Robust Statistics

Main idea: parameter estimate maintains approximate optimal performance under "small" perturbations in a "neighborhood" of an assumed model

Agenda

- Measuring Robustness
  - Sensitivity Curve
  - Influence Function
  - Breakdown Point
- Propublica Dataset Experiments
  - Mapping to fairness metrics
  - Drop row results + thoughts
- Next Steps?

Sensitivity Curve

- How much an estimate changes by adding an additional point **x** where **x** ranges in value

$$S(x) = \hat{\mu}(x_1, \ldots, x_n, x) - \hat{\mu}(x_1, \ldots, x_n)$$

Sensitivity Curve

- How much an estimate changes by adding an additional point **`x`** where **`x`** ranges in value

$$S(x) = \hat{\mu}(x_1, ..., x_n, x) - \hat{\mu}(x_1, ..., x_n)$$

- How much an estimate changes by adding **`m`** additional outliers of value **`x`**

$$S(m) = \hat{\mu}(x, ..., x, x_{m+1}, ..., x_n) - \hat{\mu}(x_1, ..., x_n)$$

Influence Function

- Asymptotic version of sensitivity curve
  - Sample size tends to infinity
  - ε contamination approaches 0, where ε is m / n

Influence Function

- Asymptotic version of sensitivity curve
  - Sample size tends to infinity
  - ε contamination approaches 0, where ε is m / n

- Assume we have a distribution **F** that is approximately known, and we are interested in the behavior of a metric over a “neighborhood”

$$(1-\epsilon)F + \epsilon G : G \in \mathcal{G}$$

$\mathcal{G}$ is the set of distributions, can be set of point mass distributions

Influence Function

x: some point
T: some statistical estimator
F: some distribution

Influence Function

x: some point
T: some statistical estimator
F: some distribution

$$IF(x,T;F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\,\delta_x) - T(F)}{\epsilon}$$

Influence Function

x: some point
T: some statistical estimator
F: some distribution

$$IF(x, T; F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\,\delta_x) - T(F)}{\epsilon}$$

*Dwork uses this as a starting point to derive some nice guarantees with respect to epsilon, delta differential privacy

Breakdown Point

- Largest proportion of atypical points that the data may contain such that an estimator $\hat{\theta}$ still gives some information about the actual distribution θ

Breakdown Point

- Largest proportion of atypical points that the data may contain such that an estimator $\hat{\theta}$ still gives some information about the actual distribution θ

- In order for the estimate to give some info, the contamination should not be able to drive the estimator to infinity or to some other boundary

$$T((1-\epsilon)F + \epsilon\,\delta_x) \in K, \forall\epsilon < \epsilon^*, \forall\delta_x$$

where K is some bounded set

Mapping to Statistical Parity

- Motivation: fairness metrics should also maintain approximate optimal behavior under small perturbations of assumed model
  - Data is dirty – entry errors, dropped and modified values during cleaning

Mapping to Statistical Parity

- Motivation: fairness metrics should also maintain approximate optimal behavior under small perturbations of assumed model
  - Data is dirty – entry errors, dropped and modified values during cleaning

- What is an "outlier" for a binary value? A flipped decision far from boundary?
  - Flip outcome label for predicted dataset or flip outcome label for training dataset?

- Can we represent robustness of these fairness metrics in terms of breakdown point and influence function?

Propublica

- Dataset Statistics: 5278 rows, 12 columns

African American: 3175	Caucasian: 2103
High: 845	High: 223
Medium: 984	Medium: 473
Low: 1346	Low: 1407

- Adding rows to an existing dataset doesn’t make sense – but removing does

Propublica Experiment Setup

Output label: score_text
Label target: “High”
Protected attribute: race

delta: proportion of overall population of 5278
epsilon: measured as absolute value of

P[score_text = “High” | race = African American] –
P[score_text = “High” | race = Caucasian]

Propublica Experiment Setup

Output label: score_text
Label target: “High”
Protected attribute: race

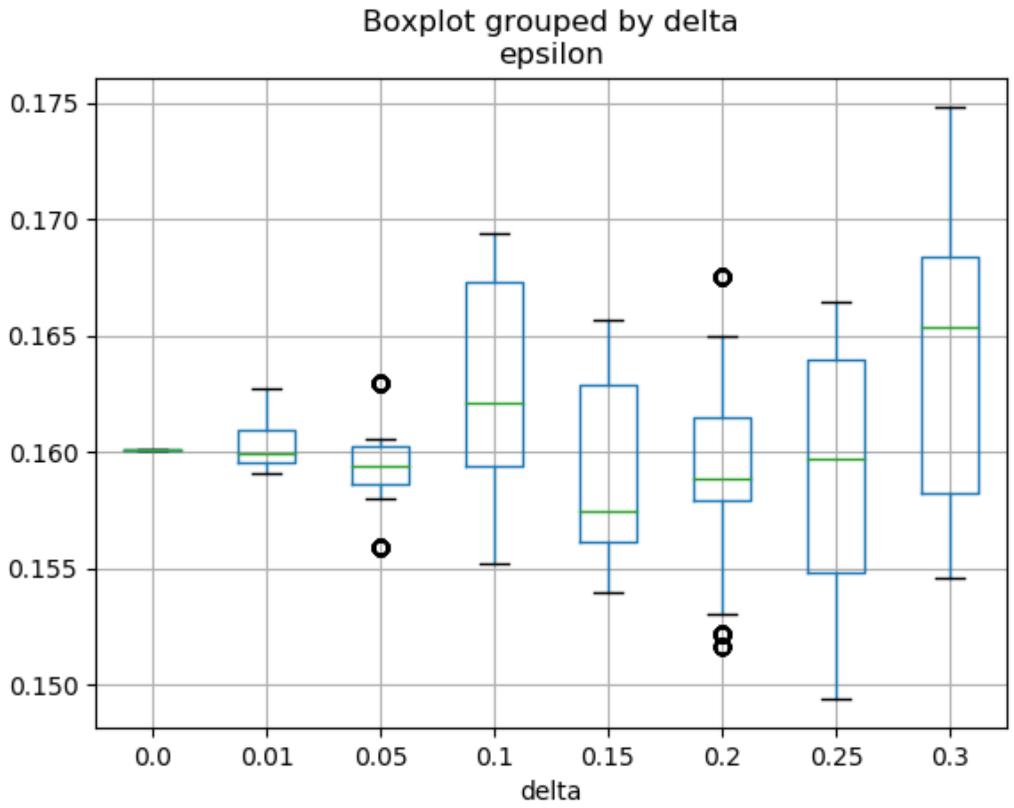
delta: proportion of overall population of 5278
epsilon: measured as absolute value of

P[score_text = “High” | race = African American] –
P[score_text = “High” | race = Caucasian]

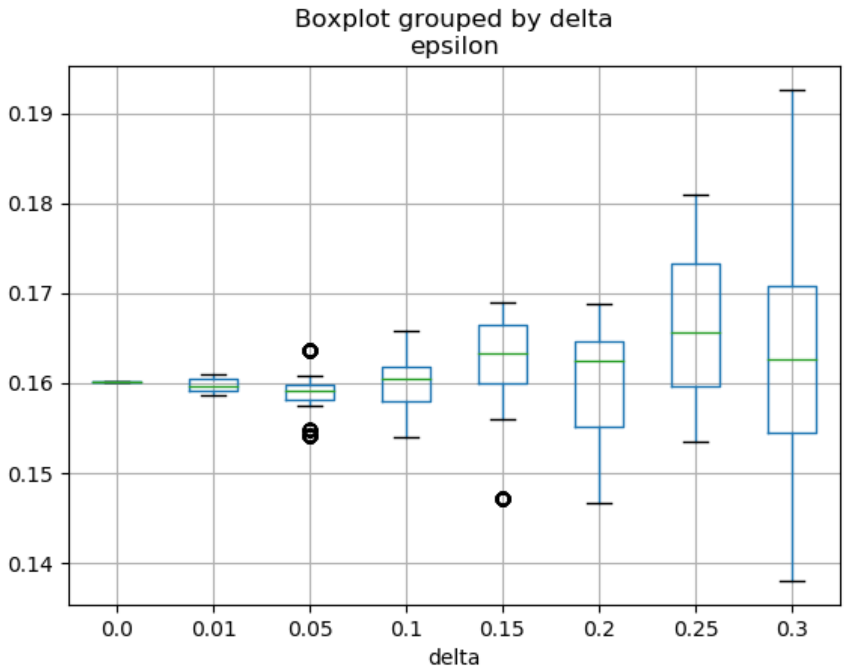
Procedure:

1. Drop rows at random
2. Drop rows at random w/in Caucasian community only
3. Drop rows at random w/in African American community only

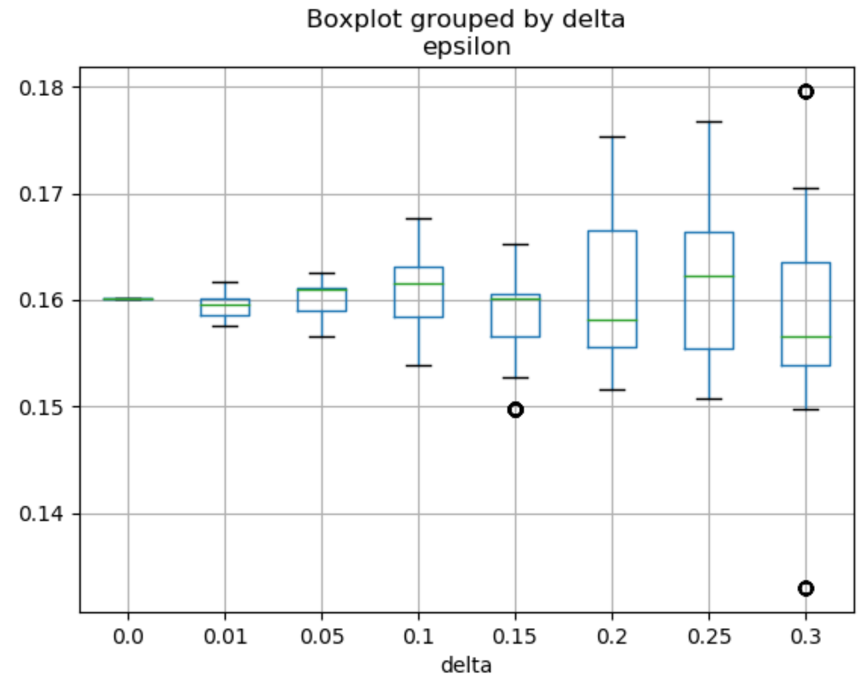
DELETE delta proportion rows at RANDOM



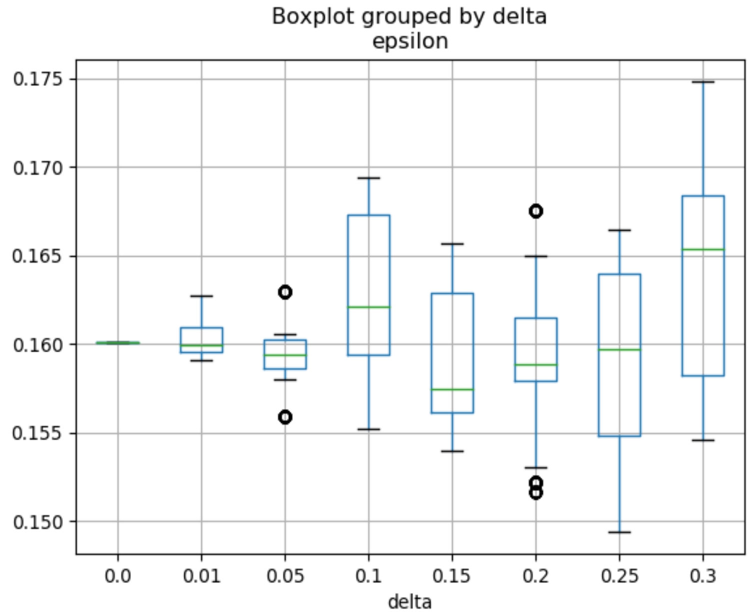
DELETE delta proportion Caucasian rows at RANDOM



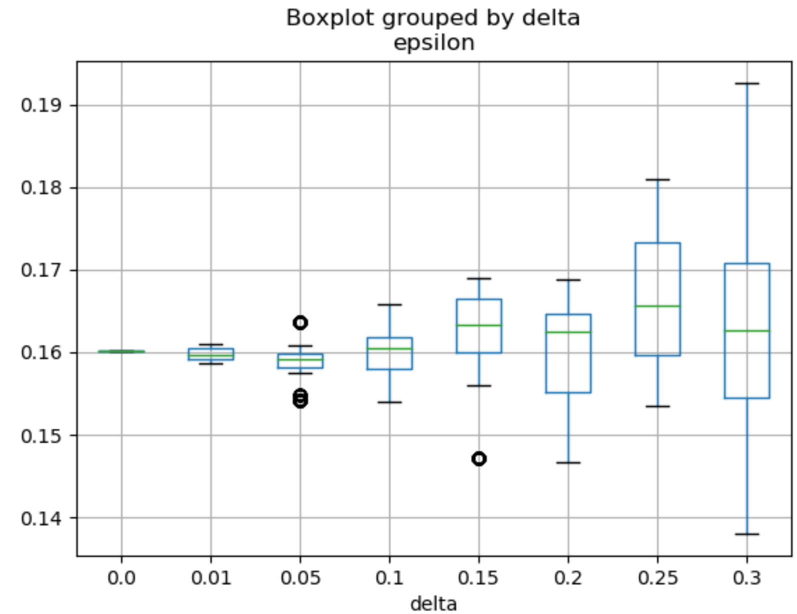
DELETE delta proportion African-American rows at RANDOM



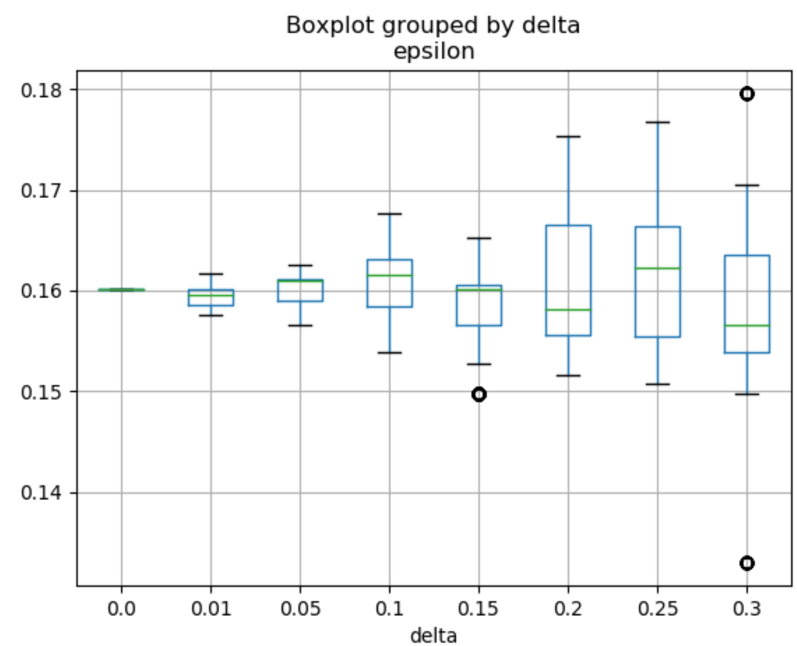
DELETE delta proportion rows at RANDOM



DELETE delta proportion Caucasian rows at RANDOM



DELETE delta proportion African-American rows at RANDOM



More specific subgroup targeting: epsilon values as delta increases

delta

	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Low//Caucasian	0.16	0.16	0.15	0.15	0.15	0.14	0.14	0.14	0.13	0.12	0.12
High//African-American	0.16	0.15	0.13	0.12	0.10	0.09	0.07	0.05	0.04	0.02	0.00
Both	0.16	0.15	0.14	0.14	0.13	0.12	0.11	0.10	0.09	0.08	0.07

Next Steps?

- Goal is to show which metrics are robust – intuition is that statistical parity probably isn't because it's a conditional probability, thus, the metric is sensitive to subgroup size

- Synthetic dataset experiments
  - For 4 subgroups (2 attributes), try different combinations of equal/unequal populations, equal/unequal probability of outcome, and satisfying/not satisfying statistical parity

- Is this an interesting problem?