Discussion points

1. Selection bias as a general lens.

2. Adaptivity.

3. Information usage.

Selection bias

- Dataset $D \sim \mathcal{P}$.
- Set of tests/hypotheses $\{\phi_1(D), ..., \phi_m(D)\}$.
- Some selection protocol $T : D \to i$.
- Bias due to selection: $|\phi_T - \mu_T|$, where $\mu_i = \mathbb{E}[\phi_i]$.

Selection bias

- Dataset $D \sim \mathcal{P}$.
- Set of tests/hypotheses $\{\phi_1(D), ..., \phi_m(D)\}$.
- Some selection protocol $T : D \to i$.
- Bias due to selection: $|\phi_T - \mu_T|$, where $\mu_i = \mathbb{E}[\phi_i]$.

Example 1. Ordered hypothesis testing (Rina).

- ϕ_i is the p value distribution of the first i hypotheses.
- T is the protocol that uses the accumulation function for deciding which first k hypotheses to report.

Selection bias

- Dataset $D \sim \mathcal{P}$.
- Set of tests/hypotheses $\{\phi_1(D), ..., \phi_m(D)\}$.
- Some selection protocol $T : D \to i$.
- Bias due to selection: $|\phi_T - \mu_T|$, where $\mu_i = \mathbb{E}[\phi_i]$.

Example 2. Data carving (Will).

- Each i index a subset of covariates and ϕ_i is the coefficients of a model using just these covariates.
- T is Lasso and selects a subset of covariates.

Selection bias

- Dataset $D \sim \mathcal{P}$.
- Set of tests/hypotheses $\{\phi_1(D), ..., \phi_m(D)\}$.
- Some selection protocol $T : D \to i$.
- Bias due to selection: $|\phi_T - \mu_T|$, where $\mu_i = \mathbb{E}[\phi_i]$.

Example 3. Adaptive queries (Cynthia + Jon).

- $\{\phi_i\}$ are all the possible queries you can make on the data.
- T is an interactive protocol that involves k rounds of adaptive queries and decides which ϕ_i to report.

How adaptive is the selection protocol?

- In FDR control and data carving settings, the analyst decides on an analysis protocol ahead of time.
- One round of selection.
- Very powerful and crisp analysis for specific settings.
  -- most powerful tests.
  -- explicit correction for bias.

How adaptive is the selection protocol?

- In FDR control and data carving settings, the analyst decides on an analysis protocol ahead of time.

- One round of selection.

- Very powerful and crisp analysis for specific settings.
  -- most powerful tests.
  -- explicit correction for bias.

Interesting challenge: what if the analyst deviates slightly from the pre-determined protocol.

Information usage of selection

- When T uses information specific to the realized dataset D, then we are at risk for bias.
- Differential privacy: control information leakage.
- Data carving: use the left-over information from selection stage.
- Ordered hypothesis testing: use side information that's independent of realized data D.

Information usage of selection

- How to quantify and measure the information usage? (Not all information usage is harmful!)
- (approximate) max-information: powerful controls on probability of bad events.
- mutual information (joint work with Dan Russo), etc.