1980 Census PUMS

- 5% sample of US population
- Construct analysis sample of
  - Males 18-64
  - Work full time (30+ hours per week) full year (40+weeks per year)
- Two variables
  - ln(weekly earnings)
  - Years of education

- Total of 1,942,028 observations
- Estimate simple bivariate regression model
- $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$
- $\text{Ln(weekly earnings)}_i = \beta_0 + \text{EDUC}_i\beta_1 + \varepsilon_i$
- Treat as a 'population'
- "Actual" rate of return to education is 0.05135


```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        educ |   1942028    12.93697    3.026103          0         20
   weekearn |   1942028    379.2444    213.3202   120.0962   1829.268
  weekearnl |   1942028    5.813814    .4863825   4.788293   7.511672

. reg weekearnl educ

      Source |       SS       df       MS              Number of obs = 1942028
-------------+------------------------------           F(  1,1942026) =       .
       Model |  46888.0464     1  46888.0464           Prob > F      =  0.0000
    Residual |  412533.2571942026  .212424168          R-squared     =  0.1021
-------------+------------------------------           Adj R-squared =  0.1021
       Total |  459421.3031942027  .236567928          Root MSE      =  .46089

------------------------------------------------------------------------------
   weekearnl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0513475   .0001093   469.82   0.000     .0511333    .0515618
       _cons |   5.149533   .0014521  3546.32   0.000     5.146687    5.152379
------------------------------------------------------------------------------
```

Rate of return to education is 0.05135
-- treat as the true population value

- From population of almost 2 million
- Sample 500 observations from population
- Estimate OLS model
- Do this 500,000 times
- Look at the distribution of $\hat{\beta}_1$
- Took 3 hours to do simulation

Distribution of $\hat{\beta}_1$

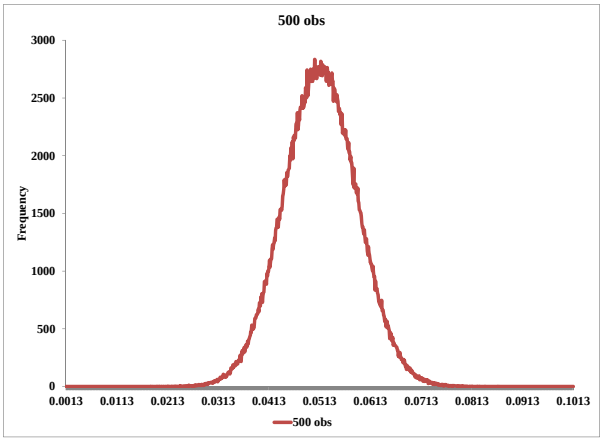
```
. sum d500

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        d500 |    500000    .0514286    .0072518   .0181311    .086526

. sum d500, detail

                            d500
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0345971       .0181311
 5%     .0395297       .0191557
10%     .0421598       .0196304       Obs              500000
25%     .0465359       .0202602       Sum of Wgt.      500000

50%     .0514128                      Mean           .0514286
                       Largest        Std. Dev.      .0072518
75%     .0563059       .0828049
90%     .0607337       .0828264       Variance       .0000526
95%     .0633986       .0839108       Skewness        .017823
99%     .0684116        .086526       Kurtosis       3.02065
```



Notice a few things

- Every time we draw a new sample, we get a different value for parameters
- The distribution of the parameter is essentially a normal distribution
- The distribution is centered on the true mean

Simulation

- Draw two independent series: x and ε
- 1000 obs
- For each draw, calculate the sample correlation coefficient
- Repeat 100,000 times
- What should we see?

Distribution of $\hat{\rho}(x, \varepsilon)$

Sample size	Mean	Stand. Dev.	Min	Max
1,000	-7.4E-5	0.032	-0.129	0.134

9