The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–13

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. 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. 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)
Comparing distribution of y when grp=0 vs. grp=1

Global test of equality of two CDFs:
    Reject at a 1% level
```

Section 2 describes syntax and usage of `distcomp`. Section 3 provides empirical examples that can be replicated with the provided do-file. Section 4 shows some of the theoretical foundations before concluding. Readers interested in related methods like one-sided or one-sample comparisons, p-values, and uniform confidence bands are referred to Goldman and Kaplan (2018) and the corresponding R code. Abbreviations are used for cumulative distribution function (CDF) and familywise error rate (FWER).

2 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 result displayed by `distcomp` shows whether or not the corresponding goodness-of-fit test rejects at the specified statistical signifiance level, similar to a two-sample Kolmogorov–Smirnov test. That is, the null hypothesis is that the two distributions 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 second result displayed shows ranges of values for which the difference between cumulative distribution functions (CDFs) is statistically significant, accounting for the "multiple testing" nature of the procedure. Instead of a single, goodness-of-fit null hypothesis, there is a group of null hypotheses, where each individual hypothesis specifies equality of the two cumulative distribution functions 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.

Syntax, options, and stored results are now shown.

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

`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%).

Scalars

`r(alpha)`	specified FWER level	`r(alpha_sim)`	simulated FWER level
`r(rej_gof)`	goodness-of-fit rejection		

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; either way, the FWER level is no greater than the specified `alpha` option. When `r(rej_gof)` is zero, then `r(rej_ranges)` is a 1-by-2 matrix with both entries missing, to indicate that no ranges are rejected.

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

3 Examples

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

3.1 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)
. distcomp wage , by(union) alpha(0.10)
Comparing distribution of wage when union=0 vs. union=1

Global test of equality of two CDFs:
    Reject at a 10% level

With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following ranges:
distcomprejranges[1,2]
         from         to
r1  2.3268917  11.610305
. quietly distcomp wage , by(union) alpha(0.01)
. disp "Reject equality at 1% level: `=r(rej_gof)'"
Reject equality at 1% level: 1
```

Three main results are displayed. The first result says that the global, goodness-of-fit null hypothesis that the two wage distributions are identical is rejected at the specified 10% level. (Due to computational complexity, p-values are only implemented in the R version.) In fact, as seen in the third (last) displayed result, this hypothesis is rejected even at a 1% level; 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 (around \$2.33/hr to \$11.61/hr). This suggests (at least) a restricted first-order stochastic dominance relationship, as defined in Condition I of Atkinson (1987, p. 751), as the empirical CDF graph below helps show.

Figure 1 shows the empirical CDFs for union wage and non-union wage. Making such graphs is strongly recommended as a complement to running `distcomp`; code to produce all graphs is provided in `distcomp_examples.do`, though not shown here (to conserve space). It is clear that union wages tend to be higher in the sample (hence the union CDF lies below the non-union CDF). However, graphs alone cannot show what the `distcomp` results above showed, i.e., that this difference is indeed statistically significant at a 10% FWER level across most (though not all) of the distribution.

3.2 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

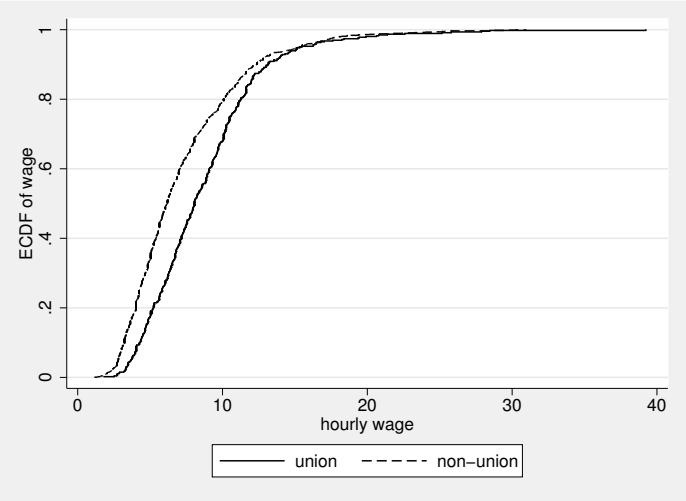

Figure 1: Empirical CDFs from NLSW.

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) // no effect
Comparing distribution of y1 when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Do not reject at a 10% level
. distcomp y2 , by(treated) // effect only above median (zero)
Comparing distribution of y2 when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Reject at a 10% level

With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following ranges:
distcomprejranges[1,2]
         from         to
r1  .93706262  3.0885596
. distcomp y3 , by(treated) // effect largest in tails, 0 at median
Comparing distribution of y3 when treated=0 vs. treated=1

Global test of equality of two CDFs:
    Reject at a 10% level

With strong control of FWER at a 10% level,
CDF equality is rejected at all points in the following ranges:
```

```
distcomprejranges[3,2]
          from         to
r1  -3.8624718  -1.9740542
r2  -1.2792215  -1.1729968
r3   1.5580958   3.2656784
```

Above, for the case where the control and treatment distributions are equal, the `distcomp` goodness-of-fit test does not reject equality. The empirical CDFs differ, but `distcomp` says these differences are not statistically significant at a 10% level.

For the 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. 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 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, `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 sample values). Given the same FWER level, more data would be required to enlarge the ranges where we are statistically confident in a CDF difference.

3.3 Example with experimental data

The following example uses data from Gneezy and List (2006) to test for distributional treatment effects. A longer version (with results from R code) 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 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, we examine heterogeneity in the treatment effect during the first few hours of work, with results seen below.

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

Global test of equality of two CDFs:
    Do not reject at a 5% level
. distcomp yfun , by(treated) a(0.05)
Comparing distribution of yfun when treated=0 vs. treated=1
```

```
Global test of equality of two CDFs:
    Reject at a 5% level

With strong control of FWER at a 5% level,
CDF equality is rejected at all points in the following ranges:
distcomprejranges[1,2]
    from    to
r1     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 has to be 14% before rejecting equality in the range $[56, 58]$, near the upper end of the distribution.

For the fundraising task, even though the sample sizes are again small, the treatment effect is statistically significant at a 5% FWER level (two-sided). Specifically, `distcomp` identifies the range of \$8 to \$14, 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.

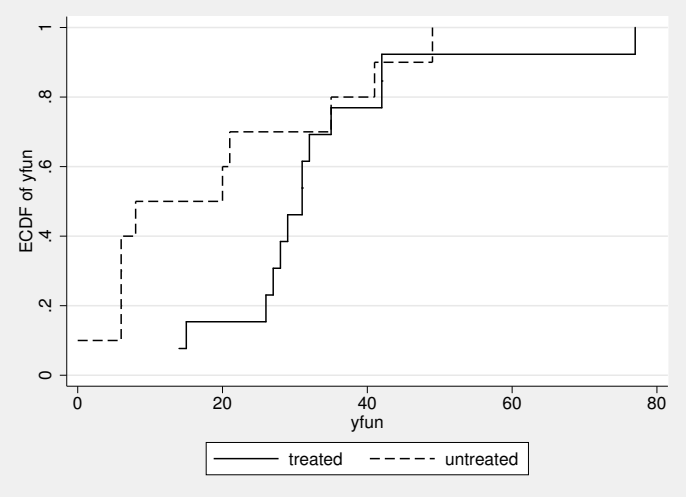

Figure 2: Empirical CDFs from experiment (fundraising).

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

3.4 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

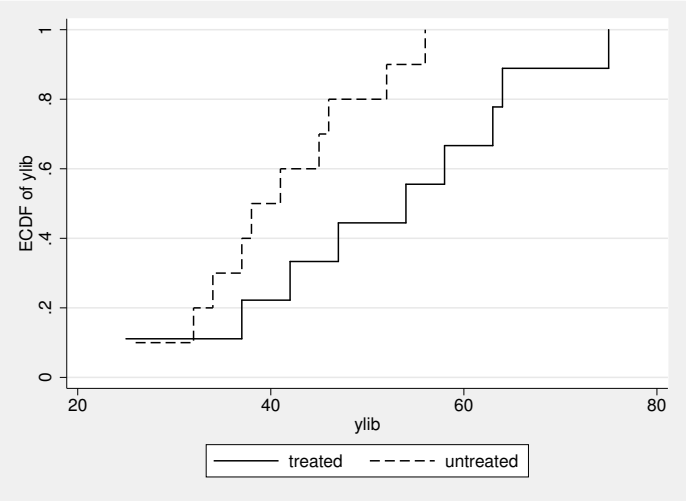

Figure 3: Empirical CDFs from experiment (library data entry).

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.

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 = (demmv>=R0)
. distcomp demvoteshfor2 if demmv>=R0-h & demmv<=R0+h , by(D) a(0.10)
Comparing distribution of demvoteshfor2 when D=0 vs. D=1

Global test of equality of two CDFs:
    Reject at a 10% level

With strong control of FWER at a 10% level,
```

```
CDF equality is rejected at all points in the following ranges:
distcomprejranges[3,2]
         from         to
r1   43.21114  47.738708
r2   47.81345  51.700489
r3  51.836891  56.647652
```

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.

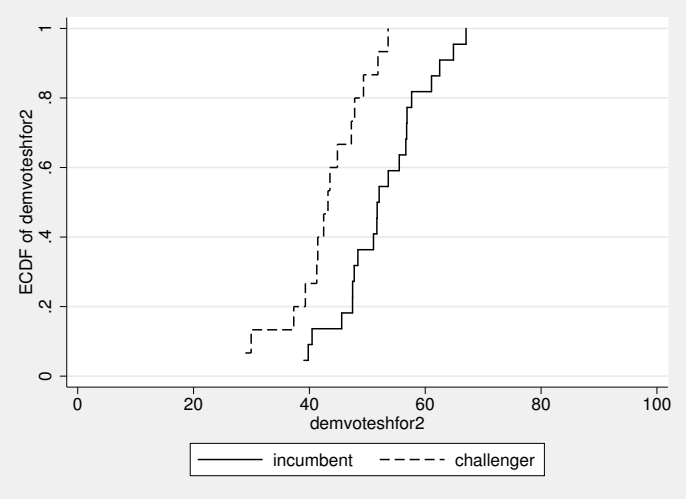

Figure 4: Empirical CDFs from regression discontinuity design.

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.

4 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 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 goodness-of-fit 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}: $F_X(r) = F_Y(r)$.

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:

$$\text{FWER} \equiv \Pr\left(\text{reject any } H_{0r} \text{ with } r \in \mathcal{T}\right). \tag{1}$$

"Weak control" of FWER at level α requires FWER $\leq \alpha$ if each H_{0r} is true, i.e., if the goodness-of-fit null hypothesis H_0: $F_X(\cdot) = F_Y(\cdot)$ 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 goodness-of-fit test that rejects H_0: $F_X(\cdot) = F_Y(\cdot)$ 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 $B^{\tilde{\alpha}}_{k,n}$ denote the $\tilde{\alpha}$-quantile of the Beta$(k, n+1-k)$ distribution, defining $B^{\tilde{\alpha}}_{0,n} = 0$ and $B^{\tilde{\alpha}}_{n+1,n} = 1$ 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 $\tilde{\alpha}$,

$$\widehat{\ell}_X(r) = B^{\tilde{\alpha}}_{n_X \widehat{F}_X(r), n_X}, \quad \widehat{u}_X(r) = B^{1-\tilde{\alpha}}_{n_X \widehat{F}_X(r)+1, n_X}, \tag{2}$$

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 $X_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, $i = 1, \ldots, n_X$, and $Y_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, $i = 1, \ldots, n_Y$, 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 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. 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). This 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}.

5 Conclusion

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.

6 Acknowledgements

There would be no Stata command without the work of Matt Goldman (Microsoft 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.

7 References

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

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* XXX(XXX): XXX–XXX. 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).