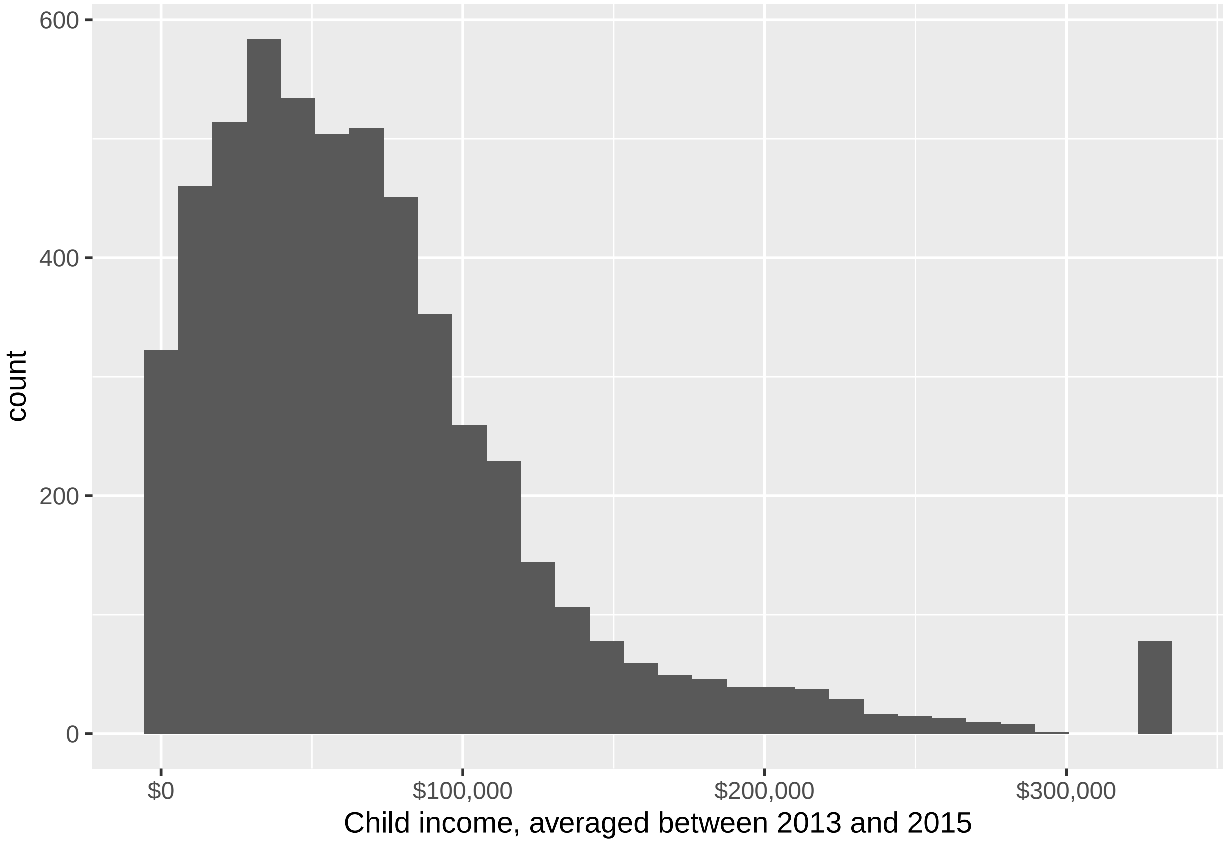
EC50_Lab1

2023-02-04

```r
#Creating histogram of kid_income
income_histogram <- nlsy97 |>
                    ggplot() +
                    geom_histogram(aes(kid_income)) +
                    labs(x = "Child income, averaged between 2013 and 2015") +
                    scale_x_continuous(labels = scales::dollar_format())
income_histogram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
#Calculating mean of kid_income
mean_kidincome <- mean(nlsy97$kid_income, na.rm = TRUE)
mean_kidincome
```

```
## [1] 70499.94
```

The sample mean of `kid_income` is $70499.94.

```r
#Creating indicator variable for values of kid_income below the mean and calculating the proportion of
nlsy97 <- nlsy97 |>
```

```
  mutate(below_mean = if_else(nlsy97$kid_income < mean_kidincome, 1, 0))

mean(nlsy97$below_mean)
```

```
## [1] 0.5960627
```

The sample mean of below_mean is .596, suggesing that just under 60% of the observations have a value of kid_income below its mean. This is because the dataset is skewed with a number of outlier values for kid_income, as seen in the histogram with the high number of right-side values of over $300,000.

```
#Calculating the median of kid_income
median_kidincome <- median(nlsy97$kid_income, na.rm = TRUE)
median_kidincome
```

```
## [1] 58750
```

The sample median of kid_income is $58750.

```
#Calculating the standard deviation of kid_income
sd_kidincome <- sd(nlsy97$kid_income, na.rm = TRUE)
sd_kidincome
```

```
## [1] 59552.02
```

The sample standard deviation of kid_income is $59552.02.

```
#Creating indicator variables for values of kid_income within 1 and 2 SDs of the mean and calculating t
upper_bound1sd <- mean_kidincome + sd_kidincome
upper_bound1sd
```

```
## [1] 130052
```

```
lower_bound1sd <- mean_kidincome - sd_kidincome
lower_bound1sd
```

```
## [1] 10947.92
```

```
upper_bound2sd <- mean_kidincome + (sd_kidincome * 2)
upper_bound2sd
```

```
## [1] 189604
```

```
lower_bound2sd <- mean_kidincome - (sd_kidincome * 2)
lower_bound2sd
```

```
## [1] -48604.1
```

```
nlsy97 <- nlsy97 |>
          mutate(within1sd = ifelse(kid_income <= upper_bound1sd & kid_income >= lower_bound1sd, 1, 0),
                 within2sd = ifelse(kid_income <= upper_bound2sd & kid_income >= lower_bound2sd, 1, 0))
nlsy97
```

```
## # A tibble: 5,486 x 19
##    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      3  150000       0      16       1     800   63000      12      12      1
##  2      4   76500       0      13       0      NA   11700      12      12      1
##  3     11   53250       0      16       1    1200   34500      14      12      1
##  4     12   71500       0      14       0      NA       0      12      12      0
##  5     13   60000       0      10       0      NA       0       6      12      0
##  6     14   17500       0      11       0    1600       0      12      12      0
##  7      16   37500       0       16       0    1000   13000       11       12       0
##  8      18   43000       0       13       0      NA       0       15       12       0
##  9      19   37500       1       12       0    1200       0       15       12       0
## 10      21   19000       0        8       0      NA       0       13       12       0
## # ... with 5,476 more rows, 9 more variables: black <dbl>, hispanic <dbl>,
## #   white <dbl>, region <dbl+lbl>, age2015 <dbl>, cohort <dbl>,
## #   below_mean <dbl>, within1sd <dbl>, within2sd <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
```

```
mean(nlsy97$within1sd)
```

```
## [1] 0.7867299
```

```
mean(nlsy97$within2sd)
```

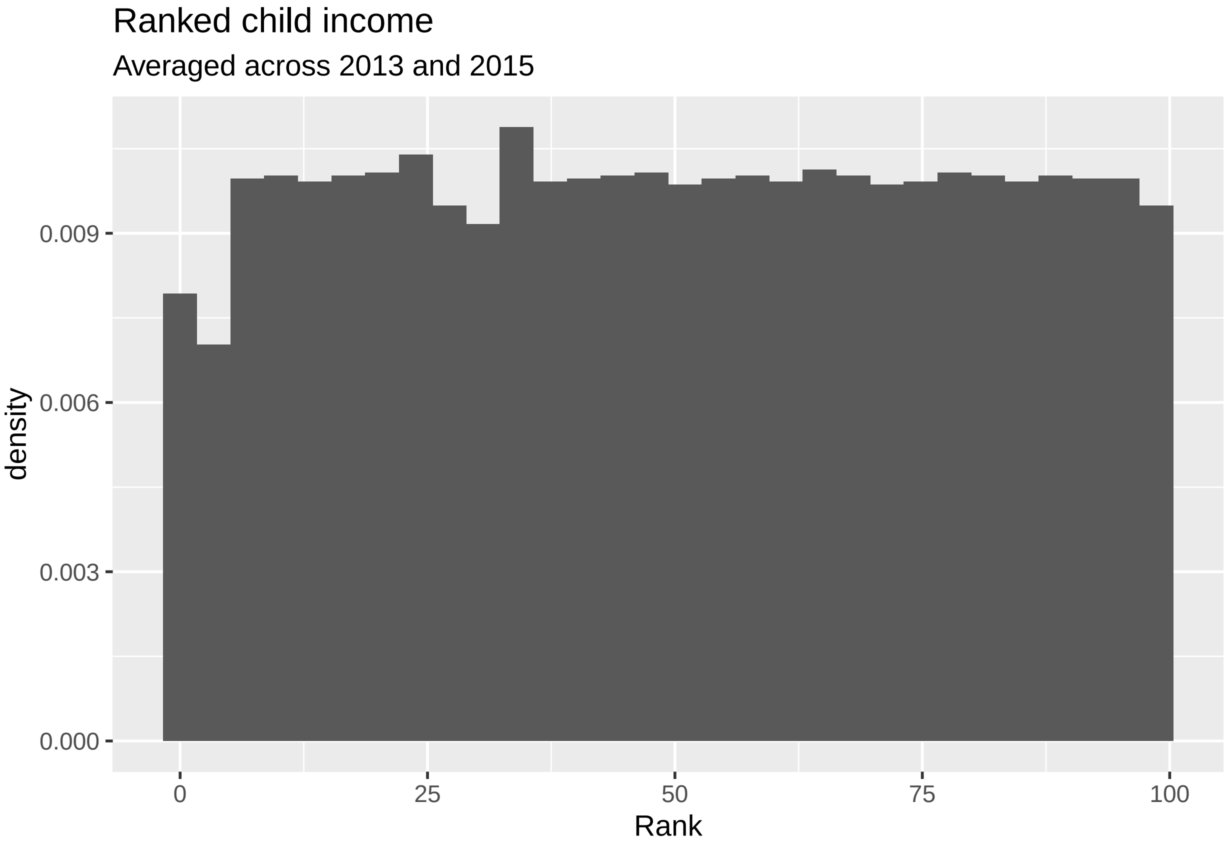
```
## [1] 0.948961
```

Roughly 79% of observations are within one SD of the mean of `kid_income` and almost 95% of observations are within two SDs of the mean of `kid_income`. This sample is not quite normally distributed.

```
#Creating ranked percentiles for all observations of kid_income
nlsy97 <- nlsy97 |>
          mutate(ranked_kidincome = rank(kid_income))

max_rank <- max(nlsy97$ranked_kidincome)

nlsy97$kid_inc_rank <-  (nlsy97$ranked_kidincome/max_rank)*100

percentile_rank <- function(variable){
  r <- ifelse(is.na(variable), NA, rank(variable, ties.method = "average"))
  100*r/max(r, na.rm = TRUE)
}

nlsy97$kid_inc_rank <- with(nlsy97, percentile_rank(kid_income))
view(nlsy97)

#Displaying the top 10 and bottom 10 observations ranked by kid_inc_rank
toprank <- nlsy97 |>
  arrange(desc(kid_inc_rank)) |>
  slice(1:10)

bottomrank <- nlsy97 |>
  arrange(kid_inc_rank) |>
  slice(1:10)
```

```
#Creating a histogram of kid_inc_rank, showcasing an approximately uniform distribution
rank_histo <- nlsy97 |>
  ggplot() +
  geom_histogram(aes(x = kid_inc_rank, y = after_stat(density))) +
  labs(title = "Ranked child income",
       subtitle = "Averaged across 2013 and 2015",
       x = "Rank")

rank_histo
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
#Calculating the mean and median of kid_inc_rank
mean(nlsy97$kid_inc_rank)
```

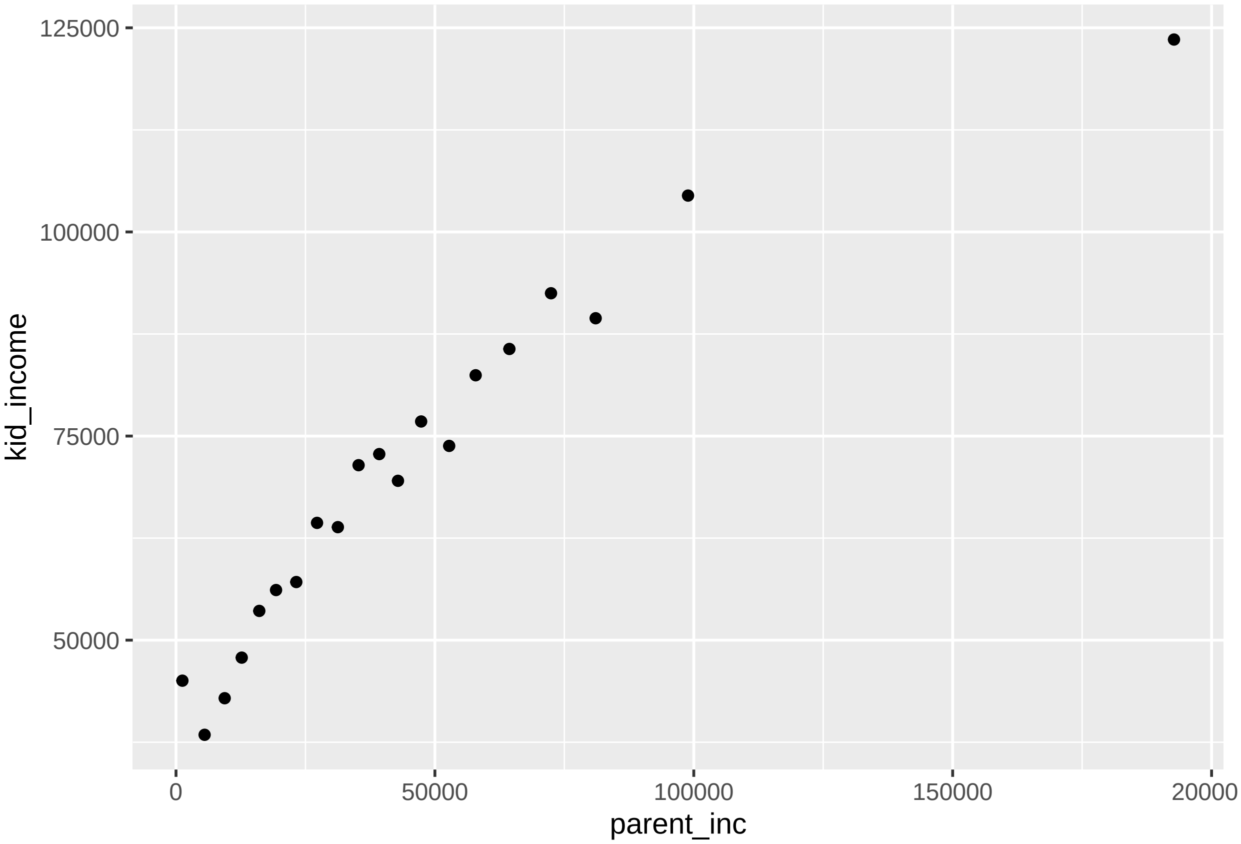
```
## [1] 50.08672
```

```
median(nlsy97$kid_inc_rank)
```

```
## [1] 50.1141
```

As shown above, the sample mean and median for `kid_inc_rank` are approximately equal, 50.08 and 50.11.

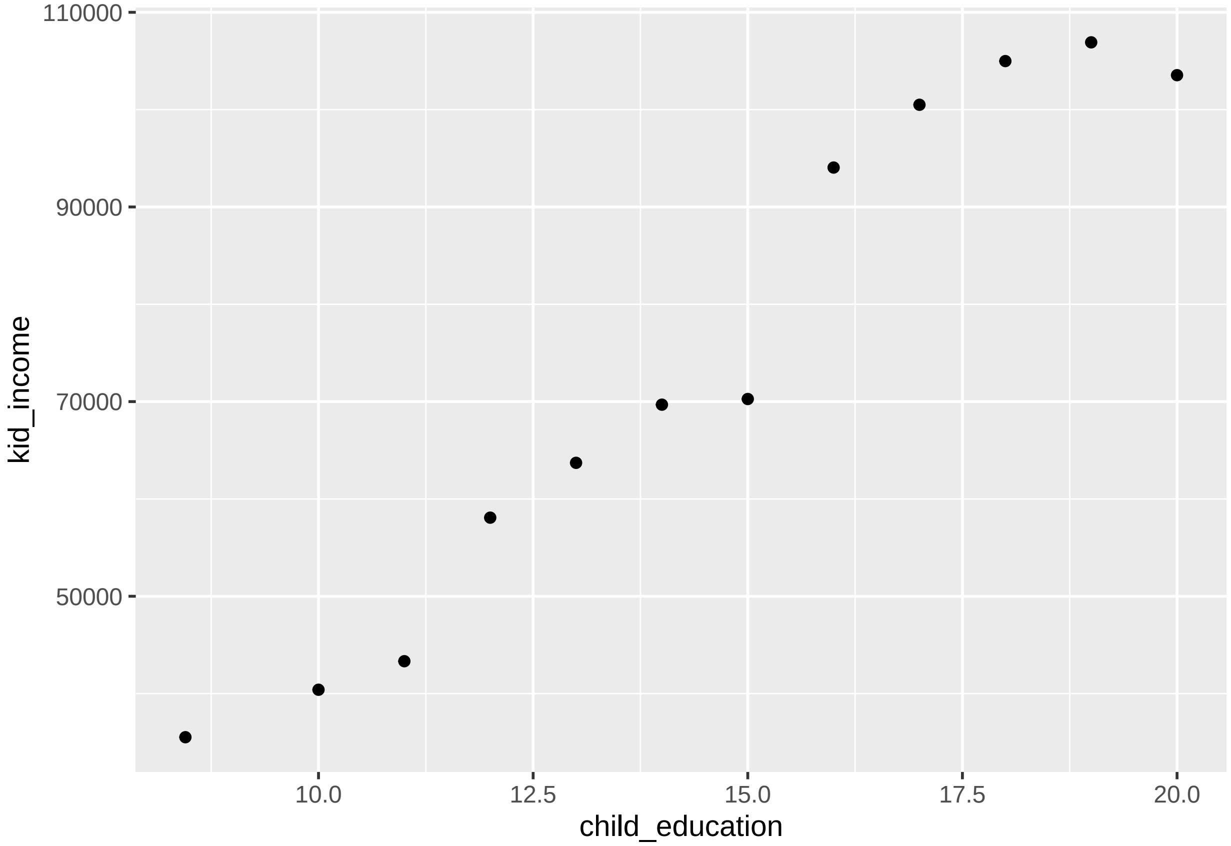
```
#examining the relationship between kid income and parents' income

linearscatter <- nlsy97 |>
  ggplot(aes(x = parent_inc, y = kid_income)) +
  stat_binmean(n = 20, geom = "point")
linearscatter
```



```
ggsave("linearscatter.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
#examining the relationship between kid income and years of education

nonlinearscatter <- nlsy97 |>
  ggplot(aes(x = child_education, y = kid_income)) +
  stat_binmean(n = 20, geom = "point")
nonlinearscatter
```



```
ggsave("nonlinearscatter.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
#ensuring reproducibility
set.seed(21519588)

#creating treatment and control groups
nlsy97 <- nlsy97 |>
  mutate(random_number = runif(length(parent_inc)),
         treatment_group = ifelse(random_number >= 0.5, 1, 0))

#Report total number of observations in treatment group
sum(nlsy97$treatment_group)
```

```
## [1] 2752
```

```
#Report total number of observations in control group
sum(1-nlsy97$treatment_group)
```

```
## [1] 2734
```

```
#Reporting summary statistics for all variables by treatment group
options(dplyr.width = Inf)
nlsy97 |>
  group_by(treatment_group) |>
  summarise_all("mean")
```

```
## # A tibble: 2 x 23
##   treatment_group id_num kid_income incarcerated child_education child_college
##             <dbl>  <dbl>      <dbl>        <dbl>           <dbl>         <dbl>
## 1               0  4620.     70975.       0.0947            13.8         0.290
## 2               1  4566.     70028.       0.104             13.8         0.299
##   child_sat parent_inc mother_education father_education female black hispanic
##       <dbl>      <dbl>            <dbl>            <dbl>  <dbl> <dbl>    <dbl>
## 1        NA     46852.             12.7             12.7  0.495 0.261    0.208
## 2        NA     45976.             12.7             12.7  0.507 0.269    0.189
##   white region age2015 cohort below_mean within1sd within2sd ranked_kidincome
##   <dbl>  <dbl>   <dbl>  <dbl>      <dbl>     <dbl>     <dbl>            <dbl>
## 1 0.595   2.68    33.0  1982.      0.594     0.787     0.947            2755.
## 2 0.605   2.63    33.0  1982.      0.598     0.786     0.951            2732.
##   kid_inc_rank random_number
##          <dbl>         <dbl>
## 1         50.3         0.248
## 2         49.9         0.749
```

```
nlsy97 |>
  group_by(treatment_group) |>
  summarise_all("sd")
```

```
## # A tibble: 2 x 23
##   treatment_group id_num kid_income incarcerated child_education child_college
##             <dbl>  <dbl>      <dbl>        <dbl>           <dbl>         <dbl>
## 1               0  2526.     59866.        0.293            2.99         0.454
## 2               1  2541.     59245.        0.306            3.01         0.458
##   child_sat parent_inc mother_education father_education female black hispanic
##       <dbl>      <dbl>            <dbl>            <dbl>  <dbl> <dbl>    <dbl>
## 1        NA     46375.             2.46             2.37  0.500 0.439    0.406
## 2        NA     45808.             2.52             2.35  0.500 0.444    0.392
##   white region age2015 cohort below_mean within1sd within2sd ranked_kidincome
##   <dbl>  <dbl>   <dbl>  <dbl>      <dbl>     <dbl>     <dbl>            <dbl>
## 1 0.491  0.985    1.39   1.39      0.491     0.409     0.225            1580.
## 2 0.489  0.988    1.41   1.41      0.490     0.410     0.215            1588.
##   kid_inc_rank random_number
##          <dbl>         <dbl>
## 1         28.8         0.144
## 2         29.0         0.143
```

Random assignment reduces the risk of bias influencing the results of any experiment. Subconscious or conscious bias when allocating groups through human judgment could lead to groups that are not approximately equal, which can lead to confounding factors that contributes to the fundamental problem of causal inference. Bias could lead to less than verifiable or reproducible results that claim to show a causal link or effect between two or more variables. Reducing bias strengthens the internal validity of any experiment's results and makes for a more robust experiment.