# Opportunity Insights Data Task
# October 2021

Please complete the tasks below to the best of your ability. While we prefer that you use Stata, you are free to use your preferred statistical software. Please complete the tasks within the allotted time as indicated in the email.

Please send us your answers and graphs in a PDF file, along with the code file which you ran to get your results. We shouldn't have to run the code to see your results; include all graphs and answers in the PDF file. Make graphs look sufficiently professional to include in an academic presentation (clearly labelled axes, legends, etc). Your code should be carefully commented so that we can read it easily.

Good luck!

## Task 1

Merge together the "county_outcomes" and "cty_covariates" datasets.  Present a graph overlaying two binned scatter plots – the first showing the relationship between average future family income for kids with parents at the 25th percentile of the national income distribution and the single parent share in their county, and the second showing the same relationship for kids at the 75th percentile. Weight each binned scatter plot by the number of kids in the county and display the respective correlation coefficients from the underlying data on the graph. **Note:** A binned scatter plot partitions the data into bins depending on the x-variable, and then presents a scatter plot of the mean of the x-variable and y-variable in each of these bins. If you are using Stata, the "binscatter" package may be helpful. For R, you can use the "binsreg" package.

## Task 2

For this task, please use unweighted correlations throughout.

a) Using a format of your choice, display the county-level correlations between family income for kids with parents at the 25th/75th percentile and the following list of variables:
   - Single parent share
   - Median household income
   - Poverty Rate
   - Employment Rate
   - Population Density

b) Note that the dataset contains the county-level standard errors for each of the following variables:
   - Average future family income for White kids with parents at the 25th/75th percentile.
   - Average future family income for Black kids with parents at the 25th/75th percentile.

- Average future family income for Asian kids with parents at the 25th/75th percentile.
- Average future family income for Hispanic kids with parents at the 25th/75th percentile.

Calculate the average squared county-level standard error for each variable. Letting this average squared standard error represent the magnitude of the noise component of the variance in each variable, calculate the reliability of each of the variables listed in part (b). Present these eight reliability statistics in a table and provide a brief intuitive explanation for why some variables have higher reliability than others.
**Note:** The reliability of a variable is the proportion of the variation in that variable that is driven by true underlying variation (signal variation), as opposed to the variation caused by noise. To calculate it for each variable, divide the noise component of the variance by the overall variance, and subtract this quantity from 1.

c) Calculate the signal correlations between all the variables listed in part (b), presenting your results in a table. To do this, divide each of the raw correlations by the product of the square roots of the two relevant reliabilities. **Note:** The signal correlation is the correlation between two hypothetical versions of the variables without measurement error.

d) Describe how the signal correlations are different from the raw correlations and provide a short explanation for why this is the case.

## Task 3

To protect the confidentiality of individuals in our datasets, we follow the differential privacy literature and apply noise to statistics we publish.

A key concept in differential privacy is the global sensitivity of the function – that is, the maximum amount that the function can change if you alter one observation in any theoretically possible dataset. For example, suppose your dataset consists of N observations of a variable X, where X is bounded between 0 and 100. Then the global sensitivity of the sum of these observations is 100, which is achieved either by changing an observation of 100 to 0, or an observation of 0 to 100. For the rest of the question, continue to assume that you observe N draws of X, where X is bounded between 0 and 100.

(a) What is the global sensitivity of the sample mean in this case? Take N as given.
(b) What is the global sensitivity of the sample median?
(c) What is the global sensitivity of the sample variance?
(d) Download the “inventor_rate_by_college” dataset. Building on your answer from (a), write code to make the “inventor rate” variable differentially private at the level epsilon = 1. You should not need to submit this new dataset, just the code. **Note:** epsilon is a tuning parameter controlling the level of privacy protection. To make a dataset differentially private at the level epsilon = 1, you need to add noise from a Laplace distribution with location 0 and scale equal to the global sensitivity divided by epsilon.
(e) Assume that there is no sampling variation in the underlying data. Produce a graph showing the tradeoff between epsilon and top-tail accuracy (the proportion of observations in the top 10% of the new dataset that were also in the top 10% of the underlying data) for values of epsilon between 0.01 and 1.