Journal of Development Economics 182 (2026) 103815

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular Article

Winsorizing and trimming in RCTs

Till Wicker[1]

Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands

ARTICLE INFO

Dataset link: 10.17632/6ffk227kkp.1

JEL classification:
C18
C21
C81

Keywords:
Winsorizing
Trimming
Biased treatment effect
Type I errors
Type II errors
RCT

ABSTRACT

Winsorizing and trimming are used to minimize the effects of outliers on estimated treatment effects. In Randomized Controlled Trials (RCTs), the typical approach winsorizes/trims the tails of the whole sample, pooling together treatment and control groups. This can have as a consequence that observations from treatment and control groups are disproportionately winsorized/trimmed. An alternative approach, *Stratified* Winsorizing/Trimming, winsorizes treatment groups separately, ensuring that an equal proportion of observations are winsorized/trimmed per experimental arm. A formal framework and Monte Carlo simulations of an RCT illustrate that *Stratified* Winsorizing/Trimming reduces the treatment effect bias and risk of Type II errors compared to the traditional approach, although at the cost of a greater likelihood of Type I errors. Applications to Angelucci et al. (2023) and Jack et al. (2023) illustrate that the chosen winsorizing/trimming technique can affect the magnitude and statistical significance of treatment effects. Practical guidelines for researchers conducting RCTs that want to winsorize/trim outliers are discussed.

1. Introduction

Researchers are concerned with the role of measurement errors and outliers in the estimation of variables and treatment effects. For example, Gollin and Udry (2021) find that measurement errors and productivity shocks explain between half and two-thirds of the variance in productivity among farmers in Uganda and Tanzania. While one literature strand focuses on designing surveys to minimize the occurrence of measurement errors, another strand focuses on dealing with measurement errors — in particular, outliers — once the data is collected.[2] The most common approach to mitigating the role of outliers is to winsorize or trim the tails of the sample distribution. Winsorizing entails "replacing any values bigger than a certain percentile with the value of the data point at that percentile itself", while trimming consists of "replacing the outliers with a missing value" (World Bank, 2023).[3]

Winsorizing and trimming is most commonly observed in Randomized Controlled Trials (RCTs): 50% of papers published in the Top-5 economics journals between 2019–2023 that winsorized or trimmed were RCTs.[4] Authors of RCTs typically winsorize/trim the whole sample, pooling together all experimental arms. However, some recent papers — Benson et al. (2023), Muralidharan et al. (2023), and Bedoya et al. (2023) — winsorize/trim experimental arms separately.

This paper explores the advantages and disadvantages of winsorizing/trimming the whole pooled sample versus separate treatment arms in an RCT (*Pooled* vs. *Stratified* Winsorizing/Trimming). After outlining the two techniques in Section 2, a formal framework in Section 3 derives their effects on estimated treatment effect biases, statistical power, and Type I and II errors. Monte Carlo simulations of an RCT in Section 4 illustrate the effects of both winsorizing/trimming techniques on a study's estimated treatment effect bias and the likelihood of Type I and II errors.[5] The simulations reveal that compared to the pooled approach of winsorizing/trimming the whole sample, *Stratified* Winsorizing/Trimming increases the likelihood of Type I errors, while reducing both the bias on the treatment effect estimate and the

E-mail address: t.n.wicker@tilburguniversity.edu.

[1] I am grateful to Giuseppe Musillo, Anaya Dam, Juan Segnana, Hazal Sezer, Manon Delvaux, Christoph Walsh, Ashley Wong, Daan van Soest, Patricio Dalton, and David McKenzie for their helpful comments and suggestions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[2] For example, the Journal of Development Economics released a Special Issue on Measurement and Survey Design.

[3] Other terminology used includes truncating (both for winsorizing and trimming), and replacing data with empty observations (for trimming).

[4] I define an RCT as a study design where units of observation are randomized into different experimental groups that receive differing interventions. As of February 21st, 2025, 32% of Pre-Analysis Plans Accepted during a Stage 1 Review at the Journal of Development Economics specify that they intend to winsorize or trim the data.

[5] The focus of the formal framework and simulations is on winsorizing, as this is more commonly applied in the academic literature. However, the same intuition and results hold for trimming, see Appendices A and B.

https://doi.org/10.1016/j.jdeveco.2026.103815
Received 7 March 2025; Received in revised form 30 April 2026; Accepted 1 May 2026
Available online 9 May 2026

likelihood of Type II errors. The two approaches to winsorizing/trimming are subsequently applied to two published RCTs (Angelucci et al., 2023; Jack et al., 2023) in Section 5 to illustrate that the chosen winsorizing/trimming method can impact both the magnitude and statistical significance of estimated treatment effects in published papers.[6] Section 6 discusses practical guidelines associated with winsorizing/trimming the pooled sample versus experimental arms separately in an RCT, including Stata and R code, before Section 7 concludes.

By focusing on one of the most common method of dealing with outliers, this paper contributes to the literature on the importance of outliers and measurement errors in the estimation of variables and their relationships. While quantile treatment effects are often used to highlight the heterogeneity of treatment effects across a sample distribution, trimming and winsorizing are used to reduce the effects of outliers. For example, Angrist and Krueger (2000) apply trimming to matched employer-employee data and conclude that "a small amount of trimming could be beneficial" to reduce the effect of outliers. Bollinger and Chandra (2005) illustrate that winsorizing and trimming can result in biased regression estimates, by inducing a sample selection bias: the remaining sample post-winsorizing/trimming is no longer representative of the underlying population (Heckman, 1979; Goldberger, 1981; Heckman, 1990). Crump et al. (2009) formalize this concern in the context of treatment effect estimation, showing that trimming observations based on propensity scores changes the estimand and target population, and derive optimal trimming rules to minimize the asymptotic variance of the resulting estimator.[7] This paper contributes to this literature by identifying an additional potential bias with the *Pooled* approach to winsorizing/trimming the whole sample as a result of the unequal winsorizing/trimming of experimental groups in an RCT, and illustrates the advantages and disadvantages of both winsorizing/trimming techniques on biased estimates of treatment effects, and the likelihood of Type I and II errors.

More recently, Broderick et al. (2023) and Young (2019) have placed renewed emphasis on how outliers and *high leverage* observations can affect average treatment effects. Broderick et al. (2023) show that dropping less than 1% of observations can change the magnitude and sign of estimated treatment effects of published economics papers. Young (2019) illustrates that, across 53 papers published in AEA journals, removing just a single observation results in 35% of treatment effects that were statistically significant at the 1% level to no longer be as statistically significant. This paper contributes to the literature on the sensitivity of treatment effect estimates to outliers by illustrating how the winsorized/trimmed outliers can affect the treatment effect estimate, through empirical applications to Angelucci et al. (2023) and Jack et al. (2023). Across both papers, treatment effect estimates change by 53.21% on average as a result of *Stratified* Winsorizing/Trimming instead of the *Pooled* approach of winsorizing/trimming the whole sample adopted in both published papers. Reporting treatment effects as a result of both winsorizing/trimming techniques can complement the "Approximate Maximum Influence Perturbation" of Broderick et al. (2023) to strengthen the robustness of treatment effect estimates.

Based on the Monte Carlo simulations and applications to Angelucci et al. (2023) and Jack et al. (2023), this paper offers six practical guidelines for researchers conducting RCTs who want to winsorize/trim outliers, further outlined in Section 6:

1. Data collected during different time periods/survey rounds should be winsorized/trimmed separately.

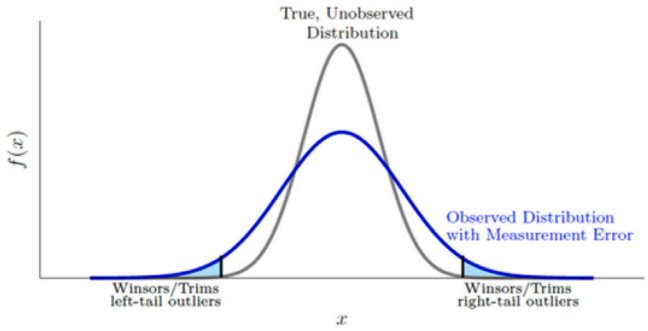


Fig. 1. Winsorizing/Trimming the whole sample.

2. Data collected before observations were randomized into different experimental arms (e.g., baseline survey) should be treated as one sample and hence the recommended technique is *Pooled Winsorizing/Trimming*.
3. For data collected post-randomization, there is no clear winner between winsorizing/trimming the entire sample (*Pooled*) vs. experimental arms separately (*Stratified*). Instead, reporting both techniques provides a more robust estimation of the treatment effect.
4. Reporting the proportion of winsorized/trimmed observations per experimental arm in a paper's appendix can alleviate concerns that observations in certain subgroups are disproportionately winsorized/trimmed.
5. For Pre-Analysis Plans of RCTs, the recommendation is to pre-specify that both approaches to winsorizing/trimming will be used at a pre-specified percentile cut-off, in order to provide further robustness that treatment effect estimates are not driven by outliers.

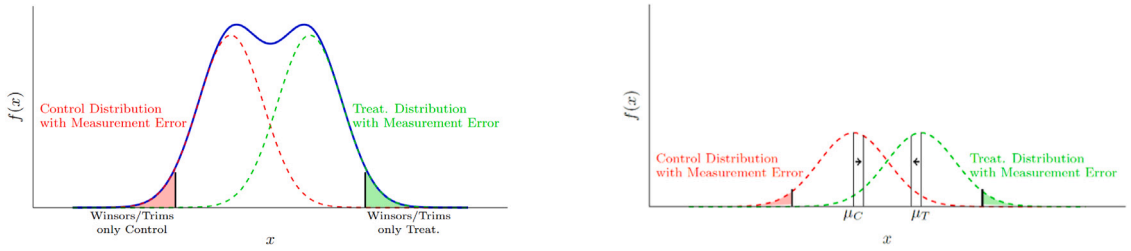
2. Winsorizing and trimming: The intuition

Outliers, particularly in self-reported data, can arise for a variety of reasons: enumerator fatigue, human error, or misreporting, to name a few. Regardless of their reason, outliers can result in the sample distribution differing from the true, unobserved population distribution. Similarly, outliers — in particular *high leverage observations* (Broderick et al., 2023) — can bias treatment effect estimates. Therefore, authors frequently winsorize/trim outliers (the shaded region in Fig. 1) such that the observed sample distribution more closely reflects the true, unobserved population distribution.

The most common approach to winsorizing/trimming is to define an upper and/or lower percentile bound beyond which observations are considered outliers and hence winsorized/trimmed. However, some studies use different criteria for winsorizing/trimming their data, informed by the underlying data generating process. For example, Allcott et al. (2020) winsorize individual's willingness-to-accept to abstain from Facebook at $170, as that was the upper bound of the distribution of Becker–DeGroot–Marschak offers made. de Mel et al. (2019) trim a firm's number of workers at 5, in order to be powered to detect small changes in the outcome variable, due to a long right tail. Fafchamps et al. (2012) trim observations above 10,000 Ghanaian cedi, arguing these are likely due to currency errors. For situations like these, a clear rationale exists to winsorize/trim at a certain value. However, often outcome variables are winsorized/trimmed at the 95th or 99th percentile to account for right-tailed outliers, without an understanding of the data generating process and cause of the outliers. Particularly since the emergence of Pre-Analysis Plans, researchers pre-specify how they will deal with outliers, without understanding the underlying nature of these outliers, and hence rely on rules of thumb.

[6] Appendix D applies both winsorizing techniques to Schilbach (2019) and Augsburg et al. (2015).

[7] Relatedly, Khan and Tamer (2010) and Ma and Wang (2020) show that trimming observations whose propensity scores are close to zero can address problems of instability and poor inferential behavior in inverse-probability-weighted estimators, though at the cost of changing the effective sample population.



(a) Treatment vs. Control (b) Treatment Effect Bias

Fig. 2. Winsorizing/Trimming: by subgroup.

The *Pooled* approach to winsorizing and trimming treats the RCT sample as being drawn from one distribution, even once observations randomized into different experimental arms receive different interventions. If the measurement error is uncorrelated with the experimental group (e.g., the result of an enumerator error, or white noise) — as is typically the case — when authors winsorize/trim, the expectation is that the likelihood of outliers and measurement errors is the same across experimental groups. However, if the subgroups have different distributions — for example due to a non-zero treatment effect — winsorizing or trimming the whole sample can disproportionately winsor/trim the tails of the distribution of each experimental group.

Fig. 2(a) illustrates this in the case of an RCT where the treatment group experiences a positive treatment effect compared to the control group.[8] Winsorizing/trimming the left and right tail of the pooled sample distribution results in winsorizing/trimming the left tail of the control group distribution, and the right tail of the treatment group distribution. If the measurement error is uncorrelated with the experimental arm, the differential winsorizing/trimming of outliers in the treatment and control distributions as a result of the traditional approach to winsorizing/trimming moves the means of both experimental groups inwards, and can generate a biased treatment effect. This is illustrated in Fig. 2(b).

An alternative winsorizing/trimming technique — *Stratified* Winsorizing/Trimming — instead winsorizes/trims each experimental group separately, as illustrated in Fig. 3. By ensuring that an equal proportion of observations are winsorized/trimmed from each treatment arm (and an equal proportion of left- and right-tailed observations per treatment arm), the distribution of each treatment arm more closely reflects the underlying population distribution of these subgroups (see Fig. 3).

The next section formalizes this intuition, before Section 4 illustrates, using Monte Carlo simulations, the effects of winsorizing the whole RCT sample vs. experimental arms separately on treatment effect estimate biases, a study's statistical power (Type II errors), and Type I errors.

3. Formal framework

In this section I present a formal framework and testable propositions for the effects of *Pooled* and *Stratified* winsorizing in an RCT with a treatment and a control group. Appendix B contains derivations of the propositions, along with extensions, and propositions for pooled and stratified trimming.

Consider an RCT with n individuals randomly assigned to a treatment and a control group. $D_i \in \{0,1\}$ denotes the treatment assignment of individual i, where $D_i = 1$ is treatment and $D_i = 0$ is control. Let $Y_i^*(1)$ and $Y_i^*(0)$ denote the *clean latent* potential outcomes, without noise. The estimand of interest is the clean latent average treatment effect:

$$\tau^* \equiv \mathbb{E}\left[Y_i^*(1) - Y_i^*(0)\right] \tag{1}$$

[8] Alternatively, Fig. 2(a) can also illustrate the case of Wave I vs. Wave II of a survey.

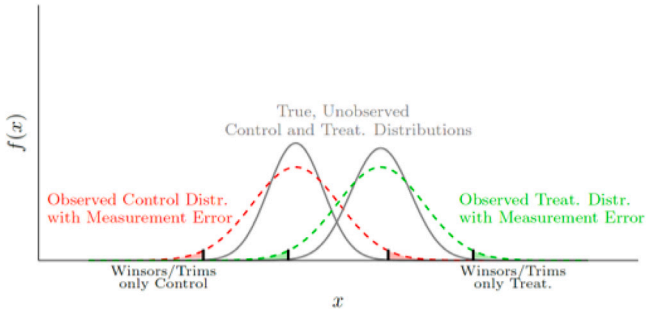


Fig. 3. Stratified Winsorizing/Trimming.

The observed potential outcomes are contaminated by measurement error/noise:

$$Y_i(1) = Y_i^*(1) + U_i(1), \qquad Y_i(0) = Y_i^*(0) + U_i(0) \tag{2}$$

so the realized observed outcome is

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{3}$$

$$= D_i(Y_i^*(1) + U_i(1)) + (1 - D_i)(Y_i^*(0) + U_i(0)) \tag{4}$$

and the observed ATE is $\tau \equiv \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$

Assumption 1 (*RCT Assignment and Mean-Zero Contamination*).

1. $D_i \perp\!\!\!\perp \left(Y_i^*(0), Y_i^*(1), U_i(0), U_i(1)\right)$;
2. $\mathbb{E}[U_i(d)] = 0$ for $d \in \{0,1\}$;
3. The arm-specific distributions F_d of $Y_i(d)$ are continuous with densities f_d that are strictly positive at the quantiles used below, and $\mathbb{E}[Y_i(d)^2] < \infty$.

Assumption 1 offers a baseline set-up for an RCT with two experimental groups, where random assignment is independent of potential outcomes and outliers. This is the baseline case, as researchers typically assume that the likelihood of outliers is uncorrelated with the treatment assignment. Assumption 1 furthermore establishes that the distributions of the outcome variable have well-defined first and second moments, and outliers have on average zero mean. For the baseline specification, we will further assume that individuals are evenly randomized across treatment and control groups, such that $n_0 = n_1 = \frac{n}{2}$. Therefore, the pooled observed distribution of the whole sample is $F_P(y) = \frac{1}{2}F_1(y) + \frac{1}{2}F_0(y)$.

Under Assumption 1, the difference in observed means is unbiased for the latent ATE because:

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i^*(1) - Y_i^*(0)] + \underbrace{\mathbb{E}[U_i(1) - U_i(0)]}_{=0} = \tau^* \tag{5}$$

Next, let us denote the u-quantile for any cumulative distribution function F, by $q_F(u) \equiv \inf\{y \in \mathbb{R} : F(y) \geq u\}$.

In the baseline specification, we explore two-sided symmetric winsorizing, and hence will let $\alpha \in (0, 1/2)$ denote a symmetric two-sided tail fraction in order to define the cutoff values for both *Pooled* and *Stratified* winsorizing. The *Pooled* winsorizing left- and right-tail cutoffs are $a_P = q_{F_P}(\alpha), b_P = q_{F_P}(1-\alpha)$, while the *Stratified* winsorizing left- and right-tail cutoffs are $a_d = q_{F_d}(\alpha), b_d = q_{F_d}(1-\alpha), d \in \{0,1\}$.

I subsequently define the winsorizing map as $w(y; a, b) = a\mathbf{1}\{y < a\} + y\mathbf{1}\{a \le y \le b\} + b\mathbf{1}\{y > b\}$, and for both experimental arms $d \in \{0,1\}$, let $m_d^W(a,b) \equiv \mathbb{E}\big[w(Y_i(d); a, b)\big]$. Hence y is set equal to a if $y < a$ (left-tail winsorizing) and y is set equal to b when $y > b$ (right-tail winsorizing).

Then the population treatment effects after pooled and stratified winsorizing are

$$\tau_W^P \equiv m_1^W(a_P, b_P) - m_0^W(a_P, b_P) \tag{6}$$

$$\tau_W^S \equiv m_1^W(a_1, b_1) - m_0^W(a_0, b_0) \tag{7}$$

with $(a_d^S, b_d^S) = (a_d, b_d)$ and $(a_d^P, b_d^P) = (a_P, b_P)$. P stands for *Pooled* winsorizing and S stands for *Stratified* winsorizing.

The corresponding sample estimators are

$$\hat{\tau}_W^P = \bar{W}_{1,P} - \bar{W}_{0,P} \tag{8}$$

$$\hat{\tau}_W^S = \bar{W}_{1,S} - \bar{W}_{0,S} \tag{9}$$

where $\bar{W}_{d,P}$ is the sample mean in experimental arm d after winsorizing with *Pooled* empirical cutoffs, and $\bar{W}_{d,S}$ is the sample mean in experimental arm d after *Stratifying* winsorizing for each experimental arm-specific empirical cutoffs.

3.1. Bias relative to latent ATE

Proposition 1 (*General Bias Decomposition*). *Let $j \in \{P, S\}$ index pooled or stratified procedures.*

$$\tau_W^j - \tau^* = \Delta_1^W(a_1^j, b_1^j) - \Delta_0^W(a_0^j, b_0^j) \tag{10}$$

where $(a_d^S, b_d^S) = (a_d, b_d)$, $(a_d^P, b_d^P) = (a_P, b_P)$, and

$$\Delta_d^W(a,b) \equiv \mathbb{E}\big[(a - Y_i(d))\mathbf{1}\{Y_i(d) < a\}\big] - \mathbb{E}\big[(Y_i(d) - b)\mathbf{1}\{Y_i(d) > b\}\big] \tag{11}$$

The bias induced by either winsorizing technique can therefore be decomposed into two within-arm distortions: one distortion from winsorizing the treatment arm ($\Delta_1^W(a,b)$) and one distortion from winsorizing the control arm ($\Delta_0^W(a,b)$). *Pooled* and *Stratified* winsorizing differ because they apply different cutoff values to these two experimental arm distributions. In particular, *Pooled* winsorization uses common cutoffs taken from the combined distribution, whereas *Stratified* winsorization lets the cutoffs vary with each arm's own distribution. This difference can be written as:

$$\tau_W^P - \tau_W^S = \big[\Delta_1^W(a_P, b_P) - \Delta_1^W(a_1, b_1)\big] - \big[\Delta_0^W(a_P, b_P) - \Delta_0^W(a_0, b_0)\big] \tag{12}$$

which isolates the extra between-arm distortion, and hence bias, that arises as a result of imposing common pooled cutoffs in *Pooled* winsorization compared to arm-specific cutoffs in *Stratified* winsorization.

3.1.1. A location shift

The cleanest benchmark is that treatment shifts the outcome distribution by a constant amount but does not otherwise change its shape. This mirrors the intuition outlined in Section 2. Formally, this means that treatment acts as a pure additive translation: the treated distribution is the control distribution shifted horizontally by τ^*. Equivalently, every quantile increases by the same amount, while the variance, skewness, kurtosis, and tail shape remain unchanged.

This benchmark is useful because it isolates the role of the winsorization rule itself. If the two arm distributions differ only by a horizontal shift, then any difference between *Pooled* and *Stratified* winsorization must come from the fact that *Pooled* winsorization imposes common cutoffs on two distributions centered at different locations, whereas *Stratified* winsorization allows the cutoffs to shift together with each arm's distribution.

Assumption 2 (*Location-shift Benchmark*). The observed arm-specific distributions satisfy

$$Y_i(1) \overset{d}{=} Y_i(0) + \tau^* \tag{13}$$

A sufficient primitive condition is that $Y_i^*(1) \overset{d}{=} Y_i^*(0) + \tau^*$ and $U_i(1) \overset{d}{=} U_i(0)$.

Proposition 2 (*Stratified Winsorizing Is Unbiased Under a Location Shift*). *Under Assumptions 1 and 2,*

$$\tau_W^S = \tau^* \tag{14}$$

Hence stratified winsorizing is exactly unbiased for the latent ATE in the benchmark location-shift case.

Why does Proposition 2 hold? Under a pure location shift, the treated arm is equal in distribution ($\overset{d}{=}$) to the control arm, shifted by τ^*. The arm-specific quantiles therefore shift by the same amount, so $a_1 = a_0 + \tau^*$ and $b_1 = b_0 + \tau^*$. Stratified winsorization thus clips the same quantile regions in each arm. Equivalently, the winsorization map changes with the additive treatment effect:

$$w(y + \tau^*; a_0 + \tau^*, b_0 + \tau^*) = w(y; a_0, b_0) + \tau^*$$

Taking expectations yields $m_1^W(a_1, b_1) = m_0^W(a_0, b_0) + \tau^*$, and hence

$$\tau_W^S = m_1^W(a_1, b_1) - m_0^W(a_0, b_0) = \tau^*$$

So under a location shift, stratified winsorization preserves the treatment effect exactly because it winsorizes the corresponding parts of the two distributions.

Proposition 3 (*Pooled Winsorizing Attenuates a Pure Location Shift*). *Suppose Assumptions 1 and 2 hold and $\tau^* > 0$. Then:*

1. *the pooled cutoffs lie between the arm-specific cutoffs,*

$$a_0 < a_P < a_1 = a_0 + \tau^*, \qquad b_0 < b_P < b_1 = b_0 + \tau^* \tag{15}$$

2. *the pooled winsorized treatment effect admits the exact representation*

$$\tau_W^P = \int_0^{\tau^*} \mathbb{P}\big(a_P - s < Y_i(0) < b_P - s\big)\, ds \tag{16}$$

3. *consequently,*

$$0 \le \tau_W^P \le \tau^* \tag{17}$$

and the bias is

$$\tau_W^P - \tau^* = -\int_0^{\tau^*} \mathbb{P}\big(Y_i(0) \le a_P - s \text{ or } Y_i(0) \ge b_P - s\big)\, ds \le 0 \tag{18}$$

If $\tau^* < 0$, all inequalities reverse, so pooled winsorizing attenuates the effect toward zero regardless of sign.

Eq. (16) is easiest to read as an accumulation of infinitesimal treatment shifts. The variable s is a dummy integration variable that runs from 0 to the latent treatment effect τ^*. For each intermediate shift s, an untreated observation contributes to the pooled winsorized treatment effect only if, after being shifted by s, it still lies inside the pooled winsorization window. The indicator $a_P - s < Y_i(0) < b_P - s$ captures exactly that event. Integrating over s sums these surviving incremental contributions across the full treatment shift. Eq. (18) then measures the part of the treatment effect that is lost because pooled winsorization clips observations near the common cutoffs. In this sense, pooled winsorization attenuates the treatment effect toward zero by discarding part of the incremental shift in the tails.

Propositions 2 and 3 illustrate the estimated treatment effects for *Stratified* and *Pooled* winsorizing when the treatment induces a shift in the distribution without changing its shape. *Stratified* winsorization does not introduce an additional bias to the estimated treatment effect, while *Pooled* winsorization introduces a bias that attenuates the ATE toward zero. This captures the intuition of Figs. 2 and 3.

3.2. Variance

For tractability, this subsection treats the winsorizing cutoffs as fixed population objects. Appendix B then shows how the formulas change once the cutoffs themselves are estimated from the sample.

Let

$$\sigma^2_{W,d,j} \equiv \mathrm{Var}\big(w(Y_i(d); a_d^j, b_d^j)\big), \qquad d \in \{0,1\},\ j \in \{P,S\} \tag{19}$$

capture the within-arm population variance of the winsorized outcome.

Proposition 4 (*Fixed-cutoff Asymptotic Variance: Winsorizing*)**.** *If the winsorizing cutoffs are treated as fixed population objects, and the treatment share satisfies $n_0 = n_1 = \frac{n}{2}$ under random assignment, then*

$$\sqrt{n}\big(\hat{\tau}^j_W - \tau^j_W\big) \xrightarrow{d} N\big(0, \Omega_{W,j}\big), \qquad \Omega_{W,j} = 2\big(\sigma^2_{W,1,j} + \sigma^2_{W,0,j}\big) \tag{20}$$

Eq. (20) gives the asymptotic distribution of the estimator around the winsorized population estimand (τ^j_W), using the usual difference-in-means variance formula applied to winsorized outcomes. $\Omega_{W,j}$ captures the asymptotic variance parameter. Once the cutoffs are treated as fixed, $\hat{\tau}^j_W = \bar{W}_{1,j} - \bar{W}_{0,j}$ behaves just like an ordinary treatment-control difference in means, except that the raw outcomes are replaced by winsorized outcomes. In other words, Eq. (20) says that once cutoffs are fixed, winsorized estimators behave like standard difference-in-means estimators with modified within-arm variances. Appendix B derives the expression for $\Omega_{W,j}$ in unbalanced cases (when $n_0 \neq n_1$).

3.3. Statistical power and Type II errors

The fixed-cutoff formulas immediately imply trade-offs with respect to the mean squared error (MSE) of the estimator relative to the latent ATE τ^* and statistical power.

Corollary 1 (*MSE Trade-Off*)**.** *For $j \in \{P,S\}$, define $B_j \equiv \tau^j_W - \tau^*$. Then*

$$\mathrm{MSE}(\hat{\tau}^j_W) = B^2_{W,j} + \frac{\Omega_{W,j}}{n} + o(n^{-1}) \tag{21}$$

where $o(n^{-1})$ denotes a remainder term that is negligible relative to n^{-1} (meaning that $n \cdot o(n^{-1}) \to 0$ as $n \to \infty$). Thus Eq. (21) is a first-order approximation of the MSE, with the leading terms given by squared bias and variance. Under the location-shift benchmark, $B_{W,S} = 0$, so stratified winsorizing is preferred to pooled winsorizing in MSE whenever

$$n\, B^2_{W,P} > \Omega_{W,S} - \Omega_{W,P} \tag{22}$$

Pooled winsorizing is preferred in MSE only if its variance advantage over stratified winsorizing is large enough to offset its squared bias.

Corollary 2 (*Type II Error and Asymptotic Power*)**.** *Consider a two-sided Wald test of level κ based on $\hat{\tau}^j_W$ with a consistent standard error. The large-sample power against a fixed alternative is approximately*

$$\pi_{W,j}(\tau^*) \approx 1 - \Phi\left(z_{1-\kappa/2} - \frac{\sqrt{n}\,|\tau^j_W|}{\sqrt{\Omega_{W,j}}}\right) + \Phi\left(-z_{1-\kappa/2} - \frac{\sqrt{n}\,|\tau^j_W|}{\sqrt{\Omega_{W,j}}}\right) \tag{23}$$

The right-hand side of Eq. (23) is increasing in the signal-to-noise ratio, so the probability of failing to reject the false null (Type II error) is decreasing in the signal-to-noise ratio $\frac{\sqrt{n}\,|\tau^j_W|}{\sqrt{\Omega_{W,j}}}$.

Under the location-shift benchmark, stratified winsorizing has higher asymptotic power than pooled winsorizing whenever

$$\frac{|\tau^*|}{\sqrt{\Omega_S}} > \frac{|\tau^P|}{\sqrt{\Omega_P}} \tag{24}$$

Why does Type II error fall as the signal-to-noise ratio rises? The Wald statistic is approximately normal under the alternative, with a center that moves farther away from zero when the numerator $\sqrt{n}\,|\tau^j_W|$ increases relative to the noise term $\sqrt{\Omega_{W,j}}$. A larger treatment effect, a larger sample size, or a smaller asymptotic variance all make the test statistic more likely to cross the critical threshold. Hence statistical power rises and Type II error falls as the signal-to-noise ratio increases.

Corollaries 1–2 summarize the empirical trade-off. *Stratified* winsorizing is bias-dominant in the clean location-shift benchmark because it preserves the additive treatment effect exactly, whereas *Pooled* winsorization attenuates the treatment effect toward zero. However, *Stratified* winsorizing is not automatically variance-dominant. Variance is governed by the within-arm dispersion of the winsorized outcomes, not by bias. Common pooled cutoffs can sometimes compress both arm distributions more strongly and thus reduce variance, even though they also introduce an attenuation bias to the treatment effect estimate. Hence it is possible for *Stratified* winsorization to dominate on bias while *Pooled* winsorization enjoys a variance advantage.

The preferred method therefore depends on whether the variance reduction from pooling is large enough to compensate for its bias in MSE or statistical power terms. By contrast, as n grows, the non-vanishing attenuation bias from pooled procedures matters more and more for MSE and statistical power, because the variance term shrinks at rate $1/n$ while the bias term does not.

3.4. Type I errors

Up to this point, the winsorization cutoffs a_P, b_P, a_d, b_d have been treated as fixed population quantiles. In practice, however, researchers replace these with empirical quantiles computed from the sample. This introduces an additional source of randomness: the treatment effect estimator now varies not only because the winsorized sample means vary, but also because the cutoff estimates themselves vary from sample to sample. This distinction matters most for Type I error and for comparing *Pooled* versus *Stratified* winsorizing under the null hypothesis.

At the population level, if the null of no treatment effect is true and the arm distributions coincide, $F_1 = F_0 \implies \tau^P_W = \tau^S_W = 0$. Hence, if one treats the cutoffs as fixed and uses the correct variance, both *Pooled* and *Stratified* procedures would have asymptotic size equal to the chosen significance level (the nominal level).

Once the cutoffs are estimated, however, the estimator admits a first-order expansion of the form

$$\hat{\tau}^j_W - \tau^j_W = \underbrace{(\text{fixed-cutoff mean-difference term})}_{\text{sampling variation in winsorized means}} + \underbrace{(\text{cutoff-estimation term})}_{\text{sampling variation in estimated quantiles}} + o_p(n^{-1/2})$$

and hence decomposes the sampling-noise around the winsorized population estimand τ^j_W.

The first term is the same source of variation studied in Proposition 4. The second term appears only because empirical quantiles are random and therefore transmit additional sampling noise into the winsorized estimator. The third term ($o_p(n^{-1/2})$) captures a remainder term that is negligible relative to n^{-1}.

This distinction explains why *Pooled* and *Stratified* winsorizing procedures can have different finite-sample Type I error rates under the null. *Pooled* winsorization estimates common cutoffs from the full sample. When $F_1 = F_0$, the resulting cutoff noise enters the treatment and control means symmetrically, so it cancels the first-order components of the sampling error in the treatment-control contrast. *Stratified* winsorization instead estimates separate cutoffs in the two experimental arms. The cutoff-estimation errors therefore differ across experimental arms and need not cancel, which can inflate finite-sample rejection rates under *Stratified* winsorization even when the null is not rejected. Appendix B formalizes this point using the full asymptotic linear representation.

See Appendix B for proofs and extensions of the benchmark framework outlined in this section. The appendix derives the full large-sample variance when winsorization cutoffs are estimated from the

data, extends the results to unequal assignment shares and to trimming, and considers cases in which treatment changes dispersion as well as location. Taken together, these extensions show that the location-shift case studied here is a useful benchmark: when treatment mainly shifts the distribution, stratified procedures remain attractive because they align the cutoff rule with each arm's own distribution, while pooled procedures attenuate the effect toward zero and are more stable under the null because common cutoff-estimation noise cancels in the treatment-control contrast. Once treatment also changes the shape of the distribution, however, neither pooled nor stratified procedures uniformly dominates, so the comparison becomes more context dependent.

4. Monte Carlo simulations

Monte Carlo simulations replicate an RCT where 500 participants are randomly assigned to a control and a treatment group. The estimated regression is $Y_i = \alpha + \beta_1 T_i + \varepsilon_i$, where T_i is an indicator equal to one if the participant is assigned to the treatment group, and zero otherwise. β_1 therefore is an unbiased estimate of the treatment effect.

The outcome variable Y_i for each of the 500 observations in all RCT simulations consists of a value drawn from a distribution with known mean and standard deviation, and an error term that is standard normally distributed ($\sim N(0,1)$). The error term thus introduces noise (and outliers) in the simulated outcome variable, and these outliers are uncorrelated with assignment to the treatment or control group. Therefore, the simulations reflect the baseline set-up of an RCT outlined in Assumption 1.

The outcome variable Y_i is subsequently winsorized at the 90% level (top and bottom 5%), using the *Pooled* approach of winsorizing the whole sample, as well as *Stratified* Winsorizing separately by treatment arm. The focus for this section, in line with the formal framework in Section 3, is on winsorizing, however Appendix A reproduces simulations for trimming, with qualitatively similar results. Appendix A also reproduces the simulations when the underlying data-generating process reflects other distributions aside from a normal distribution (e.g., Log-Normal, Poisson, Gamma, Uniform) for both winsorizing and trimming, along with simulations where additional outliers are artificially induced, and simulations that vary the skewness of treatment and control distributions — hence simulating data generating processes where treatment effects are small for the majority of observations but have a very large (or negative) treatment effect on a few observations. Results are qualitatively identical.

4.1. Biased treatment effects

The stylistic example of Fig. 3 illustrates how winsorizing the entire sample distribution can differentially winsor observations from experimental groups if their underlying distributions differ. This in turn can bias the treatment effect estimate by attenuating the true treatment effect toward zero. To test this, I run 10,000 simulations of the RCT with 500 subjects divided across a treatment and control group. The outcome variable is normally distributed and has a normally distributed error term ($\sim N(0,1)$). In Figs. 4(a) and 4(b), the normal distributions of the outcome variable of the treatment and control groups are characterized by a mean that is uniformly, randomly drawn from [0,0.5] ([0,2] for Fig. 4(b)), with a standard deviation of 1.

Each simulation generates a treatment effect estimate (β_1) with the two approaches to winsorizing. The resulting deviation is measured as the difference between the winsorized treatment effect and the non-winsorized treatment effect, normalized by the standard deviation of the control group of the non-winsorized sample. The non-winsorized treatment effect estimate is an unbiased estimate of the latent average treatment effect (as the simulations satisfy Assumption 1), and hence the deviation can also be interpreted as a treatment effect bias in expectation.

Results are presented in Fig. 4. The horizontal white line means there is no treatment effect deviation as a result of winsorizing. Values above the white horizontal line indicate that winsorizing induces a positive deviation on the treatment effect estimate, while values below the horizontal line indicate a negative deviation.

Fig. 4(a) shows that *Stratified Winsorizing* on average results in a smaller treatment deviation compared with the pooled approach to winsorizing for small and moderate treatment effects, ranging from Cohen's d = $[-0.5, 0.5]$ (Cohen, 1988). Fig. 4(b) reproduces Fig. 4(a) for larger treatment effects in the range of Cohen's d = $[-2, 2]$. Both figures reflects the benchmark case outlined in Assumption 2, where the treatment effect induces a location-shift.

While differences between the two approaches to winsorizing are not statistically significantly different (paired t-test), *Stratified Winsorizing* generates a smaller mean deviation, smaller spread, and the deviation does not increase or flip sign with the treatment effect (K-S test, $p < 0.001$). In cases of a positive treatment effect, the *Pooled* approach to winsorizing can underestimate the treatment effect. When the treatment effect is negative, the *Pooled* approach on average underestimates the true negative treatment effect by generating a positive deviation on the treatment effect estimate. Thus, *Pooled* winsorization attenuates the estimated treatment effect toward zero, while *Stratified* winsorization does not. This captures the intuition of Figs. 2 and 3 and Propositions 2 and 3.

Fig. 4(c) shows the effects of 10,000 random draws of the mean and standard deviation of the treatment and control distributions of the outcome variable, with a range $(0, 4)$. Each observation furthermore contains a normally distributed error term ($\sim N(0,1)$) *Stratified* Winsorizing outperforms the *Pooled* approach to winsorizing, with a statistically insignificant smaller mean bias (paired t-test, p = 0.492), but a statistically significantly smaller spread (K-S test, $p < 0.001$). Fig. 4(d) fixes the treatment effect to Cohen's d = 0.5, but varies the share of the sample belonging to the treatment group from 5% to 95%. Compared with the *Pooled* approach to winsorizing, the bias arising from *Stratified Winsorizing* is consistent across the range of sample allocations, and smaller in magnitude. This difference in bias is highly statistically significant (paired t-test and K-S test, $p < 0.001$).

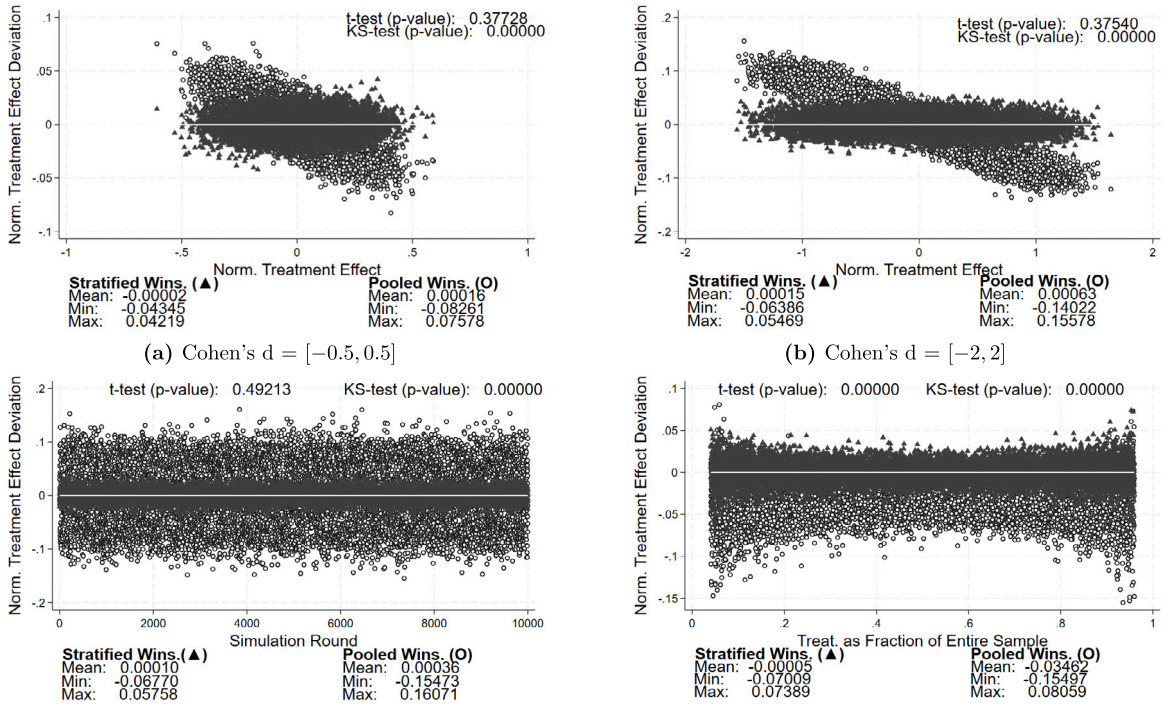
4.1.1. What is driving these results?

To understand the reduced treatment effect bias from *Stratified Winsorizing* compared with the *Pooled* approach of winsorizing the whole sample, emphasis is placed on the observations that are winsorized, and the share of winsorized observations that are from the treatment and control group. *Stratified Winsorizing* ensures that a proportional share of observations are winsorized from the control and treatment groups. The simulations ensure that outliers are uncorrelated with treatment status, and thus the likelihood of an observation being winsorized should be uncorrelated with treatment status too. As treatment and control groups are equally sized in the simulations (aside from Fig. 4(d)), proportional winsorizing would result in 50% of the winsorized observations being from the treatment group.

Fig. 5(a) plots a histogram of the share of winsorized observations that are from the treatment group when using the *Pooled* approach of winsorizing the whole sample. In some simulations, 100% of winsorized observations are from the treatment group, while in other simulations, 0% of winsorized observations are from the treatment group. In only 10.16% of the 10,000 simulations underlying Fig. 4(c) does the *Pooled* approach to winsorizing result in equal proportions of observations from the control and treatment group being winsorized.

Figs. 5(b) and 5(c) plot the fraction of left- and right-tailed observations that are winsorized from the treatment group using the *Pooled* and *Stratified* approaches to winsorizing, as a function of the treatment effect size.[9] *Stratified Winsorizing* ensures that control and treatment

[9] The data is based on the 10,000 simulations underlying Fig. 4(b).



(a) Cohen's d $= [-0.5, 0.5]$ (b) Cohen's d $= [-2, 2]$

(c) Varying Mean and S.D. (d) Varying Treatment Share

Fig. 4. Varying normalized treatment effect.

groups are winsorized proportionately, irrespective of the size of the treatment effect. This results in 50% of winsorized observations being from the treatment group. The *Pooled* approach to winsorizing, on the other hand, winsorizes control and treatment groups disproportionately. When treatment effects are negative, a larger share of left-tailed observations are winsorized from the treatment group, while a smaller share of right-tailed observations are winsorized from the treatment group, compared with the control group. When treatment effects are positive, the effect is reversed, and disproportionately more right-tailed observations are winsorized from the treatment group.

The intuition for these results can be traced back to Fig. 2: the larger the treatment effect, the more right-tailed observations of the treatment group are winsorized when using the *Pooled* approach to winsorizing, and the fewer left-tailed observations of the treatment group are winsorized. The line of best fit of the fraction of winsorized right-tailed observations from the treatment group has a slope of 0.42, implying that a 0.1 standard deviation increase in the treatment effect size results in the percentage of winsorized observations from the treatment group increasing by 4.2%.

Ensuring that a proportional share of observations are winsorized from the control and treatment groups also means that *Stratified Winsorizing* reduces the average distance of the winsorized variable from the nearest non-winsorized variable.[10] This can be explained by comparing Figs. 2 and 3, which illustrate how the two approaches to winsorizing differ. *Stratified Winsorizing* ensures that only values greater than the 95th percentile of each treatment arm's distribution are winsorized.[11] The *Pooled* approach to winsorizing instead can result in values smaller than the 95th percentile of a treatment arm's distribution getting winsorized, which increases the distance between the value of the winsorized and non-winsorized observations.[12] In the simulations underlying Fig. 4(c), *Stratified Winsorizing* reduced the average distance of a winsorized from a non-winsorized observation by 8.03%, compared with the pooled approach to winsorizing. This difference in distance between the two winsorizing techniques is highly statistically significantly ($p < 0.001$, paired t-test and K–S test, see Appendix A.1.4).

4.2. Type II errors and statistical power

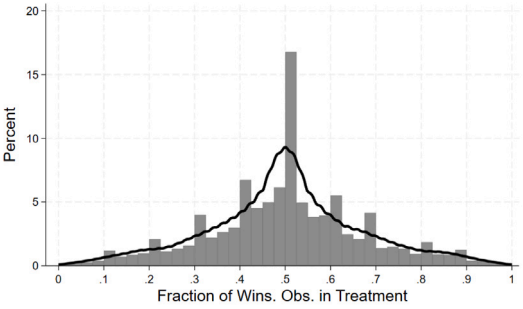
Both approaches to winsorizing can affect the likelihood of Type II errors, and thus a study's statistical power. To simulate this, 1000 iterations are run, each consisting of 1000 simulations of an RCT with 500 observations. In each iteration, the sample size is 500 subjects, equally divided across treatment and control groups. Two-sided t-tests of independent observations are performed, with a significance level of $\alpha = 0.05$. The distribution of the outcome variable of the control group is characterized by a standard normal distribution, while the distribution of the outcome variable of the treatment group is a normal distribution with a standard deviation of 1, but a non-zero mean. Additionally, the outcome variable includes a standard normal error term, which induces outliers. The resulting distributions are winsorized at the 90% level (top and bottom 5%), using both winsorizing techniques.

For each iteration, statistical power is calculated as the percentage of simulations in which the treatment effect is statistically significant. This is performed separately for the whole sample, and the winsorized sample using the *Pooled* and *Stratified* approach. Fig. 6 reports the

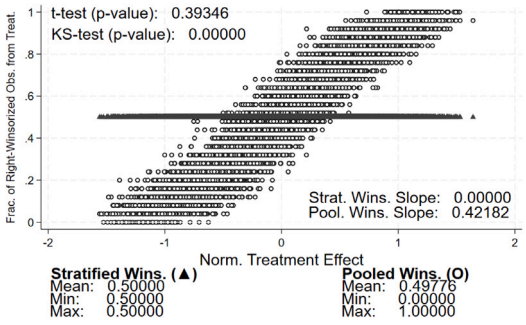
[10] For example, a distribution is winsorized at the 5th and 95th percentile. If an observation at the 99th percentile had an initial value of 10 (which would get winsorized), and an un-winsorized observation at the 95th percentile had a value of 5, the distance in absolute value would be $|10 - 5| = 5$.

[11] The focus here is on the right tail, however the intuition is identical for the left tail (5th percentile).

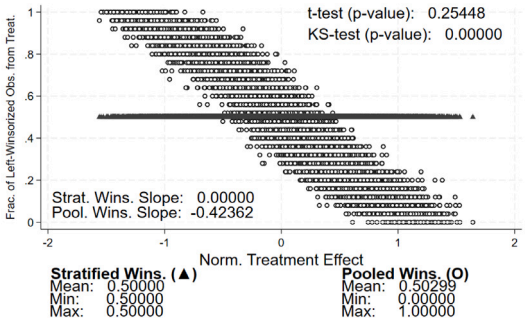
[12] For example, if Fig. 1 simulates winsorizing the sample at the 5th and 95th percentile, then Fig. 2 showcases that the traditional approach to winsorizing would winsorize the 10th percentile and below of the Control group, and the 90th percentile and above of the treatment group. The assumption here is that control and treatment have an equal sample size.



(a) Fraction of Wins/ Observations in Treatment Group, Standard Wins. Approach



(b) Share of Left-Tailed Winsorized Observations from Treatment Group



(c) Share of Right-Tailed Winsorized Observations from Treatment Group

Fig. 5. Monte Carlo simulations: Winsorized observations.

percentage improvements in the study's statistical power as a result of the two approaches to winsorizing, compared with no winsorizing.

For Fig. 6(a), the mean of the treatment effect is a uniformly drawn value between $d = [0, 0.5]$. In Fig. 6(b), the mean treatment effect is Cohen's d=0.2, while the variance of the treatment's normal distribution varies uniformly between 0 and 2. In Fig. 6(c), the mean equals the variance of the treatment group's distribution, and is a value between $(0, 0.5]$. Fig. 6(d) keeps the treatment group's outcome variable distribution fixed ($\sim N(2, 1)$), but varies the sample size of the distribution from 100 to 800 (with the sample being evenly split between treatment and control group).

What is consistent across Fig. 6 is that *Stratified Winsorizing* outperforms the *Pooled* winsorizing technique, in terms of statistical power and hence the likelihood of Type II errors, particularly in simulations with a small treatment effect or small sample size, which typically have lower levels of statistical power.

4.3. Type I errors

The null hypothesis of the simulated RCT regression is that $\beta_1 = 0$, hence that treatment and control groups are from the same underlying distribution. With a significance level $\alpha = 0.05$, the expectation is that Type I errors — where the null hypothesis of no treatment effect is incorrectly rejected — occur in 5% of the cases. A concern with *Stratified Winsorizing* is that Type I errors can emerge with a greater likelihood if the researcher assumes that the sample distribution consists of subgroups, while in fact it does not. In that case, *Stratifying* winsorizing per treatment arm can lead to distortions, and increase the likelihood of Type I errors. This is formalized in Section 3.4 and Appendix B.

Table 1 reports the likelihood with which Type I errors occur. Results are based on 1000 iterations, each consisting of 1000 simulations of the RCT with 500 observations. The control and treatment groups are drawn from the same distribution, and hence the latent treatment effect is zero. Two-sided t-tests of independent observations

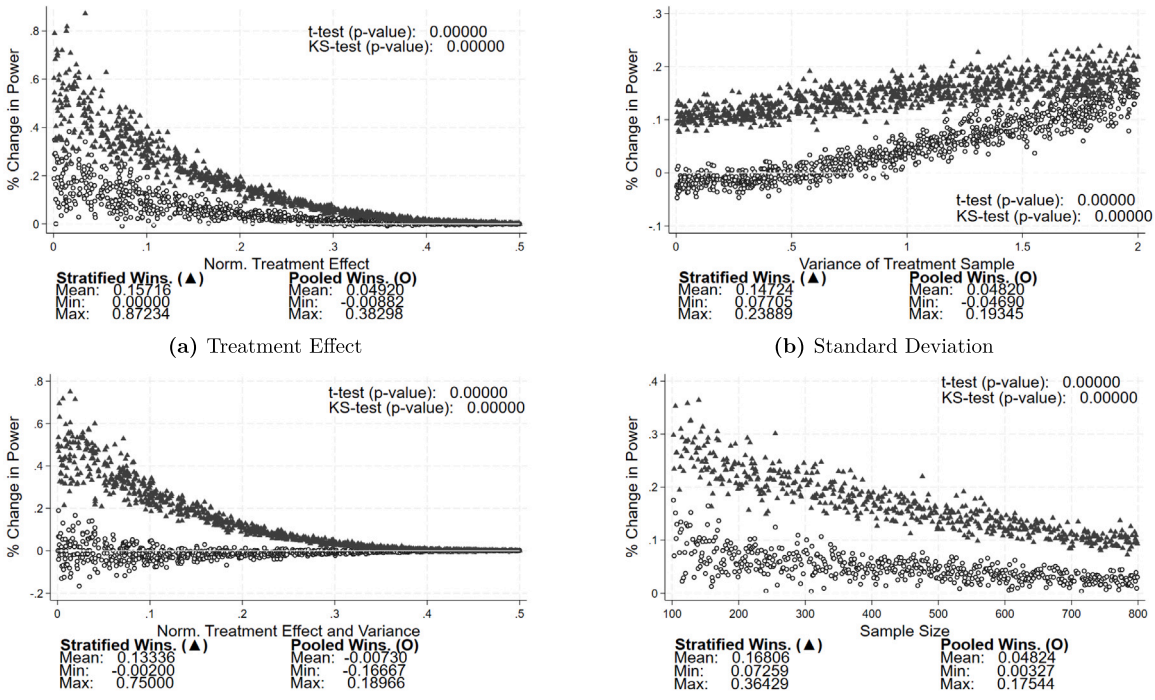
Table 1
Winsorizing and type I errors.

	Normal Distr.	Log-Normal Distr.	Skew-Normal Distr.	Gamma Distr.
A. Frequency of Type I errors				
No Winsor	0.050	0.048	0.050	0.050
Pooled Wins.	0.050	0.050	0.050	0.050
Stratified Wins.	0.075	0.108	0.069	0.081
p-value No vs. Pool.	0.28	0.00	0.68	0.90
p-value No vs. Strat.	0.00	0.00	0.00	0.00
p-value Pool. vs. Strat.	0.00	0.00	0.00	0.00
B. Percentage of *No Winsor* Type I errors included				
Pooled Wins.	85.24	61.90	88.14	80.98
Stratified Wins.	99.18	92.78	99.25	98.37
p-value Pool. vs. Strat.	0.00	0.00	0.00	0.00

are performed to estimate treatment effects, with a significance level of $\alpha = 0.05$. Therefore, Type I errors are expected in 5% of the cases. Simulations are conducted for Normal, Log-Normal, Skew-Normal, and Gamma distributions.

As Table 1, Panel A illustrates, *Stratified Winsorizing* increases the probability of Type I errors in instances where the sample distribution is not composed of subgroups. While the frequency of Type I errors is not statistically significantly different when outliers are not winsorized compared to when the *Pooled* sample is winsorized, *Stratified Winsorizing* results in statistically significantly more cases of Type I errors.

Panel B of Table 1 uncovers an interesting dynamic: while the likelihood of Type I errors is higher when using the *Stratified Winsorizing* technique, the likelihood of a Type I error when there is no winsorizing also being a Type I error when winsorizing is greater using the *Stratified Winsorizing* than the *Pooled* approach of winsorizing. The observation that not all of the same Type I errors are documented when winsorizing vs. not is in line with Bollinger and Chandra (2005), who argue that the remaining sample after winsorizing differs from the sample without



(a) Treatment Effect (b) Standard Deviation

(c) Mean and Standard Deviation (d) Sample Size

Fig. 6. Effects of Winsorizing on statistical power.

winsorizing. This can affect not only the treatment effect estimates (and hence Type II errors) but also the likelihood of Type I errors.

5. Applications to Angelucci et al. (2023) and Jack et al. (2023)

The framework and Monte Carlo simulations demonstrate that both approaches to winsorizing can affect an RCT's estimated treatment effect, and the likelihood of Type I and II errors. In this Section, I illustrate how the two approaches to winsorizing/trimming can affect the statistical significance of treatment effect estimates, using Angelucci et al. (2023) and Jack et al. (2023) as examples. Appendix D performs similar analysis on Schilbach (2019) and Augsburg et al. (2015). These studies were chosen due to their different types of data (administrative vs. self-reported), monetary and non-monetary outcomes, uses of trimming and winsorizing, and the availability of their data and code. The regression tables first replicate the findings of the respective paper in Panel A, before using alternative winsorizing techniques in Panels B and C. Panels D-F illustrate the percentage of observations winsorized/trimmed for each of the experimental arms and survey rounds using the different winsorizing/trimming techniques.

5.1. Angelucci et al. (2023, JDE)

Angelucci et al. (2023) conducted an RCT among women in the Democratic Republic of Congo, randomizing access to a multifaceted program including financial support, training, and social support. The study measured the intervention's impact on various outcomes, immediately after the program ended (endline), and one year later (follow-up). Data was winsorized at the 5th and 95th percentiles.[13] Table 2.(i) reports the OLS-estimated treatment effects for Total Monthly Earnings, Earnings Net of Costs, and Total Business Costs.[14] Panel A replicates the findings of Table 4 in Angelucci et al. (2023) by using the *Pooled* winsorization technique, while Panels B and C present OLS regression results for the *Stratified* Winsorizing per Treatment and *Stratified* Winsorizing per Treatment*TimePeriod approaches, respectively.

Table 2 reports treatment effects of the intervention for both the endline survey (after the end of the intervention), and the follow-up (one year later). Compared with Panel A, *Stratified* Winsorizing by Treatment, and *by Treatment*TimePeriod* (Panels B and C, respectively) result in larger treatment effect estimates, with greater statistical significance. This suggests that the *Pooled* approach to winsorizing attenuates the treatment effect estimates toward zero.

Table 2.(ii) illustrates that this downward bias is driven by an over-winsorizing of right-tailed observations from the treatment group, as it reports the percentage of observations winsorized in the treatment and control groups of Angelucci et al. (2023), as well as the percentage of observations winsorized at endline and the post-endline follow-up. Panel D demonstrates that the *Pooled* approach to winsorizing differentially winsorizes control and treatment observations, with a greater percentage of treated observations being winsorized than observations in the control group. The discrepancy between the percentage of observations winsorized in the control and treatment group is reduced as a result of *Stratified* Winsorizing by Treatment, as shown in Panel E.

However, Table 2.(ii) also illustrates that the *Pooled* winsorizing approach and *Stratified* Winsorizing by Treatment technique differentially winsorize observations from different survey rounds. Both techniques winsorize endline observations and 1-year follow-up observations differentially — although it is unlikely that the measurement error was systematically higher during the endline survey. This is addressed by Panels C and F, which winsorize the data stratified by *Treatment*TimePeriod*, to further ensure that not only are observations from different treatment groups winsorized proportionately, but also across survey rounds.

Compared with Panel A, treatment effects reported in Panel C are larger in magnitude, and statistically more significant. This is driven by the winsorized observations being evenly distributed across treatments, and survey rounds, as shown in Table 2.(ii). Panels C and F highlight

[13] Nevertheless, only right-tailed observations are winsorized. This is because for all three outcome variables, over 50% of observations equaled 0, the lower bound. Hence no winsorizing took place at the left tail.

[14] These outcome variables were chosen, as they were the only ones that were winsorized in the replication package.

Table 2
Angelucci et al. (2023).

Table 2.(i) OLS Treatment effect estimates

	Total Monthly Earnings		Earnings Net of Costs		Total Business Costs	
	Endline (1)	Follow-up (2)	Endline (3)	Follow-up (4)	Endline (5)	Follow-up (6)
A. Pooled Winsorizing (following Angelucci et al., 2023)						
Treatment	0.202*	0.467***	0.0714	0.191**	0.180**	0.321***
	(0.106)	(0.120)	(0.0704)	(0.0773)	(0.0731)	(0.0859)
B. Stratified Winsorizing by Treatment						
Treatment	0.365***	0.585***	0.146**	0.263***	0.429***	0.577***
	(0.114)	(0.118)	(0.0727)	(0.0768)	(0.0771)	(0.103)
C. Stratified Winsorizing by Treatment*TimePeriod						
Treatment	0.309***	0.681***	0.166**	0.249***	0.301***	0.635***
	(0.112)	(0.126)	(0.0699)	(0.0776)	(0.0672)	(0.107)

Table 2.(ii) % of Treat. and Control Obs. Winsorized

	Total Monthly Earnings (1)	Earnings Net of Costs (2)	Total Business Costs (3)
D. Pooled Winsorizing			
% of Endline Control Obs. Winsorized	3.70	5.70	3.70
% of Endline Treatment Obs. Winsorized	5.78	10.11	5.78
% of Follow-up Control Obs. Winsorized	2.20	5.20	2.80
% of Follow-up Treatment Obs. Winsorized	5.68	9.53	5.87
E. Stratified Winsorizing by Treatment			
% of Endline Control Obs. Winsorized	5.50	7.00	5.10
% of Endline Treatment Obs. Winsorized	4.43	9.05	4.14
% of Follow-up Control Obs. Winsorized	3.80	7.30	3.80
% of Follow-up Treatment Obs. Winsorized	4.81	8.28	4.72
F. Stratified Winsorizing by Treatment*TimePeriod			
% of Endline Control Obs. Winsorized	4.00	7.20	4.40
% of Endline Treatment Obs. Winsorized	4.81	9.14	4.72
% of Follow-up Control Obs. Winsorized	3.80	7.10	4.20
% of Follow-up Treatment Obs. Winsorized	4.33	8.28	4.33

Notes: The Table reproduces and extends Table 4 of Angelucci et al. (2023). Standard errors are in parentheses, and clustered at the level of the treatment group. Stratified Winsorizing by Treatment winsorizes the sample separately for treatment and control, while Traditional Winsorizing winsorizes the entire sample. Stratified Winsorizing by Treatment*TimePeriod winsorizes the sample separately for treatment and control observations at endline and follow-up separately. Results are reported without corrections for multiple hypothesis testing. Variables are winsorized at the 5th and 95th percentiles. Consumption refers to the previous week. Business costs include the discounted use value of large purchases. * p<0.1, ** p<0.05, *** p<0.01.

the importance of not only stratifying winsorizing by treatment, but also by the survey round when there are multiple rounds of data collection post-randomization.

5.2. Jack et al. (2023, ReStud)

Jack et al. (2023) conducted an RCT among Kenyan farmers and offered four different loan offers to purchase a water harvesting tank, with varying degrees of asset collateralization. To measure the intervention's impact on milk sales based on administrative data, the researchers use a ITT difference-in-differences approach, and trim the data at the 1, 5, and 10% level (only the right tail) to account for outliers.

Table 3.(i), Panel A reproduces Table 6 of Jack et al. (2023) by reporting treatment effects using the trimming approach adopted in the paper. The researchers do not trim pre- vs. post-randomization observations separately, but instead trim the pooled observations across both time periods. In Panels B and C instead, I trim pre- vs. post-randomization observations separately. In both panels, pre-randomization observations are not trimmed separately by experimental arm, and instead pooled. This is in line with recommendation #2 from Section 1 which argues pre-randomization observations should not be stratified winsorized/trimmed.

Panel B reports treatment effects using the *Pooled Trimming* technique for post-randomization observations, and Panel C reports treatment effects using the *Stratified* Trimming by Treatment technique for post-randomization observations.

Table 3.(ii), Panels D-F demonstrate the importance of winsorizing/trimming separately by survey round. Panel D illustrates that about half as many pre-randomization observations are trimmed compared to post-randomization observations. This is corrected for in Panels E–F, when pre-randomization observations are trimmed separately from post-randomization observations. This can have consequences for the magnitude of treatment effects and their statistical significance, as illustrated by looking at Columns 1 and 3 across Panels A and B in Table 3.(i). This is particularly noticeable for Column 3, where the estimated treatment effect decreases by 26% and is not longer statistically significant at the 5% level.

Comparing Panels B and C in Table 3.(i) illustrates how the chosen trimming technique for post-randomization observations can further influence treatment effect estimates and their statistical significance. Panel C reports larger and more statistically significant treatment effects than Panel B, suggesting that the *Pooled* approach to trimming can have a downward bias on the treatment effect estimate.

In line with the intervention having a positive treatment effect and the authors only trimming the right-hand tail, Panel E illustrates that a disproportionately larger share of post-randomization treatment group observations get trimmed using the *Pooled* approach to trimming. Panel F shows that this is overcome using the *Stratified* Trimming by Treatment technique.

6. Practical guidelines

The formal framework, Monte Carlo simulations, and applications to Angelucci et al. (2023) and Jack et al. (2023) have illustrated

Table 3
Jack et al. (2023).

Table 3.(i) OLS treatment effect estimates

	(1) Milk Sales 1% trim	(2) Milk Sales 5% trim	(3) Milk Sales 10% trim
A. Trimming following Jack et al. (2023)			
Treat*Post	12.580*	12.749**	9.790**
	[6.419]	[5.106]	[4.389]
Treatment	−3.568	−5.960	−6.161
	[5.804]	[4.691]	[3.914]
B. Pooled Trimming of Post-Rand. Observations			
Treat*Post	13.059**	12.962**	7.214*
	[6.544]	[5.126]	[4.346]
Treatment	−3.681	−5.927	−3.327
	[5.276]	[3.976]	[3.014]
C. Stratified Trimming by Treatment of Post-Rand. Observations			
Treat*Post	15.529**	15.051***	10.833***
	[6.412]	[5.073]	[4.259]
Treatment	−3.670	−5.926	−3.324
	[5.279]	[3.976]	[3.015]

Table 3.(ii) % of Treat. and Control Obs. Trimmed

	(1) 1% trim	(2) 5% trim	(3) 10% trim
D. Trimming a la Jack et al. (2023)			
% of Pre-Rand. Control Obs. Trimmed	0.46	2.07	4.92
% of Pre-Rand. Treatment Obs. Trimmed	0.52	2.47	5.21
% of Post-Rand. Control Obs. Trimmed	0.86	5.16	10.14
% of Post-Rand. Treatment Obs. Trimmed	1.12	5.65	11.22
E. Pooled Trimming of Post-Rand. Observations			
% of Pre-Rand. Control Obs. Trimmed	0.94	4.72	10.54
% of Pre-Rand. Treatment Obs. Trimmed	1.01	5.07	9.81
% of Post-Rand. Control Obs. Trimmed	0.79	4.64	9.20
% of Post-Rand. Treatment Obs. Trimmed	1.11	5.10	10.22
F. Stratified Trimming by Treatment of Post-Rand. Observations			
% of Pre-Rand. Control Obs. Trimmed	0.94	4.72	10.54
% of Pre-Rand. Treatment Obs. Trimmed	1.01	5.07	9.81
% of Post-Rand. Control Obs. Trimmed	1.00	4.99	9.99
% of Post-Rand. Treatment Obs. Trimmed	1.00	4.99	9.99

Notes: The Table reproduces and extends Table 6 of Jack et al. (2023). The Post dummy refers to all months from June 2010 (the median loan offer date) onwards. Milk sales are reported in liters. A 1% trim means the top percentile of observations have been trimmed; similarly for the 5% and 10% trims. Standard errors clustered at household level are reported in brackets. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01.

that the decision of how to winsorize/trim observations to reduce the role of outliers in an RCT is less innocuous than it initially seems and can have effects on the treatment effect estimates. The formal framework and Monte Carlo simulations further show that the chosen winsorizing/trimming technique can affect the likelihood of Type I and II errors. This Section therefore discusses practical guidelines when considering whether and how to winsorize/trim outliers.

6.1. When to use which technique

The underlying data generating process should inform the decision of whether and how to winsorize/trim. Regarding the first decision of whether to winsorize/trim, if outliers persist across correlated outcome variables, it is unlikely these outliers are due to repeated measurement errors, and more likely represent a large treatment effect for a few observations. When treatment effects are driven by these sorts of outliers that are not due to measurement errors — for example the large effects of microcredit among the upper tails across seven studies reported by Meager (2022) — winsorizing these outliers will bias the true treatment effect. In these cases, complementing average treatment effect estimates with quantile regressions can highlight the overall effect of the intervention as well as its heterogeneity.

The second decision is how to winsorize/trim. In cases where a value beyond/below a certain value can easily be identified as outliers (e.g., the upper bound of the WTA measure of Allcott et al., 2020), authors should consider those observations outliers and winsorize/trim them accordingly. However, the majority of RCTs set arbitrary percentile thresholds (e.g., 99th or 95th percentile). In these cases, the decision of whether to winsorize/trim the whole pooled RCT sample or separately per experimental arm can affect the estimated treatment effect, and likelihood of Type I and II errors.

Both winsorizing/trimming techniques have their advantages and disadvantages in RCTs, as the formal framework and Monte Carlo simulations illustrated. While *Stratified* Winsorizing/Trimming can improve a study's statistical power and reduce the bias of treatment effect estimates, it can increase the likelihood of Type I errors compared with the *Pooled* approach of winsorizing/trimming the whole sample when the underlying distribution is drawn from the same sample. **With Randomized Controlled Trials, there is no clear winner. Instead, reporting both techniques can provide a more robust estimation of the treatment effect, while minimizing the effects of Type I and II errors.** This is because the underlying null hypothesis of RCTs is that treatment and control groups are drawn from the same distribution. Reporting treatment effects using both winsorizing/trimming techniques can strengthen the robustness of the treatment effect by illustrating that outliers are not driving the treatment effects, in line with the insights of Young (2019) and Broderick et al. (2023).

When treatment effects differ substantially as a result of the winsorizing/trimming technique used, it is important to understand why. For this, an understanding of the underlying data generating process is crucial: if differential winsorizing/trimming of experimental arms is observed during pooled winsorizing/trimming (like in Panel D of

the applications to Angelucci et al. (2023) and Jack et al. (2023)), a justification is needed. Without a clear rationale why observations from experimental arms are disproportionately winsorized/trimmed, *Stratified* Winsorizing/Trimming is likely to report treatment effect estimates closer to the true treatment effect by ensuring that equal proportions of observations are winsorized/trimmed across the subgroups.

In cases where *Stratified* Winsorizing/Trimming results in statistically significant estimates but the *Pooled* approach to winsorizing/trimming does not, authors need to be careful that the statistically significant treatment effect estimates as a result of *Stratified* Winsorizing/Trimming are not due to an increased likelihood in Type I errors. The formal framework and Monte Carlo simulations illustrated that Type I errors are more likely as a result of the *Stratified* Winsorizing/Trimming technique. In such cases, the recommendation is for the authors to report treatment effect estimates following the *Pooled* approach to winsorizing/trimming, in order to minimize the risks associated with Type I errors. Only if authors can justify why treatment effect estimates of *Stratified* Winsorizing/Trimming are more likely to be reliable (e.g., differential winsorizing/trimming of subgroups using the *Pooled* approach to winsorizing/trimming although there is no clear reason why), should they be reported as main results.

6.1.1. How to define relevant subgroups in RCTs

Data collected during different time periods/survey rounds should be treated as separate subgroups, and hence winsorized/trimmed separately. As the examples of Angelucci et al. (2023) and Jack et al. (2023) illustrates, observations of certain time periods are more likely to be winsorized/trimmed when winsorizing/trimming is not done separately per time period, despite no clear rationale existing for why outliers are more common in certain time periods. As such, it is important to winsorize/trim observations from each time period or survey wave separately.

The intuition of Section 2 and the formal framework of Section 3 suggest that subgroup choice should follow the source of potential cutoff misalignment. In an RCT, the relevant post-randomization subgroups are the experimental arms, because the estimand of interest is the treatment-control contrast. Instead stratifying on baseline covariates does not generally line up with the estimand for the main ATE (e.g., gender heterogeneity) and can introduce additional cutoff-estimation noise and hence more biased estimand. Section 3.4 further shows that unnecessary stratification can worsen finite-sample Type I errors.

Accordingly, **for post-randomization outcomes in an RCT, stratified winsorizing/trimming should by default be implemented by experimental arm and, when outcomes are measured in multiple post-randomization rounds, by experimental arm × time period.** Baseline covariates should define winsorizing/trimming subgroups only when the estimand itself is subgroup-specific, or when there is a clear substantive reason to believe that the outlier-generating or measurement process differs along that dimension.

6.2. Pre-analysis plans

While the data generating process should inform the decision of how to deal with outliers, the rise of Pre-Analysis Plans means that authors have to announce their strategy for dealing with outliers before understanding the underlying data generating process. Of all the Stage 1 accepted Pre-Analysis Plans at the Journal of Development Economics that indicated their intention to winsorize/trim their data, all bar one winsorize/trim their data at either the 95th or 99th percentile.[15] For future Pre-Analysis Plans of RCTs, a recommendation is to **pre-specify that both approaches to winsorizing/trimming will be used at a pre-specified percentile cut-off**, in order to provide further robustness that treatment effect estimates are not driven by outliers.

For papers without Pre-Analysis Plans, a documentation of how outliers are handled, including which winsorizing/trimming threshold and technique are chosen, in the paper's appendix will increase the transparency surrounding data cleaning and analysis. In addition to this documentation, **reporting the proportion of winsorized/trimmed observations per subgroup — like Tables 2.(ii) and 3.(ii) — illustrates whether experimental arms are disproportionately affected.** If both winsorizing/trimming approaches are used, reporting how the proportion of winsorized/trimmed observations per experimental arm differs by winsorizing/trimming approach can help rationalize differences in observed treatment effects.

6.3. Winsorizing/trimming covariates

This paper focuses only on cases when outliers of the dependent variable are winsorized/trimmed. However, researchers often also winsorize covariates, such as the baseline value of the outcome variable in an ANCOVA regression (McKenzie, 2012). As covariates in RCTs are typically pre-randomization variables, the recommendation is to not winsorize or trim the covariate separately by experimental arm. Instead, *Pooled* winsorization/trimming should be used for covariates. However, the decision to winsorize covariates should be independent of the decision to winsorize the outcome variables.

The same intuition holds for outcome variables collected pre-randomization that are used as outcome variables in the regression analysis. This is common in DiD regressions within RCTs, such as the empirical application to Jack et al. (2023). As this data is collected pre-randomization, the recommendation is to use *Pooled* winsorizing/trimming.

6.4. Co-existing alongside other robustness checks

Researchers often use a variety of techniques to ensure the robustness of the reported treatment effects. These include multiple hypothesis testing to account for the increased probability of obtaining false positives (Type I errors) when conducting multiple statistical tests simultaneously, using machine learning tools to select covariates (Post-double selection Lasso, Belloni et al., 2014), and reporting randomized inference p-values.

Unlike winsorizing and trimming — which takes place during the data cleaning stage — these alternative robustness measures take place during the subsequent data analysis stage. As such, *Pooled* and *Stratified* winsorizing/trimming can co-exist alongside the other robustness measures mentioned above, as well as other techniques deployed during the data analysis stage to strengthen internal and external validity (e.g., techniques proposed by Young, 2019; Broderick et al., 2023; Andrews et al., 2024).

6.5. Statistical software

Below, the code for the traditional and stratified approach to winsorizing can be found for Stata and R. Online Appendix C shows the code for the pooled and stratified approach to trimming.

6.5.1. Stata

Pooled approach to winsorizing: winsor2 *OutcomeVar*, cuts(5 95)

Stratified Winsorizing: winsor2 *OutcomeVar*, cuts(5 95) by (*StratifiedVariable*)

[15] Only Angelucci and Bennett (2024) do not winsorize/trim at the 95th or 99th percentile, and instead winsorize observations outside 1.5 times the inter-quartile range, following the suggestion of Beyer and Tukey (1981).

6.5.2. R

I developed a new R package, called WinsorByGroupR, which can be found on GitHub. Once the package is installed, the functions are as follows:

Pooled approach to winsorizing: winsor(data, value_col = "OutcomeVar", bounds = c(5, 95))

Stratified winsorizing: winsorize_by_group(data, group_col = "StratifiedVariable", value_col = "OutcomeVar", bounds = c(5, 95))

7. Conclusion

Winsorizing and trimming are frequently used to reduce the role of outliers in dependent variables, by defining a percentile beyond which observations are considered outliers and hence winsorized/trimmed. However, this paper illustrates that winsorizing and trimming in RCTs is less innocuous than it seems and can bias a study's treatment effect estimates. These findings are in line with findings by Broderick et al. (2023) and Young (2019), who show that a few observations can have large effects on treatment effect estimates. This paper further shows how the winsorizing/trimming technique used can affect the likelihood of Type I and Type II errors.

While most RCTs winsorize/trim the entire sample, recent studies — including Benson et al. (2023), Muralidharan et al. (2023), and Bedoya et al. (2023) — have winsorized/trimmed separately per experimental arm, a technique called *Stratified* Winsorizing/Trimming. A formal framework and Monte Carlo simulations of an RCT illustrate that *Stratified* Winsorizing/Trimming on average result in a smaller bias of the treatment effect estimate, compared with the *Pooled* approach of winsorizing/trimming the whole sample. Furthermore, *Stratified* Winsorizing/Trimming improved the study's statistical power, at the cost of increasing the likelihood of Type I errors.

Applications to Angelucci et al. (2023) and Jack et al. (2023) illustrate that the decision of *how* to winsorize/trim has empirical implications, as estimated treatment effects and their statistical significance change. As such, authors should carefully consider how to winsorize/trim outliers, informed by the underlying data generating process. Recommendations are provided in Section 6.

The focus of the framework, simulations, and empirical applications in this paper has been on RCTs. This is intentional, as RCTs are the most common empirical setting in which outliers are winsorized/trimmed. Expanding the discussion and analysis to observational studies and other empirical methods, such as Difference-in-Difference or Regression Discontinuity Designs, is an interesting area for future research that can provide further insights on the robustness of reported empirical findings, and the importance of properly accounting for the role of outliers.

CRediT authorship contribution statement

Till Wicker: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jdeveco.2026.103815.

Data availability

The replication package will be uploaded here: 10.17632/6ffk227kkp.1 (Mendeley Data)

References

Allcott, H., Braghieri, L., Eichmeyer, S., Gentzkow, M., 2020. The Welfare Effects of Social Media. Am. Econ. Rev. 110 (3), 629–676.

Andrews, I., Kitagawa, T., McCloskey, A., 2024. Inference on Winners. Q. J. Econ. 139 (1), 305–358.

Angelucci, M., Bennett, D., 2024. Pharmacotherapy and Weight Loss in India: A Pre-Analysis Plan. Pre-Analysis Plan. (Accessed 21 February 2025).

Angelucci, M., Heath, R., Noble, E., 2023. Multifaceted programs targeting women in fragile settings: Evidence from the Democratic Republic of Congo. J. Dev. Econ. 164, 103146.

Angrist, J.D., Krueger, A.B., 2000. Empirical Strategies in Labor Economics. In: Handbook of Labor Economics, vol. 3A, Elsevier Science, Amsterdam, pp. 1277–1366.

Augsburg, B., De Haas, R., Harmgart, H., Meghir, C., 2015. The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. Am. Econ. J.: Appl. Econ. 7 (1), 183–203.

Bedoya, G., Belyakova, Y., Coville, A., Escande, T., Isaqzadeh, M., Ndiaye, A., 2023. The Enduring Impacts of a Big Push during Multiple Crises: Experimental Evidence from Afghanistan. Technical report, World Bank Group.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. J. Econ. Perspect. 28 (2), 29–50.

Benson, A., Board, S., Meyer-ter Vehn, M., 2023. Discrimination in Hiring: Evidence from Retail Sales. Rev. Econ. Stud. 91 (4), 1956–1987.

Beyer, H., Tukey, J.W., 1981. Exploratory data analysis. Addison-Wesley publishing company reading, mass. — Menlo park, cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. Biom. J. 23 (4), 413–414.

Bollinger, C., Chandra, A., 2005. Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data. J. Labor Econ. 23 (2), 235–258.

Broderick, T., Giordano, R., Meager, R., 2023. An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Routledge.

Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96 (1), 187–199.

de Mel, S., McKenzie, D., Woodruff, C., 2019. Labor Drops: Experimental Evidence on the Return to Additional Labor in Microenterprises. Am. Econ. J.: Appl. Econ. 11 (1), 202–235.

Fafchamps, M., McKenzie, D., Quinn, S., Woodruff, C., 2012. Using PDA consistency checks to increase the precision of profits and sales measurement in panels. J. Dev. Econ. 98 (1), 51–57, Symposium on Measurement and Survey Design.

Goldberger, A.S., 1981. Linear regression after selection. J. Econometrics 15 (3), 357–366.

Gollin, D., Udry, C., 2021. Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture. J. Politi. Econ. 129 (1), 1–80.

Heckman, J., 1979. Sample Selection Bias as a Specification Error. Econometrica 47 (1), 153–161.

Heckman, J., 1990. Varieties of Selection Bias. Am. Econ. Rev. 80 (2), 313–318.

Jack, W., Kremer, M., de Laat, J., Suri, T., 2023. Credit Access, Selection, and Incentives in a Market for Asset-Collateralized Loans: Evidence From Kenya. Rev. Econ. Stud. 90 (6), 3153–3185.

Khan, S., Tamer, E., 2010. Irregular identification, support conditions, and inverse weight estimation. Econometrica 78 (6), 2021–2042.

Ma, X., Wang, J., 2020. Robust inference using inverse probability weighting. J. Amer. Statist. Assoc. 115 (532), 1851–1860.

McKenzie, D., 2012. Beyond baseline and follow-up: The case for more t in experiments. J. Dev. Econ. 99 (2), 210–221.

Meager, R., 2022. Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. Am. Econ. Rev. 112 (6), 1818–1847.

Muralidharan, K., Niehaus, P., Sukhtankar, S., 2023. General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence From India. Econometrica 91 (4), 1261–1295.

Schilbach, F., 2019. Alcohol and Self-Control: A Field Experiment in India. Am. Econ. Rev. 109 (4), 1290–1322.

World Bank, 2023. Variable construction. https://dimewiki.worldbank.org/Variable_Construction. (Accessed 23 February 2024).

Young, A., 2019. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. Q. J. Econ. 134 (2), 557–598.

Further reading

Bahadur, R.R., 1966. A note on quantiles in large samples. Ann. Math. Stat. 37 (3), 577–580, http://www.jstor.org/stable/2238725.