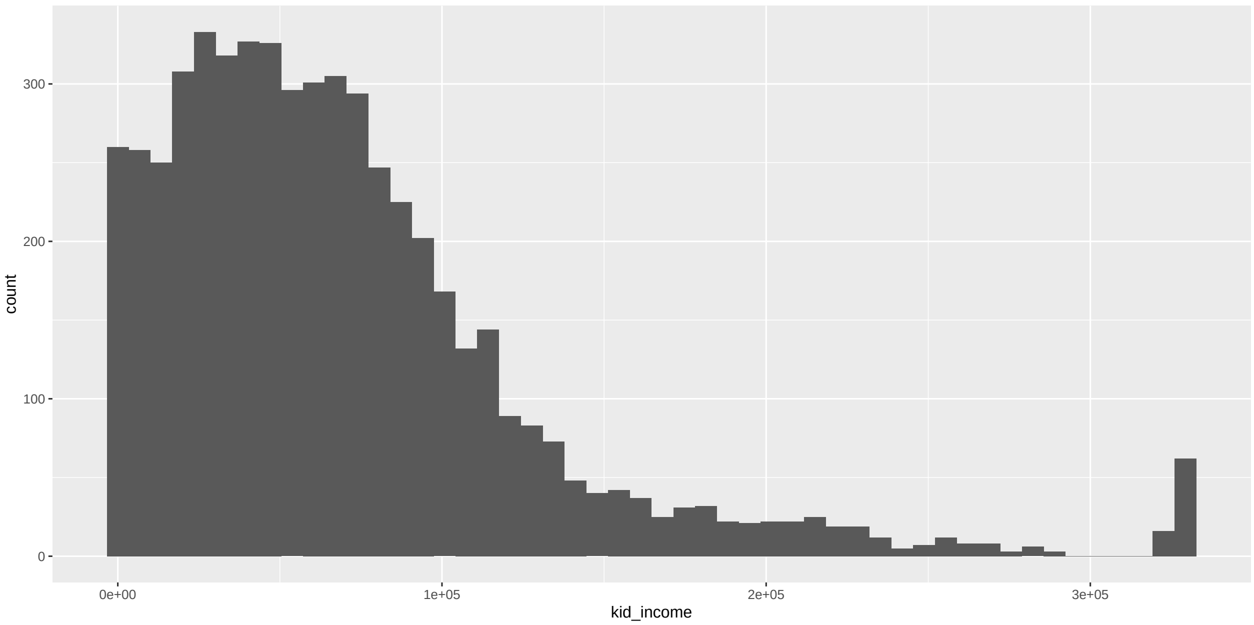
HKS SUP-135 Lab 1: Introductory Statistical Concepts and Statistical Computing

Matt Khinda

2/3/2023

Question 1: Histogram



Question 2: Mean

```
## The mean income for the sample is $70499.94
```

Question 3: Conditional Variables

3a: Below the mean

3b: Percent below mean

```
## The percent of children below the mean income is 59.60627%
```

3c: Why is it not 50%?

```
## Because the incomes (shown in the histogram above) are not evenly distributed.
```

Question 4: Median

```
## The median income is $58750
```

Question 5: Standard Deviation

```
## One stanard deviation is equal to $59552.02
```

Question 6: Within 1 or 2 Standard Deviations

```
## The percent of children within one standard deviation is 78.67299%.
```

```
## The percent of children within two standard deviations is 94.8961%.
```

Question 7: Percentile Ranks

7a: Rank incomes

7b: Sort by rank

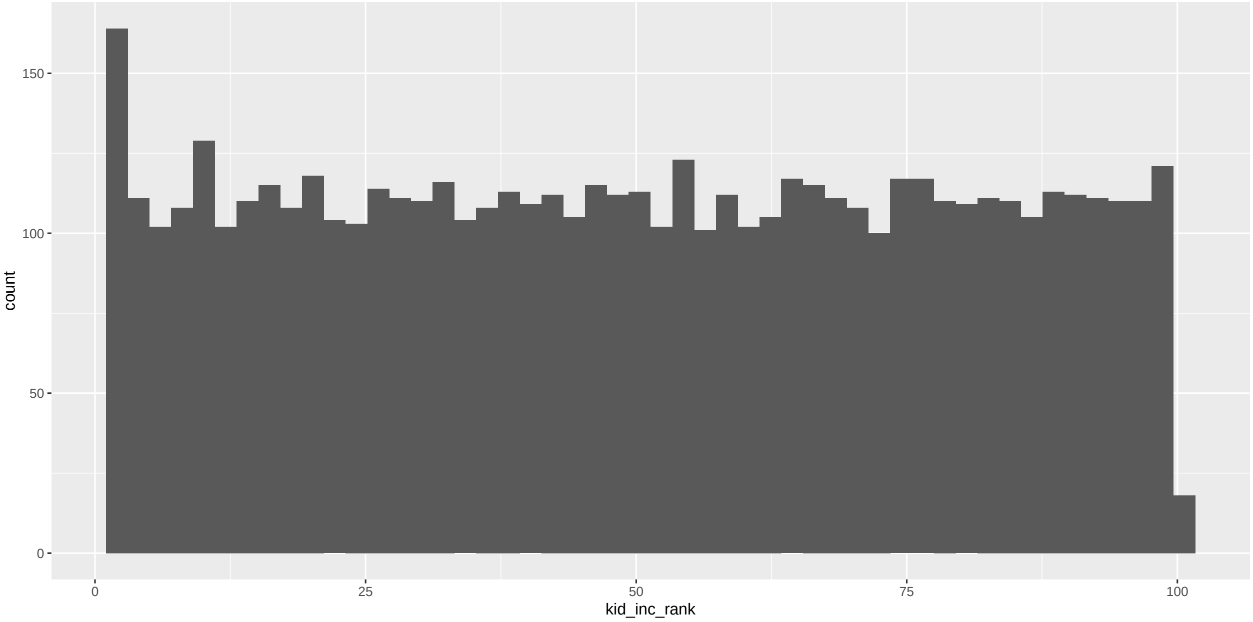
```
## # A tibble: 5,486 x 20
##    id_num kid_i~1 incar~2 child~3 child~4 child~5 paren~6 mothe~7 fathe~8 female
##     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>
##  1     46       0       0      12       0      NA       0       8      12      0
##  2     80       0       0      11       0      NA   30000      12      12      0
##  3     92       0       0      12       0      NA   37600      10      12      0
##  4    227       0       1       8       0      NA    6014      14      12      1
##  5    344       0       0      12       0      NA   25000      12      16      0
##  6    452       0       0      10       0      NA   19500      12      12      1
##  7    453       0       0       7       0      NA   19500      12      12      1
##  8    570       0       0      11       0      NA   96000      12      12      1
##  9    710       0       0      17       1    1000   55300      13      14      1
## 10    817       0       1      10       0      NA    8594      14      12      0
## # ... with 5,476 more rows, 10 more variables: black <dbl>, hispanic <dbl>,
## #   white <dbl>, region <dbl+lbl>, age2015 <dbl>, cohort <dbl>,
## #   below_mean <dbl>, sd1 <dbl>, sd2 <dbl>, kid_inc_rank <dbl>, and abbreviated
## #   variable names 1: kid_income, 2: incarcerated, 3: child_education,
## #   4: child_college, 5: child_sat, 6: parent_inc, 7: mother_education,
## #   8: father_education
```

7c: Normalize rank

7d: Browse the data

Question 8: Percentile Rank Distribution

8a: Plot percentile rank distribution

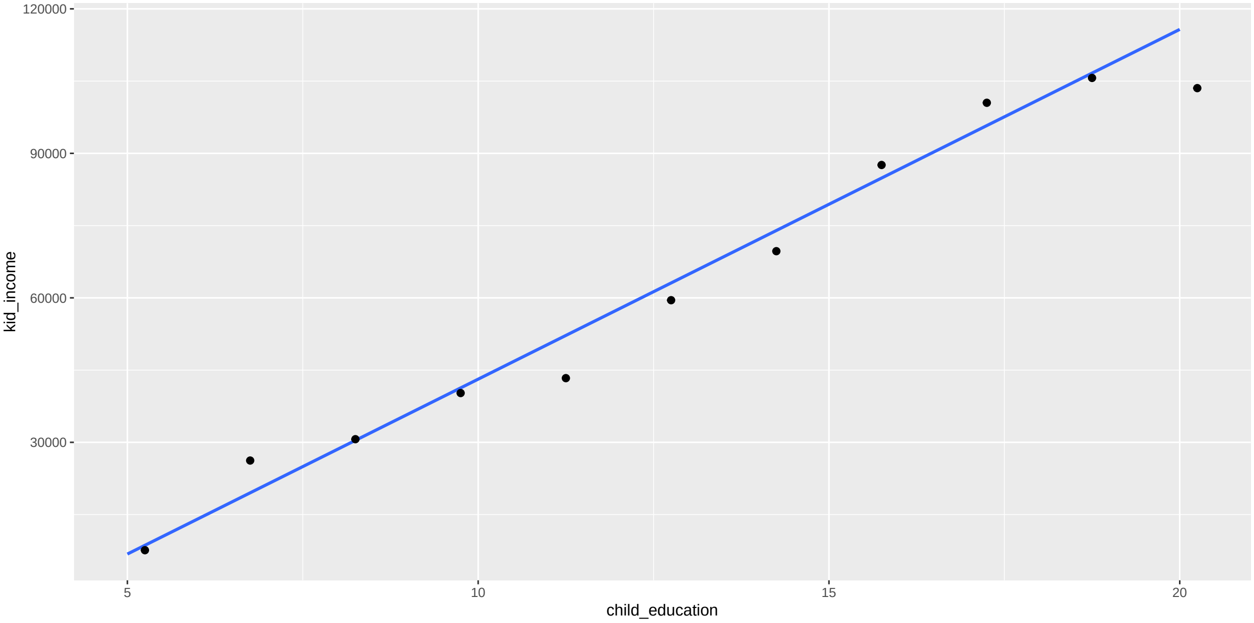


8b: Validate percentile rank mean and median

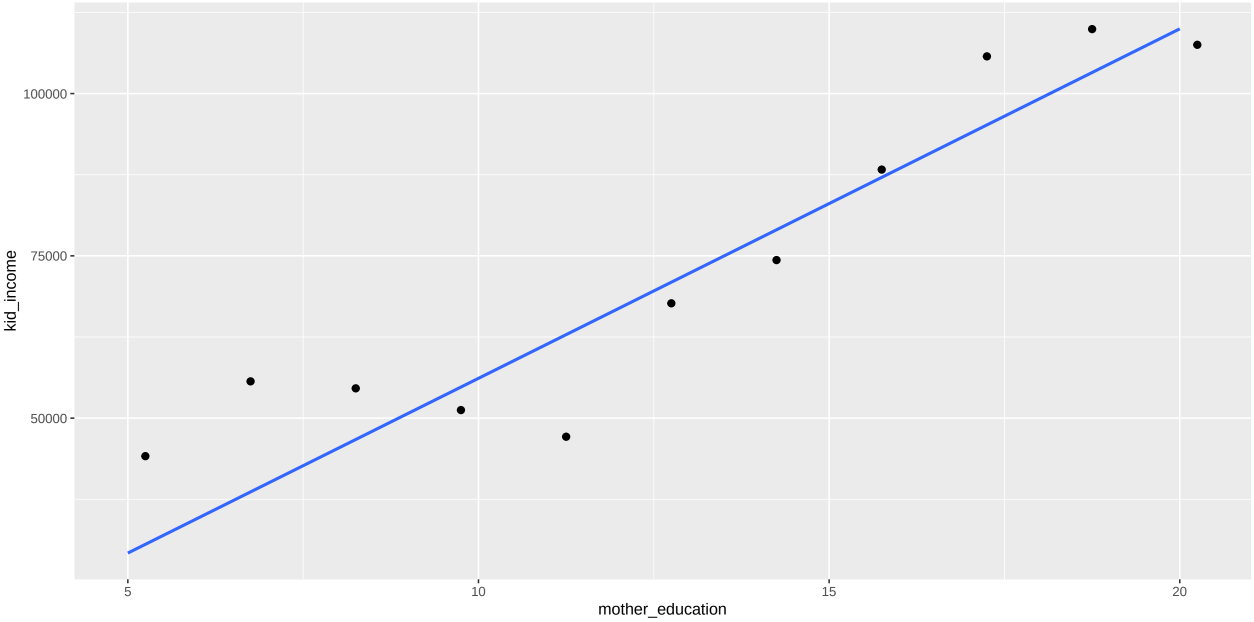
```
## The mean percentile rank is 50.08672  while the median percentile rank is 50.1141
```

Question 9: Relationships

Linear correlation



Non-linear correlation



Question 10: Randomization

10a: Generate and assign random values

10b: Determine treatment group status

```
## There are 1684 observations in the treatment group and 1649 observations in the control group.
```

10c: Treatment Group

```
## # A tibble: 1 x 44
##   id_num_mean kid_inco~1 incar~2 child~3 child~4 child~5 paren~6 mothe~7 fathe~8
##         <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1       2899.     78437.  0.0962    14.0   0.327      NA  53605.    13.1    13.1
## # ... with 35 more variables: female_mean <dbl>, black_mean <dbl>,
## #   hispanic_mean <dbl>, white_mean <dbl>, region_mean <dbl>,
## #   age2015_mean <dbl>, cohort_mean <dbl>, below_mean_mean <dbl>,
## #   sd1_mean <dbl>, sd2_mean <dbl>, kid_inc_rank_mean <dbl>,
## #   rand_val_mean <dbl>, treatment_group_mean <dbl>, id_num_sd <dbl>,
## #   kid_income_sd <dbl>, incarcerated_sd <dbl>, child_education_sd <dbl>,
## #   child_college_sd <dbl>, child_sat_sd <dbl>, parent_inc_sd <dbl>, ...
```

Control Group

```
## # A tibble: 1 x 44
##   id_num_mean kid_inco~1 incar~2 child~3 child~4 child~5 paren~6 mothe~7 fathe~8
##         <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1       2873.     78001.  0.0837    14.1   0.355      NA  54651.    12.9    13.0
## # ... with 35 more variables: female_mean <dbl>, black_mean <dbl>,
## #   hispanic_mean <dbl>, white_mean <dbl>, region_mean <dbl>,
## #   age2015_mean <dbl>, cohort_mean <dbl>, below_mean_mean <dbl>,
## #   sd1_mean <dbl>, sd2_mean <dbl>, kid_inc_rank_mean <dbl>,
## #   rand_val_mean <dbl>, treatment_group_mean <dbl>, id_num_sd <dbl>,
## #   kid_income_sd <dbl>, incarcerated_sd <dbl>, child_education_sd <dbl>,
## #   child_college_sd <dbl>, child_sat_sd <dbl>, parent_inc_sd <dbl>, ...
```

10d: Google form submisson [submitted]

10e: What is the purpose of random assigment in an experiment? Random assignment seeks to reduce or eliminate selection bias either on the part of participants (in the case of opt-in trials) or the researchers (in the case of researcher selection). For this reason, I would prefer to use random assignment to best achieve comparability.