The `distcomp` methodology achieves "strong control" of FWER, meaning FWER $\le \alpha$ regardless of which H_{0r} are true. Put differently: with strong control of FWER at a 10% level, there will be zero false positives 90% of the time.

One way to test (3) and achieve strong control of FWER is to extend the KS approach. Specifically, imagine that c is the simulated critical value given n and α. It seems natural to reject H_{0r} when $\widehat{D}(r) > c$.

The weak control of FWER comes from the KS GOF test's properties. When the GOF H_0 is true (so all H_{0r} are true), the probability of GOF rejection is α: $\Pr(\max_r \widehat{D}(r) > c) = \alpha$. Since $\max_r \widehat{D}(r) > c$ is equivalent to "at least one H_{0r} is rejected," the probability of rejecting at least one H_{0r} (i.e., FWER) is also α.

The KS approach also has strong control of FWER. The intuition is that when all H_{0r} are true (as with weak control of FWER), the opportunity for a familywise error is greatest, so FWER is highest. If fewer H_{0r} are true, then there is less opportunity for a familywise error. In the extreme, when no H_{0r} is true, it is impossible to make a familywise error. This intuition is proved correct for a certain class of methods including the KS approach and `distcomp`, although it can fail in other cases.

The KS approach controls FWER, but it distributes power unevenly. This means it detects certain types of differences between $F(\cdot)$ and $G(\cdot)$ very well, but detects others very poorly. Unless we know what type of difference to expect, it seems prudent to desire power against all differences. This is achieved by `distcomp`, without sacrificing the finite-sample strong control of FWER.

One shortcoming of KS is its implicit symmetry. The definition of $\widehat{D}(r)$ has an absolute value; positive and negative values are treated the same. This makes sense with normal distributions, but normality is only an asymptotic approximation. The finite-sample distribution of $\widehat{F}(r)$ is actually binomial, which is especially skewed (not symmetric) when r is in the tails.

KS also ignores scale differences. The variance of $\widehat{F}(r)$ is highest when r is the median, and nears zero when r is in the tails. The variance of $\widehat{D}(r)$ similarly varies with r. Even if $\widehat{D}(r)$ were normal, it would be much more likely have a large value with median r than with r in the tails, where H_{0r} is very unlikely to be rejected.

By using the finite-sample sampling distributions of $\widehat{F}(r)$ and $\widehat{G}(r)$, `distcomp` accounts for both skewness and scale differences across r, unlike KS.

3 The distcomp command

The `distcomp` command compares two distributions. One variable in the data (*varname* below) is the variable for comparison, like price or income. Another variable (*groupvar* below) takes only two distinct values, defining two groups, like an indicator/dummy for male whose value is 0 or 1, or a state abbreviation whose value is NY or CA.

The validity of two assumptions should be considered in practice. First, sampling is assumed independent and identically distributed (iid) from the two respective group population distributions, and it is assumed the groups are sampled independently. Second, the variable of interest (*varname*) is assumed to have a continuous distribution, but some amount of discreteness is ok. In particular, if there are duplicate values within each sample, but no "ties" (same value observed in both samples), then the properties remain the same. However, if there are many ties, then the theoretical results do not apply directly and the properties may change substantially. In the absence of theoretical results allowing ties, simulations suggest the method may become conservative, controlling the FWER at a level even lower than the level specified. One such simulation is included in the accompanying `distcomp_examples.do` file, in which the nominal FWER is 10% but simulated FWER is near 0%.

The first results displayed by `distcomp` are for the global (goodness-of-fit, GOF) test. This is the same type of test as the (two-sample) Kolmogorov–Smirnov test. That is, the null hypothesis is that the two cumulative distribution functions (CDFs) are identical. This could be false even if the two distributions' means are identical (and `ttest` does not reject), e.g., with normal distributions with the same mean but different standard deviation. The global test results are always reported for levels 1%, 5%, and 10%. Optionally, a p-value is also reported; it must be simulated and can substantially increase computation time with large sample sizes. (To avoid misinterpretation, when the simulated value is zero, 0.0001 is reported since there are 10^4 simulation replications.) The test is generally more powerful than Kolmogorov–Smirnov. The GOF methodology was proposed by Goldman and Kaplan (2018), refining an idea from Buja and Rolke (2006).

The second results displayed are for a multiple testing procedure. They show ranges of values for which the difference between CDFs is statistically significant, accounting for the multiple testing nature of the procedure (i.e., many different points are tested simultaneously). Instead of a single, GOF null hypothesis, there is a set of many null hypotheses. Within the set, each individual hypothesis specifies equality of the two CDFs at a different point. That is, if $F(\cdot)$ and $G(\cdot)$ are the two CDFs, then each individual null hypothesis is $H_{0x}\colon F(x) = G(x)$, and the set of such hypotheses for all possible values of x is considered. The multiple testing procedure rejects equality at certain values of x while controlling the probability of *any* type I error (false positive). The probability of any false positive is known as the familywise error rate (FWER). The `distcomp` procedure controls the finite-sample (not just asymptotic) FWER at the desired level specified by the user. The output shows the ranges of x where $H_{0x}\colon F(x) = G(x)$ is rejected. This methodology is from Goldman and Kaplan (2018).

By default, a plot is generated with the empirical CDFs of the two groups, along with the rejected ranges (if any).

The restriction of FWER levels to 1%, 5%, and 10% allows nearly instantaneous computation (when a p-value is not requested). The reason is that a table of precise "critical values" for these specific levels has been simulated ahead of time. In small samples, as with the Kolmogorov–Smirnov test, often it is impossible to attain exactly 1%, 5%, or 10%. For example, for certain sample sizes, it may only be possible to have FWER of 9.6% or 10.6%, but nothing in between. To make this transparent, the exact finite-sample FWER level has also been pre-simulated and is returned by `distcomp`. In some cases, an analytic formula (based on simulations) is used, in which case the pre-simulated exact FWER is not available. Except with very small samples, the practical difference between specified and actual FWER level is usually negligible.

Syntax, options, and stored results are now shown; prefix `by` is allowed.

`distcomp` *varname* [*if*] [*in*] `, by(`*groupvar*`)` [`alpha(`#`) pvalue noplot`]

`by(`*groupvar*`)` is required. It specifies a binary variable that identifies the two groups whose distributions are compared. (The variable does not need to have values 0 and

1 specifically; any two values are fine, like 1 and 5, or “cat” and “dog.”)

`alpha(#)` specifies the familywise error rate (FWER) level, as a decimal. The default is 0.10, i.e., 10% probability of any false positive. Other accepted values are 0.05 and 0.01 (meaning 5% and 1%).

`pvalue` requests a global p-value, computed by simulation. The computation time depends on the sample size. For prohibitively large samples, the global test’s rejections at levels 1%, 5%, and 10% (which do not require simulation) can be used to roughly approximate the p-value. For example, if only the 10% test rejects, then $0.05 < p < 0.10$.

`noplot` suppresses the plot.

Scalars

`r(rej_gof10)`	GOF rejection, 10% level	`r(p_gof)`	GOF p-value
`r(rej_gof05)`	GOF rejection, 5% level	`r(alpha_sim)`	simulated FWER level
`r(rej_gof01)`	GOF rejection, 1% level	`r(alpha)`	specified FWER level

Matrices

`r(N)`	numbers of observations	`r(rej_ranges)`	ranges with CDF equality rejected

Certain stored results may be missing in some cases. When `r(alpha_sim)` is not applicable, it is set to missing. The GOF p-value is only computed if requested. When no ranges are rejected, the `r(rej_ranges)` is a 1-by-2 matrix with both entries missing.

Also, `r(N)` shows three numbers; in order: the overall, first group, and second group number of observations used for analysis.

4 Examples

The examples in this section can all be replicated with the file `distcomp_examples.do`. Some code is omitted here to conserve space.

4.1 Simple example with built-in dataset

The following example compares the hourly wage distributions of union and non-union workers in the NLSW 1988 extract shipped with Stata, with a 10% statistical significance level.

```
. sysuse nlsw88 , clear
(NLSW, 1988 extract)
. set scheme sj
. distcomp wage , by(union) alpha(0.10) p
Comparing distribution of wage when union=0 vs. union=1

Global test of equality of two CDFs:
    Simulated p-value < .0001
    At a 10% level: reject
    At a  5% level: reject
    At a  1% level: reject

With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following
ranges of wage:
     from         to
 2.383252   2.520128
 2.568437   11.51368
 11.52979    11.5781
```

Three main results are displayed. The first result says that the global/GOF null hypothesis that the two wage distributions are identical is rejected at a 1% (and thus 5% and 10%) level. The statistical significance is even greater since the p-value is even lower than 0.01. There is very strong evidence that union and non-union wage distributions are not identical. The second result shows the range of wage values at which CDF equality is rejected while controlling the FWER at 10%. In this example, the range covers most of the wage distribution, from around the 2nd percentile ($2.38/hr) to almost the 86th percentile ($11.61/hr). This suggests a restricted first-order stochastic dominance relationship, as defined in Condition I of Atkinson (1987, p. 751). The empirical CDF graph (Figure 1) illustrates this result, too.

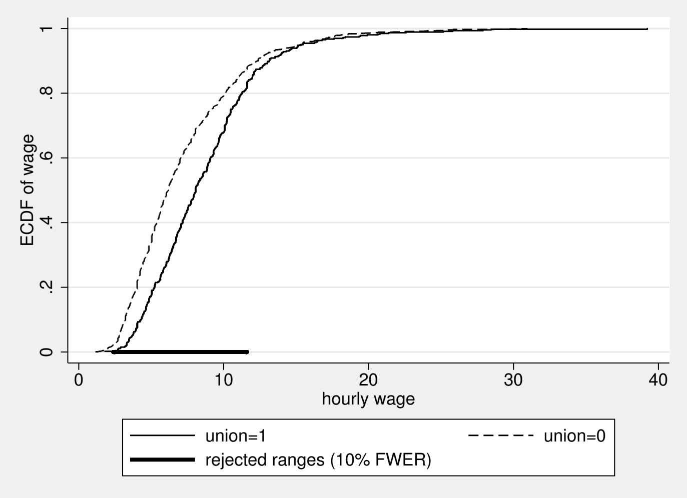

Figure 1: Empirical CDFs from NLSW.

Figure 1 shows the default plot generated by `distcomp`. It shows the empirical CDFs for union wage and non-union wage. These are step functions, but each step is very small since the sample size is moderate (461 union, 1417 non-union). The thick horizontal line near the bottom shows the ranges where CDF equality was rejected (10% FWER level). It is clear that union wages tend to be higher in the sample (hence the union CDF lies below the non-union CDF). However, the empirical CDFs alone cannot show statistical significance. The line at the bottom shows that this difference is indeed statistically significant at a 10% FWER level across most of the distribution, though not the upper tail.

Union and non-union wages can be compared for certain subgroups, with an `if` condition or `by` prefix. For example, the above analysis can be repeated for each different `race` group with `bysort race :  distcomp wage , by(union) p` or for only married individuals with `distcomp wage if married==1 , by(union) p`.

4.2 Example with simulated data

The following example uses simulated data. The code (including the random seed) to replicate the simulated data is in `distcomp_examples.do` but omitted here. The control group sample has 50 observations drawn independently from a standard normal distribution. The first treatment group also has 50 observations drawn independently from a standard normal distribution, i.e., the treatment and control population distributions are identical (but sample values differ). The second treatment group is the same for below-median individuals, but the treatment increases the outcome by two units for individuals with above-median values. The third treatment has no effect on the mean, but it affects the spread: values are drawn from a mean-zero normal distribution with standard deviation of three.

Running `distcomp` produces the following results.

```
. distcomp y1 , by(treated) p // no effect
Comparing distribution of y1 when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Simulated p-value = .395
    At a 10% level: do not reject
    At a  5% level: do not reject
    At a  1% level: do not reject
. distcomp y2 , by(treated) p // effect only above median (zero)
Comparing distribution of y2 when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Simulated p-value < .0001
    At a 10% level: reject
    At a  5% level: reject
    At a  1% level: reject

With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following
ranges of y2:
     from         to
 .9370626    3.08856
. distcomp y3 , by(treated) p // bigger in tails, 0 at median
Comparing distribution of y3 when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Simulated p-value = .0051
    At a 10% level: reject
    At a  5% level: reject
    At a  1% level: reject
```

```
With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following
ranges of y3:
     from         to
-3.862472  -1.974054
-1.279222  -1.172997
 1.558096   3.265678
```

Above, for the y1 case where the control and treatment distributions are equal, nothing is statistically significant at conventional levels. The distcomp GOF test does not reject equality even at a 10% level (the GOF p-value is 0.395), so no ranges are rejected at 10% FWER level. The empirical CDFs differ, but distcomp says these differences are not statistically significant at a 10% level.

For the y2 case where the treatment has an effect, but only above the median (which is zero), the distcomp results reflect this. First, equality of the distributions is rejected at a level well below 1% ($p < 0.0001$). Then, more specifically, distcomp says equality is rejected over the range $[0.937, 3.089]$. The distributions indeed differ over this range. They actually differ over the larger range from zero to infinity, but there is not enough data to be certain that differences closer to zero are statistically significant, and similarly for differences far in the upper tail (above 3.089).

For the y3 case where the treatment affects the standard deviation, the true CDFs differ everywhere except at zero, and again distcomp reflects this. In addition to rejecting global equality at a 1% level, distcomp identifies three specific ranges of values (that exclude zero) where the distributions differ. Similar to the second case, it is most difficult to infer a difference near zero (where the CDFs are actually equal) and far in the tails (where there are few/no observations). Given the same FWER level, more data would be required to enlarge the ranges where we are statistically confident in a CDF difference.

4.3 Example with experimental data

The following example uses data from Gneezy and List (2006) to test for distributional treatment effects. A longer version appears in Goldman and Kaplan (2018, §8.1). In brief, Gneezy and List (2006) paid control group individuals an advertised hourly wage and treatment group individuals an unexpectedly larger "gift" wage upon arrival. The "gift exchange" question from behavioral economics is whether the higher wage induces higher effort in return. The experiment is run separately for library data entry and door-to-door fundraising tasks. The sample sizes are very small: 10 and 9 for control and treatment (respectively) for the library task, and 10 and 13 for fundraising. With small samples, the finite-sample FWER control of distcomp is especially desirable. Complementing the original results of Gneezy and List (2006), Goldman and Kaplan (2018) examine heterogeneity in the treatment effect during the first few hours of work, with results seen below.

```
. distcomp ylib , by(treated) alpha(0.05) p noplot
Comparing distribution of ylib when treated=0 vs. treated=1
```

```
Global test of equality of two CDFs:
    Simulated p-value = .14105487
    At a 10% level: do not reject
    At a  5% level: do not reject
    At a  1% level: do not reject
. distcomp yfun , by(treated) a(0.05) p noplot
Comparing distribution of yfun when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Simulated p-value = .03979024
    At a 10% level: reject
    At a  5% level: reject
    At a  1% level: do not reject

With strong control of FWER at a 5% level,
CDF equality is rejected at all points in the following
ranges of yfun:
     from         to
        8         14
```

For the library task, although the sample values look very different, the sample sizes are too small for the differences to be statistically significant at a 5% FWER level (two-sided). The FWER level would have to be 14.1% (the GOF p-value) before rejecting equality in any range.

For the fundraising task, even though the sample sizes are again small, the treatment effect is statistically significant at a 5% level (two-sided). The p-value is 0.040 for CDF equality. More specifically, `distcomp` identifies the range of \$8 to \$14 as statistically significant at a 5% FWER level. (With 10% FWER, the range \$21 to \$26 is also significant.) This range is near the bottom of the distribution. Opposite the library data entry task, where the gift wage treatment seemed to have the biggest effect on the upper end of the productivity distribution, the gift wage seems to have the biggest effect on the bottom end of the productivity distribution for door-to-door fundraising.

Figures 2 and 3 show the fundraising and library data entry empirical CDFs, respectively. These are again the graphs generated automatically by `distcomp`. These graphs show the direction of the gift wage effect (higher productivity), but without `distcomp` it is unclear where the differences are statistically significant.

4.4 Example with regression discontinuity

The following regression discontinuity example uses data from Cattaneo, Frandsen, and Titiunik (2015). A longer version (with results from R code) appears in Goldman and Kaplan (2018, §8.2). In brief, the research question is about the benefit of incumbency in U.S. Senate elections. The regression discontinuity idea is essentially to consider elections where the incumbent won the prior election by a very small margin. Cattaneo, Frandsen, and Titiunik (2015) discuss a balance test-based bandwidth selection that suggests `h=0.75` percentage points is a small enough margin of victory that the outcome is (almost) as good as randomized.

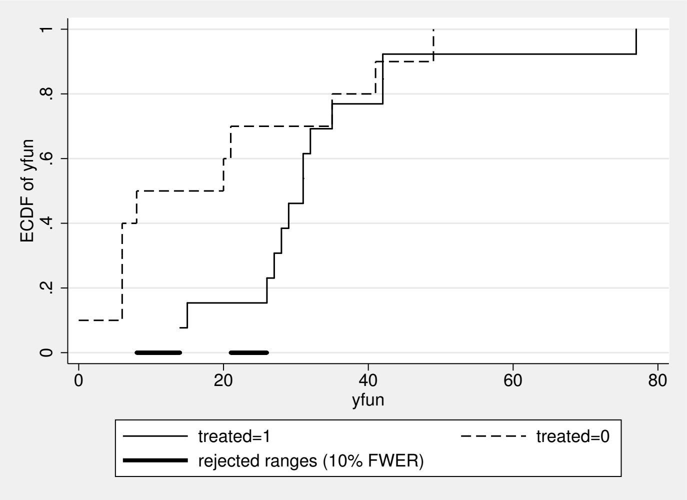

Figure 2: Empirical CDFs from experiment (fundraising).

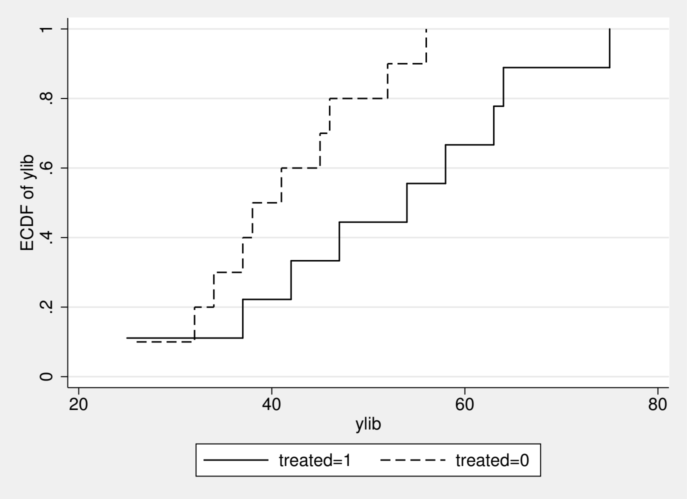

Figure 3: Empirical CDFs from experiment (library data entry).

In the following code and results, demmv is the Democratic margin of victory in the previous election for some Senate seat (in percentage points), which is negative if the Republican won. Thus, the incumbent is a Democrat if demmv exceeds the threshold R0=0. Also, demvoteshfor2 is the Democratic vote share in the current election for the same Senate seat. Below, the distribution of Democratic vote share is compared when the incumbent is a Democrat to when the incumbent is a Republican, restricting to cases where the incumbent's election was determined by a 0.75 point or smaller margin of victory.

```
. insheet using "https://sites.google.com/site/rdpackages/rdlocrand
> /r/rdlocrand_senate.csv",clear
(14 vars, 1,390 obs)
. scalar h = 0.75
. scalar R0 = 0
. gen D_incumbent = (demmv>=R0)
. distcomp demvoteshfor2 if demmv>=R0-h & demmv<=R0+h , by(D_incumb
> ent) a(0.10) p
Comparing distribution of demvoteshfor2 when D_incumbent=0 vs. D_in
> cumbent=1

Global test of equality of two CDFs:
    Simulated p-value = .00721501
    At a 10% level: reject
    At a  5% level: reject
    At a  1% level: reject

With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following
ranges of demvoteshfor2:
     from         to
 43.21114   47.73871
 47.81345   51.70049
 51.83689   56.64765
```

The results show the incumbency effect to be statistically significant across most of the distribution. With only two slim gaps, equality of the vote share distributions is rejected over the range from 43.2% to 56.6% of the vote. Of course, beyond statistical significance, it is also important to estimate the magnitude of the incumbency effect, by the usual regression discontinuity estimator.

Figure 4 shows that in the sample, the vote share distribution for incumbents (who had a very small margin of victory) first-order stochastically dominates the distribution for challengers (i.e., non-incumbents). It looks like with a larger sample size, equality could be rejected over an even larger range. The graph also gives a sense of the magnitude of the incumbency effect.

5 Methods and formulas

This section contains additional theoretical details from Goldman and Kaplan (2018). It may provide a deeper understanding for some readers, although it may also be skipped

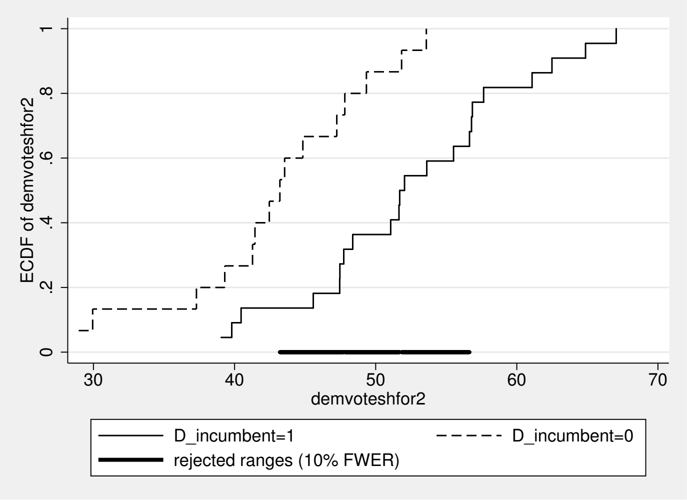

Figure 4: Empirical CDFs from regression discontinuity design.

without hindering successful application of `distcomp` in practice.

Notationally, let $F_X(\cdot)$ be the population CDF for the first group, and let $F_Y(\cdot)$ be the population CDF for the second group. The GOF null hypothesis is $H_0\colon F_X(\cdot) = F_Y(\cdot)$, i.e., $F_X(r) = F_Y(r)$ for all real numbers r. The `distcomp` command also considers the individual hypotheses $H_{0r}\colon F_X(r) = F_Y(r)$.

For more formal discussion, the following notation and definitions are helpful. These are adapted from the section on multiple testing by Lehmann and Romano (2005, §9.1). For the family of null hypotheses H_{0r} indexed by real numbers r, let $\mathcal{T} \equiv \{r : H_{0r} \text{ is true}\}$, the set of values of r for which hypothesis H_{0r} is true. The "familywise error rate" (FWER) is the probability of falsely rejecting at least one true hypothesis:

$$\text{FWER} \equiv \Pr\left(\text{reject any } H_{0r} \text{ with } r \in \mathcal{T}\right). \tag{6}$$

"Weak control" of FWER at level α requires FWER $\leq \alpha$ if each H_{0r} is true, i.e., if the GOF null hypothesis $H_0\colon F_X(\cdot) = F_Y(\cdot)$ is true, but it allows FWER $> \alpha$ if some H_{0r} are false. In this setting, weak control of FWER is equivalent to size control for the corresponding GOF test that rejects $H_0\colon F_X(\cdot) = F_Y(\cdot)$ when at least one H_{0r} is rejected. "Strong control" of FWER requires FWER $\leq \alpha$ for any $\mathcal{T}$, i.e., for any two CDFs $F_X(\cdot)$ and $F_Y(\cdot)$. Strong control implies weak control, but not vice-versa.

Strong control of FWER, even in small samples, is achieved by `distcomp`. The "rejected ranges" displayed by `distcomp` are the values of r for which H_{0r} is rejected when strongly controlling the FWER at the specified level α.

The two most important properties of `distcomp` are its strong control of finite-sample FWER and its improved sensitivity to tail differences compared to `ksmirnov`. The procedure is now described mathematically, and the achievement of these two properties is then discussed further.

Steps for computation of `distcomp` are given in Method 5 of Goldman and Kaplan (2018). The idea is to compute a uniform confidence band (detailed below) for each unknown CDF, and then reject H_{0r} for any r where the bands do not overlap. Notationally, let $B_{k,n}^{\tilde{\alpha}}$ denote the $\tilde{\alpha}$-quantile of the $\mathrm{Beta}(k, n+1-k)$ distribution, defining $B_{0,n}^{\tilde{\alpha}} = 0$ and $B_{n+1,n}^{\tilde{\alpha}} = 1$ for any $\tilde{\alpha}$, and denote sample sizes as n_X and n_Y. The uniform confidence bands for $F_X(\cdot)$ and $F_Y(\cdot)$ are respectively $[\widehat{\ell}_X(\cdot), \widehat{u}_X(\cdot)]$ and $[\widehat{\ell}_Y(\cdot), \widehat{u}_Y(\cdot)]$, where for some $\tilde{\alpha}$,

$$\widehat{\ell}_X(r) = B_{n_X \widehat{F}_X(r), n_X}^{\tilde{\alpha}}, \quad \widehat{u}_X(r) = B_{n_X \widehat{F}_X(r)+1, n_X}^{1-\tilde{\alpha}}, \tag{7}$$

and similarly replacing X with Y. Then, H_{0r} is rejected when either $\widehat{\ell}_X(r) > \widehat{u}_Y(r)$ or $\widehat{\ell}_Y(r) > \widehat{u}_X(r)$. The value of $\tilde{\alpha}$ is the largest value such that the probability of rejecting any H_{0r} (i.e., the FWER) does not exceed α when $X_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$, $i = 1, \ldots, n_X$, and $Y_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$, $i = 1, \ldots, n_Y$, which is determined by simulation.

Although $\tilde{\alpha}$ is chosen to guarantee FWER control only when both $F_X(\cdot)$ and $F_Y(\cdot)$ are standard uniform CDFs, this extends to any $F_X(\cdot) = F_Y(\cdot)$. The key insight is that at a given r, after determining $\tilde{\alpha}$ (and n_X and n_Y), rejection of H_{0r} depends only on $\widehat{F}_X(r)$ and $\widehat{F}_Y(r)$, which are step function that only increase at observed sample values. That is, rejection of H_{0r} only depends on the number of X_i below r and the number of Y_i below r, which yield $\widehat{F}_X(r)$ and $\widehat{F}_Y(r)$ when divided by n_X and n_Y, respectively. Consequently, whether or not any H_{0r} is rejected depends only on the relative order of observed X_i and Y_i values, not on the values themselves. This implies that applying any monotonic transformation to the data does not affect the FWER. When $F_X(\cdot) = F_Y(\cdot) = F(\cdot)$, we could sample X_i and Y_i from F by first drawing standard uniform random variables and then applying $F^{-1}(\cdot)$. But since $F^{-1}(\cdot)$ is a monotonic transformation, it will not affect FWER, so any $F(\cdot)$ will produce identical FWER as when X_i and Y_i are simply standard uniform themselves.

The above argument only concerns weak control of FWER; the extension to strong control is not obvious. Indeed, one of the contributions of Goldman and Kaplan (2018) is their Lemma 2 that proves weak control implies strong control for any procedure where rejection of H_{0r} depends only on $\widehat{F}_X(r)$ and $\widehat{F}_Y(r)$, which is the case here. The intuition is that if FWER is controlled when $F_X(r) = F_Y(r)$ for all r, then changing $F_X(\cdot)$ so that $F_X(r) \neq F_Y(r)$ at some r does not somehow increase the probability of rejecting the remaining H_{0r} where $F_X(r) = F_Y(r)$.

Computationally, the difficult part of the procedure is determining $\tilde{\alpha}$. Since $\tilde{\alpha}$ depends only on α, n_X, and n_Y, a large table of precise values was simulated ahead of time for the most common levels of $\alpha = 0.01, 0.05, 0.10$, and included in `distcomp`. This enables nearly instantaneous computation of `distcomp`.

For the second important property of improved tail sensitivity, it is insightful to look at the uniform confidence bands more closely. Here, we look at a single sample of X_i with CDF $F(\cdot)$. A "uniform confidence band" for $F(\cdot)$ consists of an upper function

$\widehat{u}(\cdot)$ and lower function $\widehat{\ell}(\cdot)$ that may depend on the data and satisfy $\Pr(\widehat{\ell}(\cdot) \leq F(\cdot) \leq \widehat{u}(\cdot)) \geq 1-\alpha$ for confidence level $(1-\alpha) \times 100\%$, where $\widehat{\ell}(\cdot) \leq F(\cdot) \leq \widehat{u}(\cdot)$ means $\widehat{\ell}(r) \leq F(r) \leq \widehat{u}(r)$ for all r. Such a band may be constructed by inverting the one-sample Kolmogorov–Smirnov test, but its pointwise coverage probability varies greatly with r. That is, $\Pr(\widehat{\ell}(r) \leq F(r) \leq \widehat{u}(r))$ is much larger (closer to 100%) when r is in the tails of the true distribution (i.e., when $F(r)$ is nearer zero or one) than when r is in the middle (i.e., $F(r)$ nearer 0.5). In contrast, the pointwise coverage probability of the uniform confidence band used in `distcomp` is (nearly) the same for all values of r.

The even pointwise coverage probability property of the uniform confidence bands used by `distcomp` can be seen as follows. Similar points are made by Buja and Rolke (2006, top p. 28). Let $X_{n:k}$ denote the kth order statistic in a sample of size n, i.e., the kth smallest value in the sample, so $X_{n:1} < \cdots < X_{n:k} < \cdots < X_{n:n}$. From Wilks (1962), $F(X_{n:k}) \sim \text{Beta}(k, n+1-k)$ if $F(\cdot)$ is continuous. This follows from an application of the "probability integral transform": $F(X_i) \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, and $F(X_{n:k})$ follows the same distribution as the kth order statistic from a sample of n standard uniform random variables, which is $\text{Beta}(k, n+1-k)$. Thus, since $B^{\tilde{\alpha}}_{k,n}$ is defined as the $\tilde{\alpha}$-quantile of that same distribution, $\Pr(B^{\tilde{\alpha}}_{k,n} \leq F(X_{n:k}) \leq B^{1-\tilde{\alpha}}_{k,n}) = 1-2\tilde{\alpha}$ exactly, for any k and n, irrespective of $F(\cdot)$. In the earlier expressions, $[\widehat{\ell}(\cdot), \widehat{u}(\cdot)]$ is essentially taking pointwise intervals $[\widehat{\ell}(X_{n:k}), \widehat{u}(X_{n:k})] = [B^{\tilde{\alpha}}_{k,n}, B^{1-\tilde{\alpha}}_{k,n}]$ and connecting them with a stair-step interpolation. This implies pointwise coverage probability of $1-2\tilde{\alpha}$ at every order statistic (and only somewhat larger at other points).

The $1-2\tilde{\alpha}$ probability at each order statistic contrasts the Kolmogorov–Smirnov pointwise coverage probability that is much higher in the tails. This difference translates directly to the ability to detect deviations across different values: the Kolmogorov–Smirnov sensitivity/power is concentrated in the center of the distribution, whereas `distcomp` spreads its power evenly across the whole distribution. Put differently, Kolmogorov–Smirnov implicitly uses a much larger pointwise statistical significance level for testing H_{0r} near the center of the distribution and much smaller significance level in the tails, whereas `distcomp` uses approximately the same level of statistical significance for all H_{0r}.

6 Conclusion

In addition to a two-sample goodness-of-fit test that improves upon the Kolmogorov–Smirnov test, the `distcomp` command provides a detailed, point-by-point assessment of statistically significant differences between two distributions. This is much more informative than existing goodness-of-fit tests (like the `ksmirnov` command) or t-tests for mean equality (like `ttest`) while still controlling the false positive rate, with strong control of the familywise error rate. Potential applications abound, such as descriptions of how a variable's distribution changes over time or differs between groups (geographic, socioeconomic, etc.), regression discontinuity designs, and perhaps especially in program evaluation.

7 Acknowledgements

There would be no Stata command without the work of Matt Goldman (Facebook Research) on Goldman and Kaplan (2018). Thanks to Gneezy and List (2006) and Cattaneo, Frandsen, and Titiunik (2015) for making their data public. Thanks to Colin Cameron for first encouraging me (in 2013) to write Stata commands.

8 References

Atkinson, A. B. 1987. On the Measurement of Poverty. *Econometrica* 55(4): 749–764. https://doi.org/10.2307/1911028.

Buja, A., and W. Rolke. 2006. Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data. Working paper, available at http://stat.wharton.upenn.edu/ buja/PAPERS/paper-sim.pdf.

Cattaneo, M. D., B. R. Frandsen, and R. Titiunik. 2015. Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference* 3(1): 1–24. https://doi.org/10.1515/jci-2013-0010.

Eicker, F. 1979. The asymptotic distribution of the suprema of the standardized empirical processes. *Annals of Statistics* 7(1): 116–138. https://www.jstor.org/stable/2958837.

Gneezy, U., and J. A. List. 2006. Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments. *Econometrica* 74(5): 1365–1384. https://doi.org/10.1111/j.1468-0262.2006.00707.x.

Goldman, M., and D. M. Kaplan. 2018. Comparing distributions by multiple testing across quantiles or CDF values. *Journal of Econometrics* 206(1): 143–166. https://doi.org/10.1016/j.jeconom.2018.04.003.

Lehmann, E. L., and J. P. Romano. 2005. *Testing Statistical Hypotheses*. 3rd ed. Springer Texts in Statistics, Springer. http://books.google.com/books?id=Y7vSVW3ebSwC.

Wilks, S. S. 1962. *Mathematical Statistics*. New York: Wiley.

About the author

David M. Kaplan is an assistant professor in the Department of Economics at the University of Missouri. His primary research interest is econometric methodology. In particular, he enjoys creating and advancing methods for understanding changes and treatment effects on entire distributions (instead of just averages).