Distribution-Free Assessment of Population Overlap in Observational Studies
Libra Lei (Stanford), Alexander D’Amour (Google Brain), Peng Ding (UC Berkeley), Avi Feller (UC Berkeley), and Jaspreet Sakhun (Yale)

Main (and perhaps surprising) contribution:
We develop distribution-free O-values that are valid in finite samples!

Population overlap in observational studies
Setting: binary treatment T, baseline covariates X (arbitrary), (T, X) simulate (T, X) (the only assumption!)
Strict overlap condition: \( \exists \theta_0 > 0, \quad \theta_0 \leq \theta(X) \iff \theta_0 \leq P(T = 1 | X = x) \leq 1 - \theta_0 \) a.s.
Population overlap slack: \( \theta^* = \min \{ \min(n_1, n_0), \theta_0 \} \)

Current approaches for assessing overlap:
- Informal comparisons or plug-in estimates based on estimated propensity scores
  - useful but not statistically rigorous
- “sample overlap” ≠ population overlap
  - sensitive to model mis-specification or finite sample errors
- Standard two-sample test: testing the wrong null
  \( H_0: P(X | T = 1) = P(X | T = 0) \Rightarrow H_1: \theta(X) \neq \theta_0 \)

Major challenge: \( \theta^* \) is irregular (extreme of an unknown function)

O-value
Definition. \( \hat{D} \) is an O-value if it is an upper confidence bound of \( \theta^* \), i.e. \( P(\theta^* \leq \hat{D}) \geq 1 - \alpha \)
Analogous to p-value:
- A small \( \hat{D} \) provides strong evidence against overlap
- A large \( \hat{D} \) does not necessarily imply sufficient overlap

Some practical implications:
- Strict overlap condition as a composite null hypothesis: reject if \( \hat{D} < \theta_0 \)
  \( (1 - \min(n_1, n_0)) \) estimates efficiency loss caused by the imbalance
- Assessing if trimming (say, at 0.1) is successful by comparing \( \hat{D} \) with 0.1
- Comparing different matches based on \( \hat{D} \)

Main (and perhaps surprising) contribution:
We develop distribution-free O-values that are valid in finite samples!

Step I: covariate standardization
Key observation: overlap is preserved under transformation of \( X \)
\( \theta_0 \leq \theta(X) \iff \theta_0 \leq P(T = 1 | s(X) = s) \leq 1 - \theta_0 \)
for any fixed function \( s() \)
- Let \( \theta^*_T \) be the population overlap slack for \( (T, s(X)) \)
- \( \theta^* = \theta^*_T \)

Data splitting guarantees that \( \hat{\theta}(\cdot) \perp \perp \) (second half of data)
\( \theta^* \leq \theta^*_T \) always holds even if \( \hat{\theta} \) is bad; tight if \( \hat{\theta} \) is good

From now on, we assume that \( (S_i, T_i) \) are i.i.d. with \( S_i = \delta(X_i) \in \{0, 1\} \)
Goal: construct upper confidence bounds on \( \theta^*_T \) (w/ standardized covariates)

Step II: careful balance check
Key observation: overlap \( \Rightarrow \) bounded likelihood ratio
\( \hat{b}_{\min}(\theta^*_T; \pi) \leq \frac{dP(T = 1)}{dP(T = 0)}(s) \leq \hat{b}_{\max}(\theta^*_T; \pi), \quad \forall s \in [0, 1], \)
where \( \pi = P(T = 1), \quad \hat{b}_{\min}(\theta^*_T; \pi) = \frac{\theta^*_T}{1 - \theta^*_T} \pi \quad \hat{b}_{\max}(\theta^*_T; \pi) = \frac{1 - \theta^*_T}{\theta^*_T} \pi \)

Intuition: larger \( \theta^* \) ⇒ smaller discrepancy between \( P_0 \) and \( P_1 \)

A generic strategy:
- Find an estimable “discrepancy” \( \Delta(P_0, P_1) \) and \( B_d(\theta^*_T) \)
- Compute a lower confidence bound on \( \Delta(P_0, P_1) \)
  \( \Delta(P_0, P_1) \leq B_d(\theta^*_T) \)
- \( \hat{D} = B_d(\hat{\Delta}) \) is a valid O-value:
  \( P(\hat{D} \geq \theta^*_T) = P(\hat{\Delta} \leq B_d(\theta^*_T)) \geq P(\hat{\Delta} \leq \Delta(P_0, P_1)) \geq 1 - \alpha \)

Summary of DiM, DiT, DiR, CE O-values

\[\Delta \quad B_d(\theta^*_T) \quad \Delta_d \quad \text{Two-sample test analyzation}\]

<table>
<thead>
<tr>
<th>DiM</th>
<th>T-stat</th>
<th>( x^2 \text{-divergence} )</th>
<th>Hedged capital bound (National Supported Work Demonstration Program [96])</th>
<th>T-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiT</td>
<td>LR</td>
<td>Simple algebra</td>
<td>Line-crossing (DeHeide [34]); Simes’ inequality [34]; Kolmogorov–Smirnov test</td>
<td></td>
</tr>
<tr>
<td>DiR</td>
<td>AUC</td>
<td>Generalized Neyman–Pearson</td>
<td>Hybrid bound for U-statistic (DeHeide [34]; Leurgans et al. [11])</td>
<td>Wilcoxon rank-sum test</td>
</tr>
<tr>
<td>CE</td>
<td>class</td>
<td>Formula of error</td>
<td>Same as DiT O-values (Classification-based test)</td>
<td></td>
</tr>
</tbody>
</table>

O-values for Latent data
- National Supported Work Demonstration program [96]
- Treatment group has \( n_t = 185 \) units
- 7 control groups: 6 from observational studies, 1 from an RCT
- Apply gradient boosting for DiT O-values

Sample property: Main (and perhaps surprising) contribution:

\[ \text{Proposition (Empirical Bernstein’s inequality), } \forall Z_1, \ldots, Z_n \in [0, 1] \text{ be i.i.d. with } EZ = \mu, \text{Var}(Z) = \sigma^2. \text{ Then with probability } 1 - \delta, \]
\[ |\mu - \mu| \leq \frac{2 \log \frac{n}{\delta}}{n} \quad \text{and} \quad \sigma - \delta \leq \frac{2 \log \frac{n}{\delta}}{n - 1} \]
\[ \Rightarrow \text{a lower confidence bound on } T_0 \text{ (with Bonferroni correction over } \{\mu_1, \mu_0, \sigma_0\}): \]
\[ \hat{D}_{\text{ Bonferroni}} = \sup_{\delta} \frac{1}{\delta} \left( \frac{1}{\delta} \frac{\pi(1 - \pi)}{\pi T_0} + 1 \right) \]

Theorem. With solely the i.i.d. assumption, \( P(\theta^*_T \geq \hat{D}_{\text{ Bonferroni}}) \geq 1 - \alpha \)

An illustrative simulation study
- \( X \sim N(0, 1^2) \) with \( p \in \{10, 30, 100\} \)
- \( e(x) = f(x; \beta) \)
  \( f(y) = \begin{cases} 0.1 & (y < 0) \\ 0.9 & (y > 0) \end{cases} \)
- \( \beta \) sparse; \( \epsilon \) is chosen such that \( P(e(X) = 0.1) = 0.8 \)

O-value
Practical recommendation: based on extensive numerical experiments
- Algorithm to estimate propensity scores: gradient boosting
- Type of O-value: DiT

<table>
<thead>
<tr>
<th>( n_t )</th>
<th>( n_c )</th>
<th>( \theta^*_T )</th>
<th>( \Delta_d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2922</td>
<td>0.003</td>
<td>0.75</td>
<td>2400 0.018 0.75</td>
</tr>
<tr>
<td>2369</td>
<td>0.021</td>
<td>0.75</td>
<td>253 0.234 0.45</td>
</tr>
<tr>
<td>129</td>
<td>0.143</td>
<td>0.75</td>
<td>128 0.313 0.23</td>
</tr>
</tbody>
</table>