Multiple Testing of Ordinal Stochastic Monotonicity

Qian Wu* David M. Kaplan†

October 30, 2023

Abstract

We develop methodology for testing stochastic monotonicity when the outcome variable is ordinal. Rather than testing a single null hypothesis, we use multiple testing to evaluate where the ordinal outcome is stochastically increasing in the covariate. By inverting our multiple testing procedure that controls the familywise error rate, we construct "inner" and "outer" confidence sets for the true set of points consistent with stochastic increasingness. Simulations show reasonable finite-sample properties. Empirically, we apply our methodology to the relationship between mental health and education. Practically, we provide code implementing our multiple testing procedure and replicating our empirical results.

JEL classification: C25, I10

Keywords: familywise error rate, confidence set, mental health

1 Introduction

One fundamental empirical question is whether a variable Y is "increasing" in a covariate X, which can be characterized in terms of stochastic monotonicity. For example, is mental health increasing in education? Variable Y is stochastically increasing in X if for any two X values, the conditional distribution of Y at the higher X value first-order stochastically dominates the conditional distribution of Y at the lower X value. Economically, this means the conditional distribution of Y is "better" at higher X values, in the sense of higher expected utility. However, because this is a strong property, it can be statistically difficult

*Corresponding author; School of Statistics, Southwestern University of Finance and Economics, wuqj@swufe.edu.cn

†Department of Economics, University of Missouri, kaplandm@missouri.edu

to find strong evidence in its favor, and conversely it may nearly hold but be rejected for a small violation.

We contribute to the stochastic monotonicity inference literature by proposing and justifying new methodology that focuses on ordinal outcomes and multiple testing. Economists study ordinal outcomes in variety of fields, using variables like general health (from the lowest category "poor" up to "excellent"), mental health, bond ratings, education level, consumer confidence, subjective well-being, and school ratings. However, other stochastic monotonicity tests (see below) assume a continuous outcome and only test a single null hypothesis of global stochastic monotonicity.

Instead of testing the single null hypothesis that Y is stochastically increasing in X, we consider multiple testing of the conditional CDF inequalities that jointly comprise stochastic increasingness. Specifically, each inequality checks whether the conditional CDF of Y is decreasing as X increases to the next-highest value, over all categories of Y and a finite number of X values.

Complementing global testing of a single null (which in principle can have better power against certain alternatives), multiple testing addresses the aforementioned statistical concerns in the following ways. First, if stochastic increasingness is rejected, multiple testing shows more precisely whether many inequalities are rejected, or only a few, or even just one. Second, if stochastic increasingness is not rejected, then multiple testing can be used on the reversed inequalities (conditional CDF increasing with X) to see if they can be rejected in favor of the original inequalities (decreasing with X). This gathers stronger statistical evidence in favor of stochastic increasingness than simply failing to reject the original null hypothesis, because type II error rates (false non-rejection) are not controlled. For example, non-rejection may occur due to small sample size (high statistical uncertainty) even if the point estimate suggests the null hypothesis is false.

Additionally, our multiple testing procedure can be inverted into "inner" and "outer" confidence sets. The inner confidence set contains all points at which the reversed inequalities

have been rejected, i.e., where there is strong evidence in favor of the original inequality consistent with stochastic increasingness. The outer confidence set contains all points at which the original inequalities have not been rejected. Such confidence sets are similar to those of Kaplan (2023) and to equation (1) of Armstrong and Shen (2015). The inner confidence set can be seen as a conservative estimator of the true set of inequalities consistent with stochastic increasingness, in that the inner set is contained within the true set with high asymptotic probability. Conversely, the outer set contains the true set with high asymptotic probability.

Our empirical application studies mental health, an important topic of increasing interest, specifically the relationship between depression and education. Our methodology finds points with strong evidence in favor of mental health stochastically increasing with education (depression stochastically decreasing with education), or strong evidence against it. Given our observational data, our analysis is descriptive rather than causal, but it shows an interesting pattern. Generally, we reject that individuals with higher education have worse mental health in favor of the alternative that they have better mental health (lower depression). However, this is not true when comparing high-school graduates with individuals who have some college but not a bachelor's degree. Not only can we not reject in favor of lower depression, but we reject lower depression in favor of higher depression for the some-college individuals, specifically a higher proportion with mild-or-worse depression and a higher proportion with moderate-or-worse depression. Because we control the familywise error rate, there is no concern that these rejections are simply artifacts of testing so many combinations of education level and depression category.

Besides intrinsic interest in settings like the education gradient in health or intergenerational transmission/mobility (a child's outcome Y compared to their parent's outcome X), stochastic monotonicity also appears in certain identifying assumptions or as a testable implication thereof. For example, stochastic monotonicity is implied by the combination of (semi-)monotone treatment response and exogenous treatment selection, as discussed by

Manski (1997, §3.4). As another example, Small, Tan, Ramsahai, Lorch, and Brookhart (2017) identify a weighted average treatment effect (and partially identify the average treatment effect) when the treatment is stochastically increasing in the instrument. Although testing is trivial in the binary–binary setting they focus on, our methodology would be useful for testing their stochastic monotonicity identifying assumption when the treatment and instrument are ordinal or discrete.

Our methodology contributes to the literature on stochastic monotonicity and multiple testing. Inference on stochastic monotonicity has focused on continuous Y and testing the single hypothesis of global stochastic monotonicity; for example, see Lee, Linton, and Whang (2009), Seo (2018), and Chetverikov, Liao, and Chernozhukov (2021). Our reversing the direction of null hypothesis inequalities to find stronger evidence in favor of stochastic increasingness is inspired by Davidson and Duclos (2013), who make a related argument for testing the null of non-dominance against the alternative of stochastic dominance, which can be seen as a special case of stochastic monotonicity with binary X. Although we emphasize multiple testing, testing the set of stochastic monotonicity inequalities jointly would fit in the (moment) inequality literature; we essentially follow the least favorable approach with a max-t statistic as in Section 4 (especially 4.1.1) of Canay and Shaikh (2017), whose survey includes additional references. Our proof strategies are also similar to those of Zhao (2023), who instead considers multiple testing of a function's value across a finite set of points. Kaplan and Zhao (2023) also use multiple testing with ordinal variables, but they only compare two (unconditional) distributions and focus not on ordinal categories but on a latent variable's quantiles. Surveys of the multiple testing literature can be found in Lehmann and Romano (2005b) and Romano, Shaikh, and Wolf (2010).

Paper structure Section 2 describes the setting, assumptions, methodology, and asymptotic properties. Section 3 applies our methodology empirically to mental health and education. Section 4 presents simulation results. Appendix A collects proofs. Appendix B

contains an additional empirical application with general health.

Notation and abbreviations Random and non-random vectors are respectively typeset as, e.g., $\boldsymbol{X}$ and $\boldsymbol{x}$, while random and non-random scalars are typeset as X and x, and random and non-random matrices as $\underline{\boldsymbol{X}}$ and $\underline{\boldsymbol{x}}$. The indicator function is $\mathbb{1}\{\cdot\}$, with $\mathbb{1}\{A\} = 1$ if event A occurs and $\mathbb{1}\{A\} = 0$ if not. " $\iff$ " means if and only if. Acronyms used include those for confidence set (CS), continuous mapping theorem (CMT), cumulative distribution function (CDF), familywise error rate (FWER), and multiple testing procedure (MTP).

2 Main Results

In this section, we describe the setting, assumptions, methodology, and asymptotic properties.

2.1 Setting

Consider random variables Y and X. The outcome Y is ordinal, with categories labeled $Y \in \{1, 2, \ldots, J\}$. The covariate may also be ordinal, or discrete, or a discretization of a continuous variable, as long as there is a finite number of possible values, labeled as $X \in \{1, 2, \ldots, K\}$. For example, if the covariate is ordinal with values "low," "medium," and "high," then $X = 1$ is the label for "low," $X = 2$ for "medium," and $X = 3$ for "high." Or if the covariate is discrete with points of interest 0.1, 0.2, ..., 0.9, then $X = 1$ is the label for 0.1, $X = 2$ for 0.2, ..., $X = 9$ for 0.9. In sum, the supports $\mathcal{Y}$ and $\mathcal{X}$ are

$$Y \in \mathcal{Y} \equiv \{1, \ldots, J\}, \quad X \in \mathcal{X} \equiv \{1, \ldots, K\}. \tag{1}$$

The population and empirical (estimated) conditional CDF values are respectively

$$\begin{aligned} F_x(y) &\equiv \mathrm{P}(Y \leq y \mid X = x), \\ \hat{F}_x(y) &\equiv \frac{\sum_{i=1}^n \mathbb{1}\{Y_i \leq y\}\,\mathbb{1}\{X_i = x\}}{n_x}, \quad n_x \equiv \sum_{i=1}^n \mathbb{1}\{X_i = x\}, \end{aligned} \tag{2}$$

given observations $i = 1, \ldots, n$. That is, $F_x(y)$ is the proportion of the $X = x$ subpopulation with $Y \leq y$, and $\hat{F}_x(y)$ is the proportion of the $X_i = x$ subsample with $Y_i \leq y$.

We assume iid sampling and condition on the size of each subsample n_x, which is equivalent to conditioning on all the observed X_i values (like in classical linear regression results). This is also equivalent to repeated sampling of n_x iid draws of Y_i from the corresponding conditional distribution independently for each $x = 1, \ldots, K$, which is how we formalize the setting in Assumption A1.

Assumption A1. Using the notation in (1) and (2), $n_1/n_x \to \gamma_x \in (0, \infty)$ for each $x \in \mathcal{X}$, and the sample consists of n_x values of $X_i = x$ for each $x \in \mathcal{X}$, with the corresponding Y_i values sampled iid from the corresponding population conditional distribution of $Y \mid X = x$, and all Y_i are mutually independent.

2.2 Stochastic monotonicity inequalities

Outcome Y stochastically increasing in X means that for any $x_2 > x_1$, the conditional distribution of $Y \mid X = x_2$ first-order stochastically dominates that of $Y \mid X = x_1$. Using the notation of (2), this means

$$F_{x_2}(y) \leq F_{x_1}(y) \text{ for all } (x_1, x_2, y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \text{ with } x_2 > x_1. \tag{3}$$

Strict and weak monotonicity/inequalities are statistically indistinguishable, so we do not emphasize the difference. We restrict attention to $y \in \{1, \ldots, J-1\}$ because $F_{x_2}(J) = F_{x_1}(J) = 1$ for all x_1, x_2. We also restrict attention to $x_2 = x_1 + 1$. Note (3) holds if and

only if

$$\theta_{x,y} \leq 0 \text{ for all } (x,y) \in \{1,\ldots,K-1\} \times \{1,\ldots,J-1\}, \quad \theta_{x,y} \equiv F_{x+1}(y) - F_x(y). \tag{4}$$

Although first-order stochastic dominance generally suggests one ordinal distribution is "better" than another, the relationship is more complex if utility depends not on the observed ordinal value but rather on a latent continuous variable. Given a model with fixed thresholds, latent dominance implies ordinal dominance, but not vice-versa. Thus, rejecting ordinal stochastic monotonicity implies rejecting latent stochastic monotonicity, and partial evidence of ordinal stochastic monotonicity still provides some evidence in favor of latent stochastic monotonicity, but even full ordinal stochastic monotonicity does not imply latent stochastic monotonicity. For related discussion and results, see Kaplan and Zhao (2023).

2.3 Multiple testing procedure

Based on (4), we consider multiple testing of the following family of hypotheses:

$$H_{0x,y}\colon \theta_{x,y} \leq 0 \text{ for } (x,y) \in \{1,\ldots,K-1\} \times \{1,\ldots,J-1\}. \tag{5}$$

The number of hypotheses is $(K-1)(J-1)$. A multiple testing procedure makes a binary decision (reject or not) for each hypothesis, hence $2^{(K-1)(J-1)}$ possible results. In principle, we could include even more hypotheses, based on (3). Even though (3) and (4) are equivalent, multiple testing of the corresponding inequality hypotheses is not equivalent. We use (5) because the interpretation is clear, the results are easier to communicate, and the critical value does not need to be as large as when including lots of additional hypotheses. Nonetheless, alternative treatments of the full family of inequalities based on (3) remains a question for future research.

Our multiple testing procedure (MTP) controls the asymptotic familywise error rate (FWER). There are other measures of overall false positive rate for multiple testing, like the k-FWER and false discovery proportion of Lehmann and Romano (2005a) and the false

discovery rate of Benjamini and Hochberg (1995); we choose FWER because it has a very clear interpretation and allows us to invert our MTP into confidence sets. FWER is defined as (e.g., Lehmann and Romano, 2005b, §9.1)

$$\text{FWER} \equiv \text{P}(\text{reject any true } H_{0x,y}). \tag{6}$$

Conversely, $1 - \text{FWER}$ is the probability of having zero false rejections (type I errors). Our MTP has "strong control" meaning that it controls asymptotic FWER regardless of which $H_{0x,y}$ are true or false.

Our MTP uses standard t-statistics but with a higher critical value that accounts for multiple testing. For any real scalar m, define

$$\hat{t}_{x,y}(d) \equiv \frac{\hat{F}_{x+1}(y) - \hat{F}_x(y) - d}{\hat{s}_{x,y}}, \tag{7}$$

where $\hat{s}_{x,y}$ is defined in (21) as an estimator of the standard deviation of $\hat{F}_{x+1}(y) - \hat{F}_x(y)$. As usual, in practice we compute the t-statistic centered at our hypothesized value $d = 0$, and asymptotic properties can be bounded by the behavior of the t-statistic centered at the true $d = \theta_{x,y} \equiv F_{x+1}(y) - F_x(y)$. As an important ingredient of our FWER derivations, Lemma 1 establishes the asymptotic normal distribution of random vector $\hat{\boldsymbol{t}}$ that contains all the t-statistics centered at the true $\theta_{x,y}$:

$$\hat{\boldsymbol{t}} \equiv (\hat{\boldsymbol{t}}_1, \hat{\boldsymbol{t}}_2, \ldots, \hat{\boldsymbol{t}}_{J-1})', \quad \hat{\boldsymbol{t}}_j \equiv \big(\hat{t}_{K-1,j}(\theta_{K-1,j}), \hat{t}_{K-2,j}(\theta_{K-2,j}), \ldots, \hat{t}_{1,j}(\theta_{1,j})\big). \tag{8}$$

Lemma 1. *Under Assumption A1,*

$$\hat{\boldsymbol{t}} \xrightarrow{d} \boldsymbol{t} \sim \text{N}(\boldsymbol{0}, \underline{\boldsymbol{\Sigma}}), \quad \underline{\boldsymbol{\Sigma}} = \underline{\boldsymbol{S}}' \underline{\boldsymbol{W}} \underline{\boldsymbol{S}},$$

where $\underline{\boldsymbol{W}}$ *and* $\underline{\boldsymbol{S}}$ *are defined in* (19) *and* (23)*.*

We provide some brief intuition for the critical value whose validity is formally established in the proof of Theorem 2. Intuitively, the least favorable configuration here is when all $\theta_{x,y} = 0$: every null $H_{0x,y}$ is true (and thus can contribute to FWER), but just barely, so

there is the highest probability of some estimated $\hat{\theta}_{x,y} > 0$ large enough to reject $H_{0x,y}$. In that case, $\hat{t}_{x,y}(0) = \hat{t}_{x,y}(\theta_{x,y})$, and we make a familywise error whenever the maximum (over x, y) t-statistic exceeds our critical value. Thus, using the asymptotic approximation, if we want FWER $= \alpha$, then the critical value should be the $(1-\alpha)$-quantile of the distribution of the maximum of $\boldsymbol{t}$:

$$c_\alpha \equiv (1-\alpha)\text{-quantile of } \max_{x,y} t_{x,y}, \tag{9}$$

where the $t_{x,y}$ are the elements of $\boldsymbol{t}$ and $\boldsymbol{t} \sim \mathrm{N}(\mathbf{0}, \underline{\underline{\boldsymbol{\Sigma}}})$ by Lemma 1. Because $\underline{\underline{\boldsymbol{\Sigma}}}$ is unknown, in practice it can be replaced with consistent estimator $\hat{\underline{\underline{\boldsymbol{\Sigma}}}}$:

$$\hat{c}_\alpha \equiv (1-\alpha)\text{-quantile of } \max_{x,y} \tilde{t}_{x,y}, \quad \tilde{\boldsymbol{t}} \sim \mathrm{N}(\mathbf{0}, \hat{\underline{\underline{\boldsymbol{\Sigma}}}}). \tag{10}$$

Although an analytic formula is intractable, (10) can be simulated as in our code.

Method 1. *First, compute the t-statistics $\hat{t}_{x,y}(0)$ as in* (7)*. Second, given the desired FWER level α, simulate the critical value $\hat{c}_\alpha$ in* (10) *following the steps in Section 2.5. Third, for each (x, y), to test all $H_{0x,y}\colon \theta_{x,y} \leq 0$ as in* (5)*, reject $H_{0x,y}\colon \theta_{x,y} \leq 0$ if $\hat{t}_{x,y}(0) > \hat{c}_\alpha$. Alternatively, to test the reversed family of hypotheses $H^{\geq}_{0x,y}\colon \theta_{x,y} \geq 0$, reject $H_{0x,y}$ if $\hat{t}_{x,y}(0) < -\hat{c}_\alpha$.*

Method 1 includes the reversed $H^{\geq}_{0x,y}$ because their rejection provides stronger evidence of $\theta_{x,y} < 0$ than does non-rejection of $H_{0x,y}$. For example, even if $\hat{\theta}_{x,y} = 1.7$ (our best guess is a positive $\theta_{x,y}$), if there is a small sample size or otherwise high uncertainty, we may still not reject $H_{0x,y}$. In contrast, to reject $H^{\geq}_{0x,y}$, not only must we have $\hat{\theta}_{x,y} < 0$, but it must be significantly less than zero (compared to our uncertainty) in order for the test to control the type I error rate (or FWER). In sum, non-rejection of $H_{0x,y}$ suggests the data are consistent with the hypothesis that Y is stochastically increasing in X, but rejection of $H^{\geq}_{0x,y}$ provides stronger evidence in favor of the inequalities that comprise stochastic increasingness of Y in X.

Theorem 2 theoretically justifies our MTP.

Theorem 2. *Under Assumption A1, Method 1 has strong control of asymptotic FWER at level α.*

2.4 Confidence sets

The MTP in Method 1 can be inverted into confidence sets for the true set

$$\mathcal{T} \equiv \{(x, y) : \theta_{x,y} \leq 0\}. \tag{11}$$

The goal is, with high asymptotic probability, for inner confidence set $\widehat{\mathcal{CS}}_{inner}$ to be contained within the true set $\mathcal{T}$, and for outer confidence set $\widehat{\mathcal{CS}}_{outer}$ to contain $\mathcal{T}$. That is, for confidence level $1 - \alpha$,

$$\begin{aligned} \lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) \geq 1 - \alpha, \\ \lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}) \geq 1 - \alpha. \end{aligned} \tag{12}$$

Intuitively, $\widehat{\mathcal{CS}}_{inner}$ provides a conservative assessment of the true set $\mathcal{T}$, while $\widehat{\mathcal{CS}}_{outer}$ gives a larger "estimate" describing how large the true set might be. Corresponding to the earlier discussion of stronger and weaker evidence, the inner confidence set collects points for which the reversed $H^{\geq}_{0x,y}$ is rejected (strong evidence), whereas the outer confidence set collects points for which the original $H_{0x,y}$ is not rejected (weak evidence).

Theorem 3 formally establishes the property in (12) for our confidence sets described in Method 2.

Method 2. *First run Method 1. The inner confidence set $\widehat{\mathcal{CS}}_{inner}$ collects all pairs of (x, y) for which the reversed null $H^{\geq}_{0x,y}\colon \theta_{x,y} \geq 0$ is rejected. The outer confidence set $\widehat{\mathcal{CS}}_{outer}$ collects all pairs of (x, y) for which the original null $H_{0x,y}\colon \theta_{x,y} \leq 0$ is not rejected.*

Theorem 3. *Under Assumption A1, Method 2 satisfies* (12)*.*

If instead of the set $\mathcal{T}$ in (11) we are interested in its complement $\mathcal{T}^{\complement} \equiv \{(x, y) : \theta_{x,y} > 0\}$, then the outer CS is the complement of the inner CS for $\mathcal{T}$, and the inner CS is the

complement of the outer CS for $\mathcal{T}$. This follows because the event $\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}$ is equivalent to the inner CS complement containing $\mathcal{T}^{\complement}$, and the event $\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}$ is equivalent to the outer CS complement being contained within $\mathcal{T}^{\complement}$. Thus, using the notation from (12),

$$\lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{inner}^{\complement} \supseteq \mathcal{T}^{\complement}) = \lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) \geq 1-\alpha,$$
$$\lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{outer}^{\complement} \subseteq \mathcal{T}^{\complement}) = \lim_{n\to\infty} \mathrm{P}(\widehat{\mathcal{CS}}_{outer} \supseteq \mathcal{T}) \geq 1-\alpha.$$

2.5 Critical value simulation

As noted earlier, we compute the critical value $\hat{c}_\alpha$ by simulation. First, let

$$\hat{\underline{\boldsymbol{\Sigma}}} = \hat{\underline{\boldsymbol{S}}}' \hat{\underline{\boldsymbol{W}}} \hat{\underline{\boldsymbol{S}}}, \tag{13}$$

where $\hat{\underline{\boldsymbol{W}}}$ and $\hat{\underline{\boldsymbol{S}}}$ are the sample analogs of $\underline{\boldsymbol{W}}$ and $\underline{\boldsymbol{S}}$ from (19) and (23). The consistency result in (22) implies $\hat{\underline{\boldsymbol{S}}} \xrightarrow{p} \underline{\boldsymbol{S}}$, and by the weak law of large numbers $\hat{\underline{\boldsymbol{W}}} \xrightarrow{p} \underline{\boldsymbol{W}}$, so applying the continuous mapping theorem yields $\hat{\underline{\boldsymbol{\Sigma}}} \xrightarrow{p} \underline{\boldsymbol{\Sigma}}$.

The simulation proceeds as follows, as implemented in our provided code. First, we generate a random draw (a vector) from $\mathrm{N}(\mathbf{0}, \hat{\underline{\boldsymbol{\Sigma}}})$. Second, we take the maximum of this vector. Third, we repeat this process N times, collecting all the maxima. Fourth, we take the $(1-\alpha)$-quantile of the N maximum values; this is $\hat{c}_\alpha$.

With large enough N, we can achieve arbitrarily small simulation error. Larger N improves accuracy but increases computation time. Given our simulation results (Section 4), we suggest $N = 1000$ as a default. In our empirical application, we use $N = 100{,}000$ because computation time is still only a few seconds.

3 Empirical Illustration

This section includes an empirical example to demonstrate our methodology's results and interpretations. The multiple testing and confidence set results provide a much more detailed picture than simply "reject stochastic monotonicity." All variables come from the publicly

available NHIS 2019 data (National Center for Health Statistics, 2019). Because total runtime is only a few seconds, we use $N = 100{,}000$ random draws to compute the critical value. Replication code in R (R Core Team, 2023) is provided.

Many studies have examined the relationship between reported depression and education (e.g., Bauldry, 2015; Cohen, Nussbaum, Weintraub, Nichols, and Yen, 2020; Lee, 2011). The nature of the correlation varies across different investigations. Individuals with more education may possess better mental health management skills, but they may also more readily recognize and report their psychological states. Further, the relationship is potentially different when (for example) comparing high-school dropouts and graduates than when comparing bachelor's and graduate degree holders, and the relationship may be different for severe depression than for milder forms. Our method can detect any such patterns because it is nonparametric and does not impose any restrictions. We assess evidence of where the level of depression is stochastically decreasing in education, i.e., where mental health is improving.

We use the following variables. Depression variable Y (`PHQCAT_A`) summarizes the eight-item Patient Health Questionnaire depression scale (PHQ-8) into four categories: "none/minimal," "mild," "moderate," and "severe," respectively coded as 1, 2, 3, and 4. Education variable X (recoded from `EDUC_A`) has five categories: "no high school degree," "HS degree only," "some college" (including associate's degree), "bachelor's degree," and "graduate degree" (master's, professional, or doctoral), respectively coded as 1, 2, 3, 4, and 5. We restrict the age range to 30–65 years old for a fair comparison of people with various levels of education including graduate degrees, and to concentrate on those who are eligible for the workforce. For simplicity, observations with missing values are dropped and weights are not used.

Figure 1 visualizes the data on depression and education in a mosaic plot. Each column corresponds to a category of X, and its width is proportional to the number of observations in that category. For example, the HS column is significantly wider than the no-HS column, showing that more individuals in the data have (only) a high-school degree than do not

Figure 1: Mosaic plot of depression and education.

have a high-school degree. Each cell's area is proportional to the sample proportion of individuals with that particular value of (X, Y); the text label shows that proportion as a percentage, along with the corresponding number of observations in that cell. Because the joint probability of $(X, Y) = (x, y)$ is the product of the marginal $X = x$ probability and the conditional Y probability given $X = x$, within a column corresponding to $X = x$, each cell's height is proportional to the proportion of observations with the corresponding y value within the $X_i = x$ subsample. These conditional probabilities are scaled such that the full height of each column is probability one (or 100%). For example, in the HS column, the "none/minimal" cell's height is over half the total column height, meaning that among individuals in the data who have (only) a high-school degree, over half have depression level "none/minimal." This further implies that the breaks between cells within a column show the conditional CDF values. For example, in the no-HS column, the top of the "mild" cell is

around 7/8 between the bottom and top of the column, indicating the empirical conditional CDF is around $\hat{F}_1(2) \approx 7/8$. For each level y, the mosaic plot shows the data sample follows the pattern $\hat{F}_1(y) \leq \hat{F}_2(y)$ and $\hat{F}_3(y) \leq \hat{F}_4(y) \leq \hat{F}_5(y)$ (i.e., generally lower depression at higher education), but $\hat{F}_2(y) \geq \hat{F}_3(y)$. However, we wish to learn not only about the sample but about the population relationship. Our methods help assess the strength of the evidence that these patterns hold in the population.

Table 1: Test statistics for depression versus education ($\hat{c}_{0.05} = 2.59$).

	Education category x			
Depression level y	1 (vs. 2)	2 (vs. 3)	3 (vs. 4)	4 (vs. 5)
1: none/minimal	**9.21**	−4.83	**10.18**	**3.86**
2: mild (or below)	**6.64**	−3.26	**9.47**	**2.81**
3: moderate (or below)	**4.69**	−2.17	**7.27**	**2.63**

True set: points where mental health is better (lower depression) at next-higher education level, $\{(x, y) : F_x(y) \leq F_{x+1}(y)\}$. Gray shading: outer confidence set. Bold: inner confidence set. Confidence level 95%. Critical value computed using $N = 100{,}000$ random draws.

Table 1 shows our inference results, which can be interpreted as follows. The numbers are the t-statistics for testing the null hypotheses $H_{0x,y}$: $F_x(y) \geq F_{x+1}(y)$. For example, the bottom-right number in the table is for $x = 4$ and $y = 3$, corresponding to the t-statistic for the null hypothesis that $F_4(3) \geq F_5(3)$, i.e., that the population probability of having moderate or below depression ($Y \leq 3$) is higher among individuals with (only) a bachelor's degree ($x = 4$) than among individuals who also have a graduate degree ($x = 5$). The corresponding t-statistic $\hat{t}_{x,y}(0)$ from (7) is $\hat{t}_{4,3}(0) = (\hat{F}_5(3) - \hat{F}_4(3))/\hat{s}_{4,3} = 2.63$, which is above $\hat{c}_{0.05} = 2.59$ (hence bold **2.63** in the table), so we reject $F_4(3) \geq F_5(3)$ in favor of $F_4(3) < F_5(3)$. Because there are only $J = 4$ categories of Y, we can also interpret the conclusion $F_4(3) < F_5(3)$ as $\mathrm{P}(Y = 4 \mid X = 4) > \mathrm{P}(Y = 4 \mid X = 3)$, i.e., a higher probability of severe depression in the bachelor's-only subpopulation than in the graduate degree subpopulation. More generally in the table, a t-statistic is positive when $\hat{F}_{x+1}(y) > \hat{F}_x(y)$, indicating a larger proportion of individuals with depression level y or less in the

$X = x + 1$ subsample than in the $X = x$ subsample, generally indicating less depression (better mental health) at the higher education level $x + 1$ than at x. If the t-statistic is positive and above the critical value, then there is strong evidence that this is true in the population, too. Consequently, when $\hat{t}_{x,y}(0) > \hat{c}_{0.05}$, that (x, y) is included in the 95% inner confidence set, which collects the points with strong evidence of depression decreasing with education; such t-statistics are in **bold**. The 95% outer confidence set includes (x, y) as long as there is no strong evidence in the opposite direction, i.e., as long as $\hat{t}_{x,y}(0) > -\hat{c}_{0.05}$; such t-statistics are shaded gray. Note the $\hat{t}_{x,y}(0) > \hat{c}_{0.05}$ are thus **bold and shaded** because the outer confidence set contains the inner confidence set.

Table 1 shows strong evidence of a general pattern of lower depression with higher education, but with important exceptions. Most of the t-statistics exceed our 5% FWER critical value, collectively forming a large 95% inner confidence set for the true set of points $\{(x, y) : F_x(y) \leq F_{x+1}(y)\}$ where mental health is better (lower depression) at the higher education level. This includes all depression levels y and all education levels except $x = 2$ (vs. 3): only-HS compared to some-college. The outer confidence set additionally includes cell $(x, y) = (2, 3)$ with a t-statistic that is not quite negative enough for our multiple testing procedure to reject in the opposite direction. (The negative point estimate means our best guess is that $F_3(3) - F_2(3) < 0$, but there is too much uncertainty for the multiple testing procedure to reject.) However, the t-statistics at $(x, y) = (2, 1)$ and $(2, 2)$ are negative enough to reject, so they are excluded from the outer confidence set. Thus, there is some strong evidence that the general pattern of less depression at higher education does not hold everywhere, specifically when comparing the only-HS and some-college subpopulations.

Overall, Table 1 provides much richer results than simply saying we reject the global null hypothesis of stochastic monotonicity.

4 Simulations

This section presents simulations of finite-sample FWER, for a variety of sample sizes and numbers of Y and X categories. We use the "least favorable" configuration that maximizes FWER by setting all null hypothesis inequalities to be binding, i.e., $\theta_{x,y} = 0$ for all (x, y). We provide code in R (R Core Team, 2023) to replicate our results.

The simulation proceeds as follows. For the data generating process, Y and X respectively have J and K categories. The conditional distribution of Y is uniform across the J categories, so the conditional CDF is $F_x(y) = y/J$ for each (x, y) and all $\theta_{x,y} \equiv F_{x+1}(y) - F_x(y) = y/J - y/J = 0$. After randomly drawing a dataset having n_x observations with $X_i = x$ for each x (so total sample size $n = Kn_x$), Method 1 is run, with the critical value based on N simulations, and the results are stored. This is repeated 1000 times. Because all $H_{0x,y}$ are true, the simulated FWER is the proportion of simulated datasets for which at least one $H_{0x,y}$ is rejected. Besides FWER, we report the full range (across all simulated datasets) of simulated critical values c_α for each case.

Table 2 presents the simulation results. They are divided into three sections corresponding to the different (J, K), but results are qualitatively similar in each section, showing the following patterns. First, FWER is somewhat above $\alpha = 0.05$ with the smallest sample size, closer to 0.10, suggesting that (as usual) results from smaller samples should be interpreted more cautiously. Second, as n_x increases, FWER approaches the nominal α. This reflects our procedure's exact asymptotic FWER in this least favorable configuration; FWER would be lower in other cases. Third, compared to $N = 1000$, computing the critical value using $N = 10{,}000$ draws improves stability somewhat (tighter range of critical values) but improves FWER accuracy only very slightly, so in practice we suggest using $N = 1000$ as the default.

Table 2: Simulation results.

J	K	n_x	N	α	c_α	FWER
4	4	20	1000	0.05	[2.32, 2.59]	0.098
4	4	100	1000	0.05	[2.34, 2.58]	0.068
4	4	1000	1000	0.05	[2.36, 2.57]	0.060
4	4	10,000	1000	0.05	[2.36, 2.55]	0.045
4	4	10,000	10,000	0.05	[2.43, 2.52]	0.047
4	4	10,000	10,000	0.10	[2.18, 2.25]	0.112
4	4	10,000	10,000	0.01	[2.94, 3.08]	0.010
6	5	20	1000	0.05	[2.59, 2.84]	0.092
6	5	100	1000	0.05	[2.55, 2.81]	0.075
6	5	1000	1000	0.05	[2.57, 2.79]	0.068
6	5	10,000	1000	0.05	[2.60, 2.79]	0.053
6	5	10,000	10,000	0.05	[2.70, 2.77]	0.047
6	5	10,000	10,000	0.10	[2.45, 2.50]	0.105
6	5	10,000	10,000	0.01	[3.18, 3.32]	0.007
8	10	20	1000	0.05	[2.92, 3.18]	0.099
8	10	100	1000	0.05	[2.95, 3.18]	0.074
8	10	1000	1000	0.05	[2.94, 3.20]	0.056
8	10	10,000	1000	0.05	[2.95, 3.18]	0.046
8	10	10,000	10,000	0.05	[3.04, 3.12]	0.046
8	10	10,000	10,000	0.10	[2.82, 2.88]	0.090
8	10	10,000	10,000	0.01	[3.47, 3.61]	0.007

5 Conclusion

We have proposed inference methods to provide a richer assessment of stochastic monotonicity when the outcome is ordinal, by separately considering each conditional CDF inequality that together comprise stochastic monotonicity. Our multiple testing procedure controls the asymptotic familywise error rate, and with the desired asymptotic probability, our outer confidence set contains the set of true inequalities and our inner confidence set is contained within it. In future work, to complement these frequentist confidence sets, we plan to develop Bayesian credible sets and assess their asymptotic frequentist properties. For the related *joint* testing problem, the frequentist test is much more conservative than a Bayesian test even asymptotically (e.g., Kaplan and Zhuo, 2021; Kline, 2011), but this does not translate directly to multiple testing and related confidence sets. Deriving multiple testing procedures

for continuous X and/or Y would also be valuable.

References

Armstrong, Timothy B. and Shu Shen. 2015. "Inference on Optimal Treatment Assignments." Working paper available at https://tbarmstr.github.io/.

Bauldry, Shawn. 2015. "Variation in the Protective Effect of Higher Education against Depression." *Society and Mental Health* 5 (2):145–161. URL https://doi.org/10.1177/2156869314564399.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B* 57 (1):289–300. URL https://www.jstor.org/stable/2346101.

Canay, Ivan A. and Azeem M. Shaikh. 2017. "Practical and Theoretical Advances in Inference for Partially Identified Models." In *Advances in Economics and Econometrics: Eleventh World Congress*, *Econometric Society Monographs*, vol. 2, edited by Ariel Pakes, Bo Honoré, Larry Samuelson, and Monika Piazzesi, chap. 9. Cambridge: Cambridge University Press, 271–306. URL https://doi.org/10.1017/9781108227223.009.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov. 2021. "On Cross-Validated Lasso in High Dimensions." *Annals of Statistics* 49 (3):1300–1317. URL https://doi.org/10.1214/20-AOS2000.

Cohen, Alison K., Juliet Nussbaum, Miranda L. Ritterman Weintraub, Chloe R. Nichols, and Irene H. Yen. 2020. "Association of Adult Depression With Educational Attainment, Aspirations, and Expectations." *Preventing Chronic Disease* 17 (94). URL https://doi.org/10.5888/pcd17.200098.

Davidson, Russell and Jean-Yves Duclos. 2013. "Testing for Restricted Stochastic Dominance." *Econometric Reviews* 32 (1):84–125. URL https://doi.org/10.1080/07474938.2012.690332.

Kaplan, David M. 2023. "Inference on Consensus Ranking of Distributions." *Journal of Business and Economic Statistics* XXX (XXX):XXX. URL https://doi.org/10.1080/07350015.2023.2252040.

Kaplan, David M. and Wei Zhao. 2023. "Comparing latent inequality with ordinal data." *Econometrics Journal* 26 (2):189–214. URL https://doi.org/10.1093/ectj/utac030.

Kaplan, David M. and Longhao Zhuo. 2021. "Frequentist properties of Bayesian inequality tests." *Journal of Econometrics* 221 (1):312–336. URL https://doi.org/10.1016/j.jeconom.2020.05.015.

Kline, Brendan. 2011. "The Bayesian and frequentist approaches to testing a one-sided hypothesis about a multivariate mean." *Journal of Statistical Planning and Inference* 141 (9):3131–3141. URL https://doi.org/10.1016/j.jspi.2011.03.034.

Lee, Jinkook. 2011. "Pathways from Education to Depression." *Journal of Cross-Cultural Gerontology* 26 (2):121–135. URL https://doi.org/10.1007/s10823-011-9142-1.

Lee, Sokbae, Oliver Linton, and Yoon-Jae Whang. 2009. "Testing for Stochastic Monotonicity." *Econometrica* 77 (2):585–602. URL https://www.jstor.org/stable/40263876.

Lehmann, E. L. and Joseph P. Romano. 2005a. "Generalizations of the Familywise Er-

ror Rate." *Annals of Statistics* 33 (3):1138–1154. URL https://projecteuclid.org/euclid.aos/1120224098.

———. 2005b. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 3rd ed. URL https://books.google.com/books?id=Y7vSVW3ebSwC.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6):1311–1334. URL https://doi.org/10.2307/2171738.

National Center for Health Statistics. 2019. "National Health Interview Survey." https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm. Public-use data file and documentation, accessed 2023.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2010. "Multiple testing." In *The New Palgrave Dictionary of Economics*, edited by Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan, online ed., 1–5. URL https://doi.org/10.1057/9780230226203.3826.

Seo, Juwon. 2018. "Tests of stochastic monotonicity with improved power." *Journal of Econometrics* 207 (1):53–70. URL https://doi.org/10.1016/j.jeconom.2018.04.004.

Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption." *Statistical Science* 32 (4):561–579. URL https://doi.org/10.1214/17-STS623.

van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press. URL https://books.google.com/books?id=UEuQEM5RjWgC.

Zhao, Wei. 2023. "Multiple Testing of a Function's Monotonicity." Working paper available at https://ideas.repec.org/p/umc/wpaper/2311.html.

A Proofs

A.1 Proof of Lemma 1

Proof. Under Assumption A1 and the central limit theorem (e.g., van der Vaart, 1998, Prop. 2.27), as $n_x \to \infty$,

$$\sqrt{n_x}(\hat{F}_x(y) - F_x(y)) \xrightarrow{d} \mathrm{N}\big(0, F_x(y)[1 - F_x(y)]\big). \tag{14}$$

Let $\hat{\boldsymbol{A}}$ be the estimator of vector $\boldsymbol{A}$ that contains all conditional CDFs,

$$\boldsymbol{A} \equiv (F_1(1), F_1(2), \ldots, F_1(J-1), \ldots, F_K(1), F_K(2), \ldots, F_K(J-1))'. \tag{15}$$

Under Assumption A1, by the continuous mapping theorem (CMT) (e.g., van der Vaart, 1998, Thm. 2.3),

$$\sqrt{n_1}(\hat{\boldsymbol{A}} - \boldsymbol{A}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{V}}), \tag{16}$$

where $\boldsymbol{0}$ is a vector of zeros and $\underline{\boldsymbol{V}}$ is the $(JK - K) \times (JK - K)$ block diagonal matrix

$$\underline{\boldsymbol{V}} \equiv \begin{bmatrix} \underline{\boldsymbol{V}}^1 & 0 & \ldots & 0 \\ 0 & \underline{\boldsymbol{V}}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underline{\boldsymbol{V}}^K \end{bmatrix} \tag{17}$$

where each $\underline{\boldsymbol{V}}^k$ is a $(J-1) \times (J-1)$ matrix with entry in row i, column j denoted

$$V_{ij}^k = \gamma_x F_x(\min\{i,j\})[1 - F_x(\max\{i,j\})].$$

We apply the delta method (e.g., van der Vaart, 1998, Thm. 3.1) to derive the asymptotic distribution of the full vector $\hat{\boldsymbol{\theta}}$ for testing the stochastic monotonicity inequalities. For $x \in \{1, 2, \ldots, K-1\}$ and $y \in \{1, 2, \ldots, J-1\}$,

$$\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_{J-1})', \tag{18}$$

where each $\hat{\boldsymbol{\theta}}_j \equiv (\hat{\theta}_{K-1,j}, \hat{\theta}_{K-2,j}, \ldots, \hat{\theta}_{1,j})$ and $\hat{\theta}_{x,y} \equiv \hat{F}_{x+1}(y) - \hat{F}_x(y)$ is the estimator of $\theta_{x,y}$ defined in (4). Then,

$$\sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{W}}), \quad \underline{\boldsymbol{W}} \equiv \underline{\boldsymbol{H}}' \underline{\boldsymbol{V}} \underline{\boldsymbol{H}}, \tag{19}$$

where $\underline{\boldsymbol{H}} = \frac{\partial}{\partial \boldsymbol{A}} \boldsymbol{\theta}'$ is the partial derivative of the vector $\boldsymbol{\theta}'$ with respect to $\boldsymbol{A}$ in (15), and $\underline{\boldsymbol{V}}$ is from (17). Note each element of $\underline{\boldsymbol{H}}$ is either -1, 0, or 1; more specifically, within each column of $\underline{\boldsymbol{H}}$ (the derivative of a particular $\theta_{x,y}$ with respect to $\boldsymbol{A}$), one element is -1, one element is 1, and the rest equal zero.

From (19), for each $\hat{\theta}_{x,y}$, using $n_1/n_x \to \gamma_x$ from A1,

$$\begin{aligned}
\sqrt{n_1}(\hat{\theta}_{x,y} - \theta_{x,y}) &= \sqrt{n_1}\{[\hat{F}_{x+1}(y) - \hat{F}_x(y)] - [F_{x+1}(y) - F_x(y)]\} \\
&= \overbrace{\sqrt{n_1/n_{x+1}}}^{\to\sqrt{\gamma_{x+1}}} \overbrace{\sqrt{n_{x+1}}[\hat{F}_{x+1}(y) - F_{x+1}(y)]}^{\text{use (14)}} - \overbrace{\sqrt{n_1/n_x}}^{\to\sqrt{\gamma_x}} \overbrace{\sqrt{n_x}[\hat{F}_x(y) - F_x(y)]}^{\text{use (14)}} \\
&\xrightarrow{d} \mathrm{N}(0, \sigma_{x,y}^2), \\
\sigma_{x,y} &\equiv \sqrt{\gamma_x F_x(y)[1 - F_x(y)] + \gamma_{x+1} F_{x+1}(y)[1 - F_{x+1}(y)]},
\end{aligned} \tag{20}$$

where the variances are summed because the subsamples for $X_i = x$ and $X_i = x + 1$ are independent given A1, and if $W \perp\!\!\!\perp Z$ then $\mathrm{Var}(W + Z) = \mathrm{Var}(W) + \mathrm{Var}(Z)$.

For the standard error of $\hat{\theta}_{x,y}$, which is asymptotically $\sigma_{x,y}/\sqrt{n_1}$ given (20), we use the estimator

$$\hat{s}_{x,y} \equiv \sqrt{\frac{(n_1/n_x)\hat{F}_x(y)(1 - \hat{F}_x(y)) + (n_1/n_{x+1})\hat{F}_{x+1}(y)(1 - \hat{F}_{x+1}(y))}{n_1}}. \tag{21}$$

By the consistency of the conditional CDF estimators and the CMT, as $n_1 \to \infty$,

$$\begin{aligned}
\sqrt{n_1}\hat{s}_{x,y} &= \sqrt{(n_1/n_x)\hat{F}_x(y)(1 - \hat{F}_x(y)) + (n_1/n_{x+1})\hat{F}_{x+1}(y)(1 - \hat{F}_{x+1}(y))} \\
&\xrightarrow{p} \sqrt{\gamma_x F_x(y)(1 - F_x(y)) + \gamma_{x+1} F_{x+1}(y)(1 - F_{x+1}(y))} = \sigma_{x,y},
\end{aligned} \tag{22}$$

the asymptotic standard deviation in (20). Informally, the standard error estimator $\hat{s}_{x,y}$ is "consistent," meaning formally that $\sqrt{n_1}\hat{s}_{x,y} \xrightarrow{p} \sigma_{x,y}$.

Substituting (21) into (7),

$$\hat{t}_{x,y}(\theta_{x,y}) = \frac{\hat{\theta}_{x,y} - \theta_{x,y}}{\hat{s}_{x,y}} = \frac{\sqrt{n_1}(\hat{\theta}_{x,y} - \theta_{x,y})}{\sqrt{(n_1/n_x)\hat{F}_x(y)(1-\hat{F}_x(y)) + (n_1/n_{x+1})\hat{F}_{x+1}(y)(1-\hat{F}_{x+1}(y))}}.$$

Thus, we can write the vector of properly centered t-statistics as $\hat{\boldsymbol{t}} = \sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\hat{\underline{\boldsymbol{S}}}$, where $\sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is the left-hand side of (19), and $\hat{\underline{\boldsymbol{S}}}$ is a diagonal matrix having elements $1/(\hat{s}_{x,y}\sqrt{n_1})$ matching the corresponding (x, y) from the $\boldsymbol{\theta}$ vector; equivalently, the row i, column j element of matrix $\hat{\underline{\boldsymbol{S}}}$ is $\mathbb{1}\{i = j\}\hat{W}_{ij}$. By (22), $\hat{\underline{\boldsymbol{S}}} \xrightarrow{p} \underline{\boldsymbol{S}}$, the diagonal matrix with corresponding (x, y) elements $1/\sigma_{x,y}$,

$$\underline{\boldsymbol{S}} \equiv \begin{bmatrix} 1/\sigma_{K-1,1} & 0 & \cdots & 0 \\ 0 & 1/\sigma_{K-2,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_{1,J-1} \end{bmatrix}. \tag{23}$$

By (19) and CMT, $\hat{\boldsymbol{t}} \equiv \sqrt{n_1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\hat{\underline{\boldsymbol{S}}} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \underline{\boldsymbol{S}}'\underline{\boldsymbol{W}}\,\underline{\boldsymbol{S}})$. □

A.2 Proof of Theorem 2

Proof. When testing original hypotheses $H_{0x,y}\colon F_{x+1}(y) - F_x(y) \leq 0$, the set of true hypotheses is $\mathcal{T} \equiv \{(x, y) : F_{x+1}(y) - F_x(y) \leq 0\}$. For any true $H_{0x,y}$,

$$\hat{t}_{x,y}(0) = \frac{\hat{F}_{x+1}(y) - \hat{F}_x(y)}{\hat{s}_{x,y}} \leq \frac{\hat{F}_{x+1}(y) - \hat{F}_x(y) - \overbrace{[F_{x+1}(y) - F_x(y)]}^{\leq 0 \text{ given true } H_{0x,y}}}{\hat{s}_{x,y}} \equiv \hat{t}_{x,y}(\overbrace{F_{x+1}(y) - F_x(y)}^{\theta_{x,y}}). \tag{24}$$

Thus,

$$\begin{aligned} \mathrm{FWER} &\equiv \mathrm{P}(\text{reject } H_{0x,y} \text{ for any } (x, y) \in \mathcal{T}) \\ &= \mathrm{P}(\hat{t}_{x,y}(0) > \hat{c}_\alpha \text{ for any } (x, y) \in \mathcal{T}) \\ &= \mathrm{P}(\max_{(x,y)\in\mathcal{T}} \overbrace{\hat{t}_{x,y}(0)}^{\text{use (24)}} > \hat{c}_\alpha) \end{aligned}$$

$$
\begin{aligned}
&\leq \mathrm{P}(\max_{(x,y)\in\mathcal{T}} \hat{t}_{x,y}(\theta_{x,y}) > \hat{c}_\alpha) \\
&\leq \mathrm{P}(\max_{\text{any } (x,y)} \hat{t}_{x,y}(\theta_{x,y}) > \hat{c}_\alpha) \\
&\to \mathrm{P}(\max_{\text{any } (x,y)} \overbrace{t_{x,y}}^{\text{from Lemma 1}} > c_\alpha) \\
&= \alpha, \qquad (25)
\end{aligned}
$$

where the second inequality holds because $\mathcal{T} \subseteq \{1,\ldots,K-1\}\times\{1,\ldots,J-1\}$ ("any" (x,y)), and the convergence holds by Lemma 1 and the CMT (because max is continuous), along with $\hat{c}_\alpha \xrightarrow{p} c_\alpha$, where $\hat{c}_\alpha$ was defined in (10) as the $(1-\alpha)$-quantile of the max of a random vector following the $\mathrm{N}(\mathbf{0}, \hat{\underline{\boldsymbol{\Sigma}}})$ distribution (with estimator $\hat{\underline{\boldsymbol{\Sigma}}}$) and c_α was defined in (9) as the $(1-\alpha)$-quantile of the max of $\mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}})$ (with true $\underline{\boldsymbol{\Sigma}}$). This last part follows because $\hat{\underline{\boldsymbol{\Sigma}}} \xrightarrow{p} \underline{\boldsymbol{\Sigma}}$ (given A1 and CMT) and applying the CMT (given that the max of a normal vector has a continuous, strictly increasing distribution function and thus continuous quantile function).

The asymptotic FWER bound for testing the reversed hypotheses $H^{\geq}_{0x,y}$: $F_{x+1}(y) - F_x(y) \geq 0$ follows essentially the same derivation with the same reason for each step below. Now defining $\mathcal{T}^{\geq} \equiv \{(x,y) : F_{x+1}(y) - F_x(y) \geq 0\}$,

$$
\begin{aligned}
\text{FWER} &\equiv \mathrm{P}(\text{reject } H^{\geq}_{0x,y} \text{ for any } (x,y) \in \mathcal{T}^{\geq}) \\
&= \mathrm{P}(\hat{t}_{x,y}(0) < -\hat{c}_\alpha \text{ for any } (x,y) \in \mathcal{T}^{\geq}) \\
&= \mathrm{P}(\min_{(x,y)\in\mathcal{T}^{\geq}} \hat{t}_{x,y}(0) < -\hat{c}_\alpha) \\
&\leq \mathrm{P}(\min_{(x,y)\in\mathcal{T}^{\geq}} \hat{t}_{x,y}(\theta_{x,y}) < -\hat{c}_\alpha) \\
&\leq \mathrm{P}(\min_{\text{any } (x,y)} \hat{t}_{x,y}(\theta_{x,y}) < -\hat{c}_\alpha) \\
&\to \mathrm{P}(\min_{\text{any } (x,y)} t_{x,y} < -c_\alpha) = \mathrm{P}(-\min_{\text{any } (x,y)} t_{x,y} > c_\alpha) = \mathrm{P}(\max_{\text{any } (x,y)} t_{x,y} > c_\alpha) = \alpha,
\end{aligned}
$$

also using the symmetry of the $\boldsymbol{t} \sim \mathrm{N}(\mathbf{0}, \underline{\boldsymbol{\Sigma}})$ distribution that implies $\max \boldsymbol{t} \overset{d}{=} -\min \boldsymbol{t}$. □

A.3 Proof of Theorem 3

Proof. For the inner CS, we want to derive the probability of the event that the inner CS is contained within $\mathcal{T}$, which is equivalently characterized as "$(x, y) \in \widehat{\mathcal{CS}}_{inner}$ only if $(x, y) \in \mathcal{T}$." With reversed hypotheses $H^{\geq}_{0x,y}\colon \theta_{x,y} \geq 0$ but $\mathcal{T} \equiv \{(x, y) : \theta_{x,y} \leq 0\}$,

$$\begin{aligned}
\mathrm{P}(\widehat{\mathcal{CS}}_{inner} \subseteq \mathcal{T}) &= \mathrm{P}(\text{reject any } H^{\geq}_{0x,y} \text{ only if } (x, y) \in \mathcal{T}) \\
&= \mathrm{P}(\text{reject any } H^{\geq}_{0x,y} \text{ only if } \theta_{x,y} \leq 0) \\
&= 1 - \mathrm{P}(\text{reject any } H_{0x,y} \text{ when } \theta_{x,y} > 0) \\
&\geq 1 - \mathrm{P}(\text{reject any } H_{0x,y} \text{ when } \theta_{x,y} \geq 0) \\
&= 1 - \mathrm{FWER} \geq 1 - \alpha + o(1)
\end{aligned}$$

because $\mathrm{FWER} \leq \alpha + o(1)$ by Theorem 2.

The outer CS contains all elements in the true set $\mathcal{T}$ if and only if we do not reject any original null hypothesis $H_{0x,y}\colon \theta_{x,y} \leq 0$ with $(x, y) \in \mathcal{T}$. Similar to the derivation above, the coverage probability is

$$\begin{aligned}
\mathrm{P}(\mathcal{T} \subseteq \widehat{\mathcal{CS}}_{outer}) &= \mathrm{P}(\text{no } H_{0x,y} \text{ rejected with } (x, y) \in \mathcal{T}\ ) \\
&= 1 - \mathrm{P}(\text{reject any } H_{0x,y} \text{ with } \theta_{x,y} \leq 0) \\
&= 1 - \mathrm{FWER} \\
&\geq 1 - \alpha + o(1)
\end{aligned}$$

again because $\mathrm{FWER} \leq \alpha + o(1)$ by Theorem 2. □

B General Health with Educational Attainment

Here we use the following variables from the NHIS 2019 data (National Center for Health Statistics, 2019). General health variable Y (`PHSTAT_A`) has categories "poor," "fair," "good," "very good," and "excellent," respectively (re)coded as 1, 2, 3, 4, and 5. The

education variable X is recoded based on `EDUC_A` as in Section 3. We restrict the age range to 30–65 years old as in Section 3. For simplicity, observations with missing values are dropped and weights are not used.

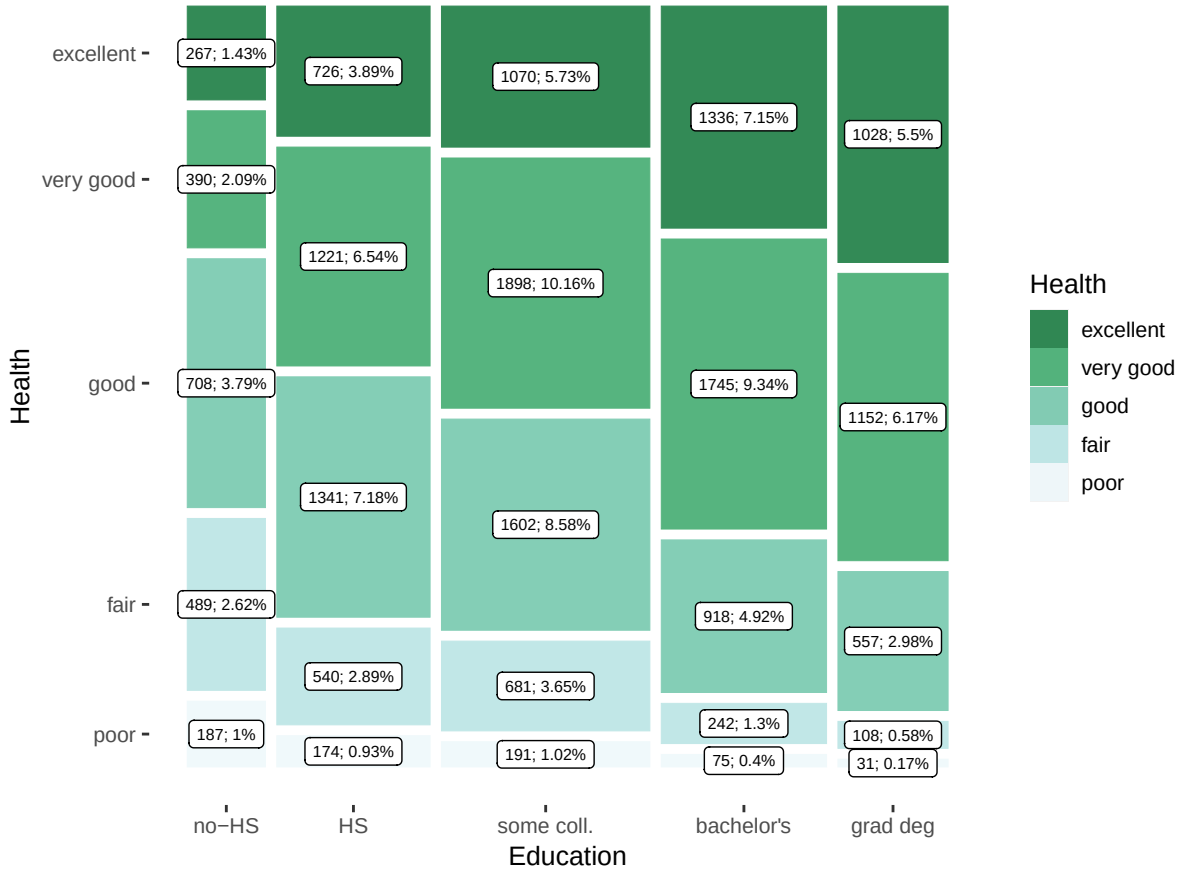

Figure 2: Mosaic plot of general health and education.

Figure 2 visualizes the data on general health and education in a mosaic plot similar to Figure 1. Here, $\hat{F}_5(y) \leq \hat{F}_4(y) \leq \hat{F}_3(y) \leq \hat{F}_2(y) \leq \hat{F}_1(y)$ for all values of y. That is, in the sample there is stochastic monotonicity: the distribution of general health is "better" at higher levels of education in the sense of first-order stochastic dominance. Our inference methods help assess the strength of the evidence that this same pattern holds in the population.

Table 3 has almost the same interpretation as Table 1 described in Section 3, with one exception. Before, better health meant lower Y, whereas here better health is higher Y, so now the set of points where health is higher at higher education is $\{(x, y) : F_x(y) \geq$

Table 3: Test statistics for general health versus education ($\hat{c}_{0.05} = 2.71$).

	Education category x			
General health y	1 (vs. 2)	2 (vs. 3)	3 (vs. 4)	4 (vs. 5)
1: poor	**−6.73**	−2.06	**−5.55**	−2.38
2: fair (or below)	**−12.68**	−2.32	**−13.64**	**−4.46**
3: good (or below)	**−12.65**	**−5.67**	**−17.48**	**−4.19**
4: very good (or below)	**−5.25**	−1.87	**−12.74**	**−4.21**

True set: points where general health is better at next-higher education level, $\{(x, y) : F_x(y) \geq F_{x+1}(y)\}$. Gray shading: outer confidence set. Bold: inner confidence set. Confidence level 95%. Critical value computed using N = 100,000 random draws.

$F_{x+1}(y)\}$, instead of $\leq$ like before. Correspondingly, now better health with higher education is indicated by a negative t-statistic. Thus, the inner confidence set corresponds to negative t-statistics below $-\hat{c}_{0.05}$, and the outer confidence set additionally allows $\hat{t}_{x,y} < \hat{c}_{0.05}$.

Table 3 shows an even stronger pattern than Table 1 of improved health with higher education. All but four of the 16 points are included in the inner confidence set, and the other four points are in the outer confidence set. The point at $(x, y) = (4, 1)$ involves comparing the two very smallest subsamples, so it is not surprising there is more uncertainty about the corresponding population comparison. The less-negative t-statistics at (x, y) values $(2, 1)$, $(2, 2)$, and $(2, 4)$ reflect the closer empirical conditional CDF values for high-school versus some-college, the same education level comparison that in Section 3 showed the opposite results of the other education level comparisons.

Overall, our results show strong evidence that general health is stochastically increasing in education for most education levels, and somewhat weaker evidence that general health for the some-college subpopulation stochastically dominates general health of the only-HS subpopulation. Compared to not rejecting the global null that health stochastically increases with education, our methodology provides much stronger and more precise results.